Baier, Franziska; Kunter, Mareike

Construction and validation of a test to assess (pre-service) teachers' technological pedagogical knowledge (TPK)

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Studies in educational evaluation 67 (2020) 100936, 45 S.

Construction and Validation of a Test to Assess (Pre-service) Teachers' Technological Pedagogical Knowledge (TPK)

Franziska Baier*[a] & Mareike Kunter[a]

*Corresponding author:

Mail: baier@psych.uni-frankfurt.de

[a]Goethe University Frankfurt

Theodor-W.-Adorno-Platz 6

60629 Frankfurt am Main

Germany

baier@psych.uni-frankfurt.de

kunter@paed.psych.uni-frankfurt.de

Acknowledgements

The authors would like to thank the experts who rated the test items and gave valuable feedback on the coding scheme.

Statements on conflict of interest and funding

- Declarations of interest: none
- This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

As society becomes increasingly digital, teachers must be trained to integrate technology effectively into their classrooms. Teachers' technological pedagogical knowledge (TPK), as defined in the TPACK framework (Koehler & Mishra, 2009), is considered an important prerequisite for effectively integrating technology. The TPACK framework has received a great deal of attention, yet few knowledge tests have been developed that directly assess TPK. However, those tests are crucial for evaluating the effectiveness of teacher education courses on technology integration. We have developed a 17-item test that covers teacher knowledge about various digital technologies as employed in teaching. Experts rated the items to represent the construct adequately. Data obtained from 245 pre-service teachers supports the test's internal structure. Concerning convergent and discriminant validity, the pre-service teachers' test scores were not related to their self-reported TPK, but to their self-reported technological knowledge. The test was sensitive to changes in pre-service teachers' TPK through teacher education courses.

Construction and validation of a test to assess (pre-service) teachers' technological pedagogical knowledge (TPK)

Teachers' professional knowledge and the instructional quality they provide are important predictors of student achievement (e.g. Jackson, Rockoff, & Staiger, 2014; Kunter et al., 2013). As all aspects of life become increasingly digital, teachers face the challenge of providing high-quality instruction by effectively integrating digital media into their teaching (Chai, Koh, & Tsai, 2013). Thus, a major aim in educational research and teacher education programs has been to research and develop teachers' knowledge about teaching with digital media (Graham et al., 2009; ISTE, 2008). The most widely used framework to study aspects of teachers' technology-related knowledge, the TPACK framework (Mishra & Koehler, 2006), extends Shulman's (1986) concept of pedagogical content knowledge (PCK), understood as an amalgam of content knowledge (CK) and pedagogical knowledge (PK), by integrating technological knowledge (TK; see Figure 1). Introducing TK triggers three additional forms of knowledge: technological pedagogical knowledge (TPK) at the intersection of PK and TK, technological content knowledge (TCK) at the intersection of TK and CK and, finally, technological pedagogical content knowledge (TPCK or, later, TPACK) at the intersection of PK, CK, and TK (Figure 1). TPACK, at the core of the framework, has been the most widely investigated as researchers have emphasized the special importance of the subject content for effective integration of technology (e.g. Yurdakul et al., 2012). However, other forms of technology-related knowledge, such as subject-independent TPK and TCK, have received less attention (Cox, 2008; Graham, 2011). This is particularly surprising for TPK, as other models emphasize the importance of generic technology-related pedagogical knowledge for teaching (e.g. Knezek & Christensen, 2016). Numerous digital learning tools can be applied effectively across subjects to enhance instructional quality (Chai et al., 2013).

Besides the lack of research into TPK, the lack of performance-based instruments (e.g. knowledge tests) to measure technology-related knowledge areas is an even more notable problem (Drummond & Sweeney, 2017). Self-report measures are often used, which assess the teachers' confidence in, or perception of, their own knowledge rather than their actual available knowledge (Hofer, Grandgenett, Harris, & Swan, 2011; Scherer, Tondeur, & Siddiq, 2017). However, without performance-based knowledge tests to measure the TPACK constructs, research questions about the importance of teachers' actual technology-related knowledge for student outcomes cannot be adequately investigated. Moreover, psychometrically sound research instruments are of practical relevance for evaluating the effectiveness of teacher education courses on technology integration (Lawless & Pellegrino, 2007) and pre-service teachers may use them to reflect on their technology-related knowledge.

Given the described research gaps, this study aims to further specify the concept of TPK. First and foremost, this involves developing and validating a knowledge test to directly measure TPK. Such a measurement can advance our understanding of teachers' knowledge about teaching with digital media – it may provide more information on the structure and development of TPK and be used in future evaluation and teacher effectiveness studies.

Theoretical Background

Conceptualization of Technological Pedagogical Knowledge (TPK)

Generally, TPK is understood to be subject-independent – it is generic knowledge about technology-enhanced teaching (Graham et al., 2009). Koehler and Mishra (2009) define TPK as "an understanding of how teaching and learning can change when particular technologies are used in particular ways. This includes knowing the pedagogical affordances and constraints of a range of technological tools as they relate to disciplinarily and developmentally appropriate pedagogical designs and strategies" (p. 65). Whereas many researchers offer solid (overlapping) definitions of TPK (Chai et al., 2013), the notion of TPK and how it is related to the other bodies of knowledge in the framework is less clear (e.g. Voogt, Fisser, Pareja Roblin, Tondeur, & van Braak, 2013). Mishra and Koehler (2006) consider general pedagogical knowledge (PK) and technological knowledge (TK) relevant to understanding TPK. Researchers define PK as domain-general knowledge of the field of teaching including, for example, knowledge about learning and instruction, classroom management and organization and learner characteristics (e.g. Shulman, 1987; Voss, Kunter, & Baumert, 2011). Mishra and Koehler (2006) define TK as "knowledge about standard technologies, […], and more advanced technologies, such as the Internet and digital video." (p. 1027), and argue that their definition resembles Fluency of Information Technology (FITness) as suggested by the Committee of Information Technology Literacy of the National Research Council (Koehler & Mishra, 2009; NRC, 1999). FITness requires a person to "understand information technology broadly enough to apply it productively at work and in his/her everyday life, to recognize when information technology can assist or impede the achievement of a goal, and to continually adapt to changes in information technology" (Koehler & Mishra, 2009, p. 64). Based on their TPACK framework, Mishra and Koehler (2006) devise TPK as the overlap between PK and TK. They take an integrative perspective by presenting TPK as the interplay between and intersection of PK and TK (Voogt et al., 2013). Thus, TPK is generated when teachers draw on and connect knowledge from other knowledge bases (i.e. PK and TK; Koehler & Mishra, 2009; Neumann, Kind, & Harms, 2018). However, a more transformative viewpoint (Angeli & Valanides, 2009) considers TPK a new, distinct body of knowledge that emerges when PK and TK are combined and that can be developed on its own. Yet, other authors simply conceptualize TPK as an extension of PK (Graham et al., 2009).

Previous Measures of Technological Pedagogical Knowledge (TPK) and Related Constructs

This study adopts a *cognitive perspective* on TPK, suggesting that it is a teacher attribute that can be measured separately from actual classroom teaching by means of questionnaires or tests (Depaepe, Verschaffel, & Kelchtermans, 2013). Other researchers have taken a more *situated perspective* on teacher knowledge and propose that this knowledge can only be assessed within the classroom context where it is enacted (Depaepe et al., 2013) by means of classroom observations, for example. A couple of observation rubrics to measure TPK or technological pedagogical content knowledge (TPACK) have already been developed (e.g. Heitink et al., 2017; Hofer et al., 2011). Reviewing those rubrics is beyond the scope of this paper, however.

In general, when measuring the TPACK constructs from a *cognitive perspective*, the focus has primarily been on questionnaires that ask teachers to self-report their own knowledge (Hofer et al., 2011; Voogt et al., 2013). Among the most widely used self-report surveys to measure the TPACK constructs (and thus TPK) is a survey developed by Schmidt et al. (2009), who have provided the first evidence for the reliability and the construct validity of the individual scales. Valtonen et al. (2017) developed another self-report instrument, the TPACK for 21st century skills. The authors underlined the construct validity of their questionnaire by finding that a six-factor CFA model (in line with the TPACK framework) fitted their data well.

Although there are many instruments for self-assessment of technology-related knowledge, the validity of these tools should be considered critically (Maderick, Zhang, Hartley, & Marchand, 2015). For example, if participants are inexperienced in a certain domain, they may not be able to accurately judge their own knowledge and thus may over- or underestimate their competence (Kuh, 2003; Maderick et al., 2015). Furthermore, self-reported knowledge has been shown to represent teachers' confidence in their knowledge rather than their actual knowledge and how it is cognitively structured and represented (Lawless, Kulikowich, & Smith, 2002; Scherer et al., 2017). Often-repeated criticism of self-report research is that it is subject to socially desirable responding (Paulhus, 2017). Even though there are correlations between (pre-service) teachers' self-rated knowledge and their knowledge as assessed by knowledge tests, they are typically small (König, Kaiser, & Felbrich, 2012; Maderick et al., 2015).

Overall, there are few performance-based knowledge tests to measure the technological aspects of the TPACK framework from a *cognitive perspective* (Drummond & Sweeney, 2017). This contrasts with the measurement of the more traditional aspects of the framework, i.e. content knowledge (CK), pedagogical knowledge (PK) and pedagogical content knowledge (PCK). For example, Voss, Kunter and Baumert (2011) and König and Blömeke (2009) developed objective tests to measure teachers' general PK.

Lachner, Backfisch and Stürmer (2019) have presented a knowledge test to more objectively measure TPK. They developed a test consisting of closed-ended questions to measure teachers' conceptual (8 items) and situational (10 items) TPK (de Jong & Ferguson-Hessler, 1996): Conceptual TPK includes knowledge about facts and concepts associated with, for example, the potential of educational technology for scaffolding learning, while situational TPK includes knowledge about strategies regarding the use of educational technology for different teaching practices (e.g. adaptive technologies; Lachner et al., 2019). Lachner et al. (2019) have provided the first evidence for the factorial validity of their instrument based on a heterogeneous sample of 248 student teachers, other students and in-service teachers. Furthermore, in a sub-sample of 120 teachers, they found a positive relationship between PK and TPK, which could be interpreted in terms of their test's construct validity. The instrument must be further validated in terms of its criterion and construct validity, however.

As this literature review shows, there is a lack of knowledge tests that more objectively measure pre-service and in-service teachers' TPK. We have developed and validated a knowledge test to measure (pre-service) teachers' TPK to help close this research gap. This is of the utmost importance, as a psychometrically sound, performance-based knowledge test will help explore the importance of technological knowledge areas defined in the TPACK framework for teachers' instructional quality and their students' achievements. Furthermore, such instruments will help researchers and teacher educators to better evaluate the effectiveness of interventions on teachers' technology-related skills in formal teacher education.

Aims and Hypotheses

This study introduces a new instrument to measure technological pedagogical knowledge (TPK). Furthermore, we employ a mixed-method approach to provide the first evidence for its content validity (Study 1) and construct validity (Study 2). A test's validity is one of the most important criteria by which its psychometric quality is judged (AERA, APA, & NCME, 2014; Hartig, Frey, & Jude, 2012). It describes the extent to which the quality of interpretation based on test scores is backed up by empirical evidence and theoretical argumentation (Hartig et al., 2012). The following validation strategy was carefully chosen based on scientific reasons (i.e. important elements of validity) in addition to practical reasons (e.g. available resources). Whereas we evaluated the quality and content validity of the individual test items (the content of the individual items indeed assesses TPK), the validity of the whole instrument was investigated in terms of its internal structure (the correlation among items can be explained by a single latent factor), convergent and discriminant validity (high/low correlation among test score and other variables, Hartig et al., 2012) and its instructional sensitivity. The instructional sensitivity reflects the extent to which a test is sensitive to an implemented curriculum (Naumann, Musow, Aichele, Hochweber, & Hartig, 2019). This was particularly important to us, as many stakeholders (e.g. lecturers) may seek an instrument to comprehensively evaluate their courses on technology integration into teaching. We formulated the following hypotheses:

- Hypothesis 1 (Content validity, Study 1): We expect educational researchers to consider the individual items clear, relevant to teaching and representative of our construct TPK.
- Hypothesis 2 (Content validity, Study 1): We expect pre- and in-service teachers to judge the individual items as clear and relevant to teaching.
- Hypothesis 3 (Internal structure, Study 2): We expect that our test items are indicators of the single latent construct TPK. The model satisfactorily explains our empirical data structure.
- Hypothesis 4 (Convergent and discriminant validity, Study 2):
    a) Based on previous studies' reports of overlap between self-reported and tested knowledge, we expect a small to moderately positive relationship between pre-service teachers' performance in our test and their self-rated TPK (convergent validity).
    b) We expect a small positive relationship between pre-service teachers' performance in our test and their self-rated technological knowledge (TK; discriminant validity).
    c) We expect a small positive relationship between pre-service teachers' performance in our test and their self-rated pedagogical knowledge (PK; discriminant validity).
- Hypothesis 5 (Instructional sensitivity, Study 2): We expect our test to be sensitive to changes in pre-service teachers' TPK during teacher education courses that explicitly aim at fostering pre-service teachers' knowledge about teaching with digital media.

Study 1: Instrument Construction and Content Validity

Specification of Technological Pedagogical Knowledge (TPK) and Test Development

Development of this test was guided by the TPACK framework (Mishra & Koehler, 2006), the common definitions of TPK found in the literature and, more generally, theories and findings from the field of instructional technology and media in learning (e.g. Clark, 2001; Kozma, 1991). We conducted extensive literature searches including journal articles, handbooks and reviews in these areas to conceptualize TPK and to create the test items. The following paragraphs present our conceptualization and operationalization of TPK based on this literature.

All instruction and learning require the selection and use of at least one medium (Reiser & Gagné, 1983), i.e. media form an essential part of learning and instruction. As there is a great variety of media and technologies to choose from, "teachers must be aware of the pedagogical affordances and constraints of a range of technological tools" (Koehler & Mishra, 2009, p.65). This kind of teacher knowledge is understood and conceptualized as TPK (based on the TPACK framework) as various definitions indicate (for an overview, see Cox, 2008). We created items to tap this knowledge, namely about the existence of various technologies useful for teaching, their capabilities (potential) in terms of designing and organizing learning and instruction, and their pedagogical opportunities and constraints. When using this definition to create items, two questions need further consideration: (1) "To which instructional technologies are we referring exactly?" and (2) "What are important, concrete aspects of classroom instruction and assessment that these technologies can enable and support?"

Concerning the first question, our understanding of TPK is limited to teaching with digital (electronic) media, i.e. excluding analog media. This is in line with several other definitions of technological knowledge (TK) and TPK in the literature (Voogt et al., 2013) and also with the notion of *computer* literacy and *information technology* literacy that Koehler and Mishra refer to in their 2009 paper. We concentrate on technologies that have already been used in teaching and, thus, can be considered a relevant aspect of practical teacher knowledge, e.g. interactive whiteboards or tablets (Torff & Tirotta, 2010; Walczak & Taylor, 2018). Furthermore, we concentrate on software applications that can be used for teaching and that are categorized based on theoretical frameworks such as the *Digital Transformation Framework for ICT Literacy* of the International ICT Literacy Panel (ETS, 2002). These software applications include word processing systems, spreadsheets, presentation and graphics software, communication applications and Internet-based search engines (Maderick et al., 2015; Senkbeil et al., 2013).

Concerning the second question, we reviewed theoretical frameworks and empirical findings that identify and systemize different important aspects of classroom teaching, learning and assessment (e.g. Brophy, 1999; Oser & Baeriswyl, 2001; Voss et al., 2011). For example, Voss et al. (2011) contend that teachers' general pedagogical and psychological knowledge entail the following aspects of classroom teaching: classroom organization (maximizing time on task), a broad repertoire of teaching methods (e.g. direct instruction or more student-centered approaches: problem-based learning, cooperative learning), classroom assessment and feedback (formative and summative) and, finally, adapting instruction to individual students' needs (e.g. low motivation, special needs students). Technology and media have been employed to enhance these exact aspects of instruction; that is, they have been used to deliver instruction (e.g. Reiser, 2001), to adapt instruction to students' individual needs (Kulik & Fletcher, 2016), to enhance student-centered learning (Morrison & Lowther, 2002), to motivate students (e.g. Yang, 1991) and to assess students' performance (Clarke-Midura & Dede, 2010).

Guided by this systematization of technologies, aspects of classroom instruction and previous use of technology in teaching and learning, we developed 17 items to measure pre- and in-service teachers' TPK. An overview of the item names and their mapping can be found in Appendix A. The items focus on one or more digital technology and software applications (e.g. as defined by the International ICT Literacy Panel; ETS, 2002) and one or more aspects of classroom teaching as defined in Voss et al. (2011). Items 4, 10, 15 and 17 directly address the two digital technologies, interactive whiteboards and tablets, commonly used in everyday teaching. The other items cover one or more of the software applications described above that can be employed be means of tablets or interactive whiteboards. For example, Items 6, 9, 11, 13 and 16 cover (among others) applications that can be used for communication and Items 1 and 14 are concerned with Internet-based search engines. Whereas some items are focused on one specific digital tool/hardware and ask about its potential for different aspects of classroom teaching (e.g. Items 4, 6, 10), other items are formulated the other way and ask how different digital tools could enable or enhance specific aspects of classroom teaching (e.g. adaptivity, assessment/feedback, Items 2, 12). Every item is concerned with the existence of digital technologies and, more specifically, their capabilities (or constraints in Item 1) for learning and instruction. Thus, we expect our test to have a unidimensional structure.

We generated items with a free-response format, as we primarily want to tap in-depth teacher knowledge and avoid creating items that were too easy (e.g. Buchholtz et al., 2013; Hohensinn & Kubinger, 2009). The stem of each item can be found in Appendix B. Along with each item, we deductively developed a coding scheme, that specifies the categories of correct (and incorrect) answers derived from the reviewed research literature. As an example, the coding scheme of Item

5 may be found in Appendix C. Moreover, our items capture both conceptual/formal and more situational/practical TPK, as we deem both types of knowledge relevant to teaching.

Next, we describe researchers' and pre- and in-service teachers' ratings of the individual items (and the corresponding coding scheme) to establish content validity of the individual test items.

Method of Study 1

Sample

Study 1 consisted of several subsamples of raters, namely four German educational researchers (two female), six in-service teachers (two female) and 13 pre-service teachers (eight female). Samples with differing levels of expertise were recruited in order to verify different aspects of content validity (Haynes, Richard, & Kubany, 1995). Educational researchers were included for their expertise in the theory and assessment of technological pedagogical knowledge (TPK), so that they could judge the items' construct fit. The in-service teachers were primarily invited to judge the practical relevance of the items to everyday teaching. The pre-service teachers were primarily (but not exclusively) asked to evaluate the items' clarity – since they were the target group of Study 2, it was important that they correctly understood the items. All participants (researchers, pre- and in-service teachers) were compensated with a coupon.

The researchers were selected based on their research profiles. In particular, researchers with several years of research experience in teaching with digital media and backgrounds in psychology and/or pedagogy were preferred. Even though we invited at least six researchers (as recommended by Haynes et al., 1995), only four (each from a different German university) had the time to evaluate the whole tool (17 item stems and their respective coding schemes) in detail.

In-service teachers were recruited from a German secondary school. These teachers had, on average, six years of work experience (*SD* = 3.44). Mathematics, foreign languages, informatics, politics, history, geography, physics and PE were among the subjects they taught. Several of the in-service teachers had tight schedules, so they reviewed different numbers of items: Items 10, 14, 15 and 17 were rated by only one teacher, Item 12 by two teachers and Item 8 by three teachers.

Pre-service teachers were recruited at a German university. They were students of primary or secondary education and were, on average, in the sixth semester of their degree programs (*SD* = 2.07). The subjects they were studying to teach included mathematics, foreign languages, German as a first language, computer science, politics, history, religion and PE. Again, not all pre-service teachers rated all items due to the time constraints of testing. Items 6, 10, 12 and 17 were rated by only five pre-service teachers and Items 8, 14 and 15 by four pre-service teachers.

Procedure

The four researchers were invited to participate in the study via e-mail. Upon agreement, they received an interactive text document with the test items, the coding scheme and scales to rate both items and coding scheme. The researchers rated the test items according to their clarity, their relevance for teaching and most importantly their construct fit on a four-point Likert scale, ranging from 1 (“disagree”) to 4 (“agree”). They also rated the correctness of the proposed answers in the coding scheme and whether they thought that any answer categories were missing.

Pre- and in-service teachers answered an online questionnaire. They first responded to each test item and then rated each item in terms of its clarity and relevance for teaching on a six-point Likert scale, ranging from 1 (“fully disagree”) to 6 (“fully agree”). They were also given the opportunity to comment on each item.

Data Analysis

The average researcher, in-service teacher and pre-service teacher ratings were calculated for each test item on each scale (fit to construct, clarity, relevance). The coding scheme was qualitatively revised based on the researchers' comments and the answers to the items provided by the pre- and in-service teachers.

Results of Study 1

Results of Researcher Rating

The researchers rated every item as clearly formulated (values between 3.25 and 4.00) except for Item 5 (Functions), which scored 2.75 (*SD* = 0.50). They also rated every item relevant for teaching (values between 3.50 and 4.00) except for Item 17 (IWB periphery) with a rating of 2.75 (*SD* = 1.26). Finally, and most importantly, they agreed that every item represented technological pedagogical knowledge (TPK; values ranged from 3.00 to 4.00) except for Item 1 (Internet search), which received a rating of 2.50 (*SD* = 1.73).

Overall, the researchers were quite positive about the test items. Most importantly, only one item (Item 1) was, on average, evaluated as non-representative of TPK. Note, however, that the standard deviation for this item is quite high, showing that researchers' opinions on this item differed substantially.

Results of Pre- and In-Service Teachers Rating

Both the pre- and in-service teachers rated every item as clearly formulated. For the pre-service teachers, the clarity ratings ranged from 4.50 to 5.77, while the in-service teachers rated the clarity between 4.00 and 6.00. Moreover, the pre-service teachers rated every item relevant for teaching. Their values ranged from 4.25 to 5.60. The in-service teachers rated every item relevant for teaching (values ranged from 4.00 to 5.50) except for Item 12 (Feedback) with a rating of 2.00 (*SD* = 1.41) and Item 14 (WebQuest) with a rating of 3.00 (only one teacher rated this item).

Overall, the pre- and in-service teachers gave the items very positive ratings. They found the items to be clearly formulated and relevant to teaching.

Revision of the Test

Across the three groups, no particular item stood out as inappropriate. Therefore, every item on the test was kept, but the coding scheme and scoring were revised. The former was revised based on the few comments that the researchers made about it as well as the answers to the items provided by the pre- and in-service teachers. The number of correct answers per item differs, with a minimum of 4 and a maximum of 13. A participant's score on an item is the sum of conceptually distinct correct answers provided for that item.

Study 2: Validation Study With Pre-Service Teachers

Method of Study 2

Sample and Design

The sample of Study 2 consisted of 255 pre-service teachers (69% female) who were recruited at a German university during the winter terms of 2017-18 and 2018-19 and the summer term of 2018. Two subsamples of pre-service teachers participated in the winter term of 2017-18 (Group 1a: n = 32 and Group 1b: n = 34), 102 pre-service teachers participated in the summer term of 2018 (Group 2) and 87 participated in the winter term of 2018-19 (Group 3). Participants in Group 2 only received 10 of the 17 technological pedagogical knowledge (TPK) items (Items 1, 4-7, 10, 11, 13, 15, 16) due to time constraints in the teacher education course.

The average participant was enrolled in their 4th semester (*SD* = 2.27) of their degree program. Participants were studying to become a teacher at different types of schools: 27.8% were preparing for primary school, 34.5% for lower and middle secondary track, 28.6% for the highest secondary track and 9% preparing for special education. Participants average grade in their secondary school leaving examination (Abitur; range: 1 [very good] – 6 [insufficient]) was *M* = 2.34 (*SD* = 0.54).

To investigate the test's instructional sensitivity, participants in Groups 1a, 1b and 3 answered every item twice. In Group 1a, the pre-service teachers took the test before and after a five-week teacher education course on teaching with digital media, whereas Group 1b served as a control group and took the test twice at an interval of five weeks without a specific intervention in between.

Group 3 took the test before and after an eleven-week teacher education course on teaching with digital media. Whereas the setting in Group 1 reflects a more controlled setting (short intervention; experimental [1a] and control group [1b]), the setting in Group 3 (no control group) shows the test's sensitivity to changes in TPK over a longer-lasting teacher education course reflecting what is usually found in practice. In Group 3, we observed significant participant turnover; of the 87 pre-service teachers only 57 participated in both the pre-test and the post-test. Nevertheless, comparisons between the two groups of participants (complete data vs. missing post-test data) only revealed a difference in terms of their semester of their degree program ($t(26) = -2.08$, $p = .05$) but no differences in terms of any other pre-test variable (i.e. TPK test performance, self-reported technological knowledge [TK], pedagogical knowledge [PK] and TPK, gender, average grade, school type).

The teacher education courses in Groups 1a and 3 included short presentations, group work, hands-on experiences and discussions of different pedagogical aspects of teaching with different digital technologies. Group 2 was only tested once at the beginning of a teacher education course on digital media and was not further involved in evaluating the test's instructional sensitivity. A detailed overview of the different samples and measurement points is provided in Appendix D.

Procedure

Most participants (Groups 1a, 2 and 3) worked on the assessment in class during their teacher education course and took approximately 30 to 45 minutes to complete the whole assessment (including the knowledge test and self-report scales). Participants of Group 1b (control group to Group 1a) were recruited outside of a specific teacher education course, worked on the assessment in a laboratory and were compensated with a coupon.

Instruments

The pre-service teachers took the new 17-item technological pedagogical knowledge (TPK) test (free-response format). For a detailed description of the test, see Study 1. We segmented the participants' answers to an item into meaningful chunks. Afterwards, three trained raters independently coded 15 percent of the material based on the coding scheme. We calculated Krippendorff's alpha as the coefficient of interrater-reliability, which is highly flexible in terms of the measurement scale and number of raters (Zapf, Castell, Morawietz, & Karch, 2016). Interrater-reliability was satisfactory to very good, ranging from .70 to .95. As a measure of internal consistency, the Omega coefficient (McDonald, 1999) is reported throughout the paper as it has been shown to be a more sensible index of internal consistency than Cronbach's alpha (e.g. Deng & Chan, 2017). The internal consistency of the test including all 17 items (at the first measurement point) was satisfactory: $\Omega = .69$.

We assessed the pre-service teachers' self-rated technological knowledge (TK), pedagogical knowledge (PK) and TPK by items adapted and translated into German from the scales by Schmidt et al. (2009). All items were rated on a six-point Likert scale, ranging from 1 ("fully disagree") to 6 ("fully agree"). Pre-service teachers' TK was assessed by five items, an example of one such item is: "I know how to solve my own technical problems". Pre-service teachers' PK was assessed by four items, an example of one such item is: "I can use a wide range of teaching approaches in a classroom setting". Pre-service teachers' TPK was assessed by five items, an example of one such item is: "I know how to use digital media effectively for learning and instruction". All scales had good internal consistencies: $\Omega = .86$ for TK, $\Omega = .82$ for PK and $\Omega = .85$ for TPK.

Data Analysis

Confirmatory factor analysis in MPlus 5.1 was used to analyze the internal structure of the test (Muthén & Muthén, 2007). As indicators of overall goodness of model fit, the chi-square statistic (good model fit = non-significant chi square value) and several descriptive fit indices are reported: the comparative fit index (CFI $\geq .97$), the root-mean-square error of approximation (RMSEA $\leq .05$) and the standardized root-mean-square residual (SRMR $< .05$; Schermelleh-Engel, Moosbrugger, & Müller, 2003). To analyze the relationships between the technological pedagogical knowledge (TPK) test and self-reported technological knowledge (TK), pedagogical knowledge (PK) and TPK, latent correlations were calculated in MPlus5.1 (Muthén & Muthén, 2007). A full information maximum likelihood (FIML) algorithm was used to estimate missing values. As already described above, Group 2 had missing values on seven items (missing completely at random).

To investigate the test's sensitivity to changes in pre-service teachers' TPK (Hypothesis 5), a sum-score was calculated across the test items for each participant in the pre- and post-test. Afterwards, repeated measures variance analyses were conducted in SPSS (Version 25). Significance testing was performed at the .05 level for every analysis.

Results of Study 2

Internal Structure of the Technological Pedagogical Knowledge (TPK) Test (Hypothesis 3)

Confirmatory factor analysis based on a sample of 245 pre-service teachers and their data at measurement point 1 (prior to any specific intervention) was used to investigate the test's internal structure (ten of the 255 pre-service teachers had participated in the post-test, only). As the measurement model with all 17 items (Model 1) did not satisfactorily fit the data, $\chi^2_{[119]} = 172.02$, $p = .00$, CFI = .79, RMSEA = .04, SRMR = .07, the model was re-specified based on the researcher ratings and non-significant parameter estimates (i.e. factor loadings). As the researchers, on average, evaluated Item 1 as non-representative of the construct, this item was removed from the test. Upon further examination, two indicators did not load significantly on the latent factor of TPK: Items 8 and 14 (see Table 1). These two items were also eliminated from the model. The measurement model, Model 2, including 14 items (excluding Items 1, 8 and 14), fit the data acceptably to well: $\chi^2_{[77]} = 88.73$, $p = .17$, CFI = .94, RMSEA = .03, SRMR = .06. Table 1 shows the standardized factor loadings. In Model 2, the factor loadings ranged between .22 and .51. Overall, these results indicate that our test items (excluding Items 1, 8 and 14) are indicators of the single latent construct TPK. Model 2 satisfactorily explains our empirical data structure.

Table 1

Standardized Factor Loadings and Their Standard Errors for Every Item

Indicator	Model 1: All Items		Model 2: Optimized Test	
	Estimate	Standard error	Estimate	Standard error
01 Internet search	.28*	.07	--	--
02 Adaptivity	.25*	.09	.24*	.09
03 Serious Games	.20*	.09	.22*	.09
04 IWB usage	.34*	.07	.33*	.07
05 Functions	.36*	.07	.39*	.07
06 LMS usage	.51*	.07	.51*	.07
07 eExam	.44*	.07	.42*	.07

08 Constructivism	.16	.09	--	--
09 Homework delivery	.45*	.08	.44*	.08
10 Tablet usage	.47*	.07	.47*	.07
11 LMS examples	.35*	.07	.37*	.07
12 Feedback	.31*	.09	.30*	.09
13 Communication	.41*	.07	.43*	.07
14 WebQuests	.08	.10	--	--
15 Inclusion	.23*	.08	.23*	.08
16 ePortfolios	.37*	.07	.36*	.07
17 IWB periphery devices	.47*	.08	.48*	.08

Note. *$p < .05$; IWB = interactive whiteboard; LMS = learning management system.

Convergent and Discriminant Validity of the Technological Pedagogical Knowledge (TPK) Test (Hypotheses 4a-c).

Before calculating latent correlations among the constructs (based on the sample of 245 pre-service teachers), the measurement models of the three self-report scales were evaluated. The measurement model for self-reported TPK fit the data well: $\chi^2_{[5]} = 4.17$, p =.53, CFI = 1.0, RMSEA = .00, SRMR = .01, as did the measurement model for self-reported technological knowledge (TK), $\chi^2_{[5]} = 9.68$, p =.09, CFI = .99, RMSEA = .06, SRMR = .02. The model fit for self-reported pedagogical knowledge (PK) was satisfactory, $\chi^2_{[2]} = 6.78$, p =.03, CFI = .99, RMSEA = .10, SRMR = .02.

Table 2 shows latent correlations among the constructs, TPK as measured by the new test (without Items 1, 8 and 14), self-reported TPK, self-reported TK and self-reported PK. TPK as measured by the test had a small, non-significant correlation with self-reported TPK (r = .15, p = .07), however, self-reported TK was significantly related to TPK as measured by the test (r = .20, p = .02). Self-reported PK and TPK as measured by the test were not correlated (r = .12, p = .18). Therefore, all things considered, self-report measures could not be used to establish convergent and discriminant validity of the test.

Table 2

Latent Inter-Correlations Among TPK Test and Self-Reported TPK, TK and PK

	1	2	3
1 TPK test			
2 TPK (self-report)	.15		
3 TK (self-report)	.20*	.36*	
4 PK (self-report)	.12	.51*	.18*

Note. *p < .05; TPK = technological pedagogical knowledge; TK = technological knowledge; PK = pedagogical knowledge.

Instructional Sensitivity of the Technological Pedagogical Knowledge (TPK) Test (Hypothesis 5).

Data from Group 1 (Groups 1a and 1b; quasi-experimental design; n = 61) and Group 3 (pre-post-test design, n = 57) were analyzed separately to test whether changes in pre-service teachers' TPK related to a university course on teaching with digital media are reflected in changes of the sum score in the test.

Results of the Quasi-Experiment (Groups 1a and 1b). Table 3 shows the means and standard deviations for both groups (EG and CG) for both measurement points for the TPK test. Shapiro-Wilk tests revealed that all variables (TPK-EG [pre/post] and TPK-CG [pre/post]) were normally distributed after removing two outliers ($W_{EG}(29) = 0.96/0.97$, $p = .37/.55$; $W_{CG}(30) = 0.97/0.96$, $p = .59/.23$). The repeated measures variance analyses in SPSS (with average grade as a covariate) showed a significant interaction between group and measurement points with the two outliers, $F(1, 58) = 6.41$, $p = .01$, $\eta_p^2 = 0.10$, and without them, $F(1, 56) = 5.17$, $p = .03$, $\eta_p^2 = 0.09$. The experimental group showed a significantly higher gain in the TPK test sum score than the control group.

Results of the Pre-Post-Test Intervention Study (Group 3). Table 3 shows the means and standard deviations for both measurement points for the TPK test. Shapiro-Wilk tests revealed that all variables (TPK[pre] and TPK[post]) were normally distributed, $W(57) = 0.99/0.98$, $p = .90/.60$. A repeated-measures variance analysis in SPSS showed a significant effect for the measurement point, $F(1, 56) = 40.82$, $p = .00$, $\eta_p^2 = 0.42$. Pre-service teachers significantly increased their sum score in the TPK test between the pre- and the post-test.

Overall, the results from the two intervention studies are in line with Hypothesis 5, supporting the sensitivity of our test to changes in pre-service teachers' TPK due to instructional interventions.

Table 3

Descriptive Statistics of the TPK Test for Both Intervention Studies

	pre			*post*		
	M	*SD*	Ω	*M*	*SD*	Ω

Group 1 (n = 61)						
EG	17.67	6.35	.73	22.10	8.70	.76
CG	17.81	6.54		18.71	5.04	
Group 3 (n = 57)						
EG	14.12	5.74	.72	18.50	5.84	.68

Note. For the TPK test a sum score was calculated across the 14 items; CG = Control group, EG = Experimental group/intervention group; TPK = technological pedagogical knowledge.

Discussion

Given the lack of psychometrically sound performance-based instruments to measure the technology-related aspects of the TPACK framework, this paper advances the field by presenting research into the creation and validation of an objective knowledge test to measure pre- and in-service teachers' technological pedagogical knowledge (TPK). Study 1 demonstrates the content validity of the test items, whereas Study 2 provides initial evidence for the test's construct validity and instructional sensitivity. The following sections discuss these results.

Interpretation of Findings

This study investigates the content validity of a new knowledge test about technological pedagogical knowledge (TPK) by means of researcher ratings and pre- and in-service teacher ratings (Hypotheses 1 and 2). Overall, the educational researchers and the pre- and in-service teachers describe the test items as relevant to teaching. Based on their review, the test items capture knowledge about important aspects of classroom teaching with digital media.

Most importantly, the four researchers in Study 1 find that every item except for one adequately represents TPK; this item was subsequently removed from the test. As indicated by the mapping in Appendix A, Item 1 stands out as the only item to ask about constraints and difficulties when employing digital media (i.e. computers and the Internet) in the classroom, whereas every other item tests knowledge about the existence, or the capabilities, of various technologies. This could be one explanation as to why researchers rated this item differently than the other items. As some definitions of TPK do include knowledge about the constraints of technology as used in teaching (Cox, 2008), further research should focus on this issue. Specifically, the test should be expanded by adding more items that test knowledge about the constraints of instructional media and that evaluate their impact on model fit. This finding may also have further implications for the definition and understanding of TPK. Knowledge about the capabilities (potential) of technology for teaching, on the one hand, and knowledge about the constraints (difficulties) of technology for teaching, on the other, may be related to either favorable or critical beliefs about teaching with technology (Chai et al., 2013; Voogt et al., 2013).

We hypothesized that our test has a unidimensional structure because all of the items (except for Item 1) are specifically concerned with the existence and capabilities of various technologies for teaching (for definitions of TPK, see Cox, 2008). The results of Study 2 (Hypothesis 3) support the hypothesized unidimensional structure of our test (good model fit) after removing three of the items. The removal of Items 8 (Constructivism) and 14 (WebQuests) was primarily data-driven, not theory-driven. Despite our belief that this does not largely impact the content validity of the test, the test as a whole (including these items) should be investigated further. Due to the smaller sample size at measurement point 2, the test's internal structure could not be examined after the interventions in Groups 1 and 3. Moreover, the internal structure was only tested with a sample of pre-service teachers. Future research should test the model's measurement invariance with a sample of in-service teachers.

The results for convergent and discriminant validity (Hypotheses 4a-c) were not as expected. No significant relationship was found between the test and self-reported TPK. Another study (Drummond & Sweeney, 2017) employing the same validation strategy has shown a small but statistically significant relationship between objective and self-reported measures of technological pedagogical content knowledge (TPACK). We do find a positive relationship between our TPK test and self-reported technological knowledge (TK), however, showing that pre-service teachers, who highly rated their TK, performed better in our test. One possible interpretation of this result could be that TK is an important prerequisite for TPK. Nevertheless, the relationships should generally be interpreted with caution, due to the uncertain validity of self-reported metrics (Maderick, Zhang, Hartley, & Marchand, 2015). This can be explained by the fact that participants with little experience (such as the pre-service teachers in this study in terms of TPK) may lack accurate information for judging their own knowledge (Kuh, 2003). Thus, future research should consider other measures to establish the test's convergent validity. Knowledge tests other than self-report scales, e.g. the one developed by Lachner et al. (2019), are certainly better suited for this purpose, but could not be implemented in this study due to limited testing time.

In terms of Hypothesis 5, the test is sensitive to two instructional interventions on TPK, a rather short and controlled one and a longer, more intensive one. Most importantly, the quasi-experiment (Group 1) shows that increases in the pre-service teachers' sum-score do not simply reflect re-test effects, but can be attributed to the intervention (systematic formation of knowledge), as the control group only shows a small gain in sum-score between the pre- and the post-test.

In sum, these results create a basis from which the test's validity can and should be further studied in the future. The results are particularly encouraging insofar as the test seems well-suited to evaluate the effectiveness of instructional interventions on technology-based teaching. The test shows little overlap with self-reported TPK, however; whether this is due to the test's lack of validity or the self-report scales cannot be fully answered at this point.

Limitations

As the items have a constructed response format, participants' answers must be carefully coded before calculating a test score; this is relatively time-consuming compared to evaluating multiple-choice questions, which are frequently used in educational assessments (Hohensinn & Kubinger, 2009). Thus, it is worth investigating whether the test can be shortened in length by reducing the number of items without restricting the test's content and construct validity. Difficulties related to the test's length (not all participants could be presented with all items) also became apparent during the two studies (see Methods sections).

A further limitation, more inherent to the construct being measured, is the future relevance, usefulness and up-to-dateness of the test items. As mentioned, some items are related to specific technology in teaching, such as interactive whiteboards. However, interactive whiteboards likely will eventually be replaced by another more advanced technology and, thus, our test items on interactive whiteboards might lose their relevance to teaching. There are two different, pragmatic courses of action that can be taken here: either the test adds new items or test items are simply updated, replacing older technologies with similar, more recent ones. From the theoretical perspective, this presents the question of to what extent technological pedagogical knowledge (TPK) should or can be measured independently of specific technologies.

This study provides the first pieces of evidence for the content and construct validity of this test. Further validation is necessary, however. In particular, there is still no evidence for the test's criterion validity. The predictive power of TPK, as measured by our test, for teachers' effective use of technology in their classrooms and for their students' outcomes requires further investigation. Relationships between teachers' knowledge (as measured by knowledge tests) and teachers' classroom behavior, as well as students' achievement, have already been shown in other areas (e.g. Kelcey, Hill, & Chin, 2019; Voss et al., 2011).

Implications and Conclusion

This study provides the first pieces of evidence that a knowledge test can validly measure technological pedagogical knowledge (TPK). Hopefully, this work will trigger further research into the construction of knowledge tests to measure the other technology-related constructs in the TPACK framework, e.g. technological content knowledge (TCK). This would create stronger connections between research into teachers' technology-related knowledge and research into the measurement of content knowledge (CK), pedagogical content knowledge (PCK) and pedagogical knowledge (PK) that adopts a cognitive perspective and has a strong focus on the development of knowledge tests (Depaepe et al., 2013; Krauss et al., 2008). Objective knowledge tests on PK, technological knowledge (TK) and TPK, for example, could be employed in an experimental study similar to that of Tröbst et al. (2018) on PCK to further illuminate the theoretical debate about the formation of and the relationships between the different knowledge areas in the TPACK framework, e.g. integrative vs. transformative perspective (e.g. Angeli et al., 2016). Tröbst et al. (2018) investigated the formation of PCK by providing pre-service teachers with instruction in PK and CK, or PCK or CK only. All three areas of knowledge were assessed in knowledge tests prior to and after the interventions. The study provided evidence for both the integrative and transformative viewpoints to some extent.

Moreover, this study shows that pre-service teachers' self-rated knowledge should be distinguished from that as assessed by a knowledge test, as no correlation between these measures was found. Future studies that use self-report instruments to measure teachers' TPK should consider this. Both aspects (self-reported and tested knowledge) may explain different portions of variance in relevant outcome measures, such as teaching performance or frequency of technology usage. In practice, combining our knowledge test with self-report scales (e.g. by Schmidt et al., 2009) could provide a more comprehensive picture of teachers' technology-related competence in teaching (Drummond & Sweeney, 2017).

This study has several important implications that go beyond measurement issues. Psychometrically sound measures of the TPACK constructs will help researchers to investigate more comprehensively the relevance of these different knowledge areas for student learning and to evaluate trainings that target their development (Lawless & Pellegrino, 2007). Thus, after further validation, this test could be used for different purposes, e.g. evaluating teacher education courses on technology integration to identify effective instructional designs for teaching TPK in a university context. Moreover, this test could help researchers to identify the effects of teachers' TPK on outcome variables such as student achievement and motivation. Last but not least, (pre-service) teachers could use this TPK test to self-reflect on their technology-related knowledge for teaching.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington D.C.: American Educational Research Association.

Angeli, C., & Valanides, N. (2009). Epistemological and methodological issues for the conceptualization, development, and assessment of ICT–TPCK: Advances in technological pedagogical content knowledge (TPCK). *Computers & Education, 52*(1), 154–168. https://doi.org/10.1016/j.compedu.2008.07.006

Angeli, C., Valanides, N., & Christodoulou, A. (2016). Theoretical considerations of technological pedagogical content knowledge. In M. C. Herring, M. J. Koehler, & P. Mishra (Eds.). *Handbook of Technological Pedagogical Content Knowledge (TPACK) for Educators* (2nd ed., pp. 11–32). New York: Routledge.

Brophy, J. E. (1999). *Teaching. Educational practices series (Vol. 1).* Brussels: International Academy of Education & International Bureau of Education. Retrieved from: http://www.ibe.unesco.org/fileadmin/user_upload/archive/Publications/educationalpracticesseriespdf/prac01e.pdf

Buchholtz, N., Leung, F. K., Ding, L., Kaiser, G., Park, K., & Schwarz, B. (2013). Future mathematics teachers' professional knowledge of elementary mathematics from an advanced standpoint. *ZDM, 45*(1), 107–120. https://doi.org/10.1007/s11858-012-0462-6

Chai, C.-S., Koh, J. H.-L., & Tsai, C.-C. (2013). A Review of Technological Pedagogical Content Knowledge. *Educational Technology & Society, 16*(2), 31–51.

Clark, R. E. (Ed.). (2001). *Learning from Media: Arguments, Analysis, and Evidence*. Greenwich, CT: Information Age Publishing.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309–328. https://doi.org/10.1080/15391523.2010.10782553

Cox, S. (2008). *A conceptual analysis of technological pedagogical content knowledge.* Unpublished doctoral dissertation. Provo, UT: Brigham Young University.

Cox S. & Graham C.R. (2009) Diagramming TPACK in practice: Using and elaborated model of the TPACK framework to analyze and depict teacher knowledge. *TechTrends, 53*, 60–69. https://doi.org/10.1007/s11528-009-0327-1

de Jong, T., & Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*(2), 105–113. https://doi.org/10.1207/s15326985ep3102_2

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement, 77*(2), 185-203. https://doi.org/10.1177/0013164416658325

Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education, 34*, 12–25. https://doi.org/10.1016/j.tate.2013.03.001

Drummond, A., & Sweeney, T. (2017). Can an objective measure of technological pedagogical content knowledge (TPACK) supplement existing TPACK measures? *British Journal of Educational Technology*, *48*(4), 928–939. https://doi.org/10.1111/bjet.12473

Educational Testing Service (ETS). (2002). *Digital transformation. A framework for ICT literacy.* Princeton, NJ: ETS.

Graham, C. R. (2011). Theoretical considerations for understanding technological pedagogical content knowledge (TPACK). *Computers & Education*, *57*(3), 1953–1960. https://doi.org/10.1016/j.compedu.2011.04.010

Graham, R. C., Burgoyne, N., Cantrell, P., Smith, L., St Clair, L., & Harris, R. (2009). Measuring the TPACK confidence of inservice science teachers. *TechTrends, 53*(5), 70–79. https://doi.org/10.1007/s11528-009-0328-0

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger, & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion [Test Theory and Construction of Questionnaires]* (pp. 143–171). Berlin, Heidelberg: Springer.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238 – 247. https://doi.org/10.1037/1040-3590.7.3.238

Hofer, M., Grandgenett, N., Harris, J. & Swan, K. (2011). Testing a TPACK-Based Technology Integration Observation Instrument. In M. Koehler & P. Mishra (Eds.), *Proceedings of SITE 2011--Society for Information Technology & Teacher Education International Conference* (pp. 4352–4359). Nashville, Tennessee, USA: Association for the Advancement of Computing in Education (AACE).

Hohensinn, C., & Kubinger, K. D. (2009). On varying item difficulty by changing the response format for a mathematical competence test. *Austrian Journal of Statistics, 38*(4), 231–239. https://doi.org/10.17713/ajs.v38i4.276

ISTE. (2008). National educational technology standards for teachers (NETS-T) 2008. Retrieved from: http://www.iste.org/standards/for-educators

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Reviews of Economics, 6*(1), 801–825. https://doi.org/10.1146/annurev-economics-080213-040845

Kelcey, B., Hill, H. C., & Chin, M. J. (2019): Teacher mathematical knowledge, instructional quality, and student outcomes: a multilevel quantile mediation analysis. *School Effectiveness and School Improvement.* https://doi.org/10.1080/09243453.2019.1570944

Knezek, G. & Christensen, R. (2016). Extending the will, skill, tool model of technology integration: adding pedagogy as a new model construct. *Journal of Computing in Higher Education, 28*(3), 307–325. https://doi.org/10.1007/s12528-016-9120-2

Koehler, M. (2011). *The TPACK Image.* Retrieved from: http://tpack.org

Koehler M. & Mishra P. (2008). Introducing TPCK. In AACTE Committee on Innovation and Technology (Eds.). *Handbook of Technological Pedagogical Content Knowledge (TPCK) for Educators* (pp. 3–29). New York: Routledge.

Koehler, M., & Mishra, P. (2009). What is technological pedagogical content knowledge (TPACK)? *Contemporary Issues in Technology and Teacher Education, 9*(1), 60–70.

König, J. & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften: Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerausbildung. [Pedagogical knowledge of pre-service teachers: Measurement and structure of the results of subject-independent teacher education] *Zeitschrift für Erziehungswissenschaft, 12*(3), 499–527. https://doi.org/10.1007/s11618-009-0085-z

König, J., Kaiser, G. & Felbrich, A. (2012). Spiegelt sich pädagogisches Wissen in den Kompetenzselbsteinschätzungen angehender Lehrkräfte? Zum Zusammenhang von Wissen und Überzeugungen am Ende der Lehrerausbildung. [Is pedagogical knowledge represented in pre-service teachers self-rated competence? On the relationship between knowledge and beliefs at the end of teacher education] *Zeitschrift für Pädagogik, 58*(4), 476–491.

Kozma, R. B. (1991). Learning with Media. *Review of Educational Research, 61*, 179–211. https://doi.org/10.3102/00346543061002179

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716–725. https://doi.org/10.1037/0022-0663.100.3.716

Kuh, G.D. (2003). *The national survey of student engagement: conceptual framework and overview of psychometric properties.* Retrieved from: http://nsse.indiana.edu/pdf/psychometric_framework_2002.pdf

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research, 86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583

Lachner, A., Backfisch, I., & Stürmer, K. (2019). A test-based approach of Modeling and Measuring Technological Pedagogical Knowledge. *Computers & Education, 142*, 103645. https://doi.org/10.1016/j.compedu.2019.103645

Lawless, K. A., Kulikowich, J. M., & Smith, E. V., Jr. (2002, April). *Examining the relationships among knowledge and interest and perceived knowledge and interest.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Lawless, K. A., & Pellegrino, J. W. (2007). Professional development in integrating technology into teaching and learning: Knowns, unknowns, and ways to pursue better questions and answers. *Review of Educational Research, 77*(4), 575–614. https://doi.org/10.3102/0034654307309921

Maderick, J. A., Zhang, S., Hartley, K., & Marchand, G. (2016). Preservice teachers and self-assessing digital competence. *Journal of Educational Computing Research, 54*(3), 326–351. https://doi.org/10.1177%2F0735633115620432

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A new framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Morrison, G. R., & Lowther, D. L. 2002. *Integrating Computer Technology into the Classroom.* Columbus, OH: Merrill Prentice-Hall.

Muthén, L.K., & Muthén, B.O. (2007). *Mplus User's Guide* (5. Ed.). Los Angeles: Muthén & Muthén.

National Research Council. (1999). *Being fluent with information technology literacy. Computer science and telecommunications board commission on physical sciences, mathematics, and applications.* Washington, DC: National Academy Press.

Naumann, A., Musow, S., Aichele, C., Hochweber, J., & Hartig, J. (2019). Instruktionssensitivität von Tests und Items. [Instructional Sensitivity of Tests and Items]. *Zeitschrift für Erziehungswissenschaft, 22*(1), 181–202. https://doi.org/10.1007/s11618-018-0832-0

Neumann, K., Kind, V., & Harms, U. (2018). Probing the amalgam: The relationship between science teachers' content, pedagogical and pedagogical content knowledge. *International Journal of Science Education, 41*(7), 847–861. https://doi.org/10.1080/09500693.2018.1497217

Oser, F., & Baeriswyl, F. (2001). Choreographies of teaching: Bridging instruction to learning. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th ed., pp. 1031–1065). Washington, DC: American Educational Research Association.

Paulhus, D. L. (2017). Socially desirable responding on self-reports. In V. Zeigler-Hill, & T. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences*. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-28099-8_1349-1

Reiser, R. A. (2001). A history of instructional design and technology: Part I: A history of instructional media. *Educational Technology Research and Development, 49*(1), 53–64. https://doi.org/10.1007/BF02504506

Reiser, R. A., & Gagné, R. M. (1983). *Selecting media for instruction.* Englewood Cliffs, NJ: Educational Technology.

Scherer, R., Tondeur, J., & Siddiq, F. (2017). On the quest for validity: Testing the factor structure and measurement invariance of the technology-dimensions in the Technological, Pedagogical, and Content Knowledge (TPACK) model. *Computers & Education, 112*, 1-17. https://doi.org/10.1016/j.compedu.2017.04.012

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23–74.

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological pedagogical content knowledge (TPACK): The development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education*, *42*(2), 123–149. https://doi.org/10.1080/15391523.2009.10782544

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The Test of Technological and Information Literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for Educational Research Online, 5*(2), 139–161.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher. 15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Torff, B. & Tirotta, R. (2010). Interactive whiteboards produce small gains in elementary students' self-reported motivation in mathematics. *Computers & Education, 2*, 379–383. https://doi.org/10.1016/j.compedu.2009.08.019

Tröbst, S., Kleickmann, T., Heinze, A., Bernholt, A., Rink, R., & Kunter, M. (2018). Teacher knowledge experiment: Testing mechanisms underlying the formation of preservice elementary school teachers' pedagogical content knowledge concerning fractions and fractional arithmetic. *Journal of Educational Psychology, 110*(8), 1049–1065. https://doi.org/10.1037/edu0000260

Valtonen, T., Sointu, E., Kukkonen, J., Kontkanen, S., Lambert, M. C., & Mäkitalo-Siegl, K. (2017). TPACK updated to measure pre-service teachers' twenty-first century skills. *Australasian Journal of Educational Technology, 33*(3), 15–31. https://doi.org/10.14742/ajet.3518

Voogt, J., Fisser, P., Pareja Roblin, N., Tondeur, J., & van Braak, J. (2013). Technological pedagogical content knowledge–a review of the literature. *Journal of Computer Assisted Learning*, *29*(2), 109–121. https://doi.org/10.1111/j.1365-2729.2012.00487.x

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing Teacher Candidates' General Pedagogical and Psychological Knowledge: Test Construction and Validation. *Journal of Educational Psychology, 103*(4), 952–969. https://doi.org/10.1037/a0025125

Walczak, S., & Taylor, N. G. (2018). Geography learning in primary school: comparing face-to-face versus tablet-based instruction methods. *Computers & Education, 117*, 188–198. https://doi.org/10.1016/j.compedu.2017.11.001

Yang, Y. C. (1991). The effects of media on motivation and content recall: Comparison of computer-and print-based instruction. *Journal of Educational Technology Systems, 20*(2), 95–105. https://doi.org/10.2190/QAB5-VRX9-38M6-FUY1

Yurdakul, I. K., Odabasi, H. F., Kilicer, K., Coklar, A. N., Birinci, G., & Kurt, A. A. (2012). The development, validity and reliability of TPACK-deep: A technological pedagogical content knowledge scale. *Computers & Education, 58*(3), 964–977. https://doi.org/10.1016/j.compedu.2011.10.0

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology, 16*(1), 93. https://doi.org/10.1186/s12874-016-0200-9

Appendix A: Overview of the Test Items of the TPK Test and Their Classification

No.	Item name	*Mapping of items based on definitions of TPK	Type of technology, media or software application	Teaching practices, methods and processes (see Voss et al., 2011)
1	Internet search (students' difficulties)	constraints	Internet Search engine	student-centered learning maximizing time on task
2	Adaptivity	capabilities support pedagogical goals	e.g. Assistive technology (on tablets) Intelligent Tutorial Systems	adaptive instruction
3	Serious Games	capabilities	Digital Games Game apps (e.g. on tablets)	student-centered learning engaging/motivating students
4	IWB usage	capabilities change of teaching	diverse software applications (e.g. text processing)	teaching methods: e.g. direct instruction
5	Functions	existence capabilities	diverse software applications (e.g. text processing, spreadsheets)	teaching methods, assessment, adaptive instruction
6	LMS usage	capabilities	Web-based software systems	organizing learning and instruction;

		support pedagogical goals support organization	Content Management Systems Communication Tools	adaptive instruction teaching methods: cooperative learning
7	eExam	capabilities	Assessment software	classroom assessment
8	Constructivism	capabilities support pedagogical goals	Several: e.g. Internet, Tablet technology	student-centered learning
9	Homework delivery	existence support organization	Communication tools, Web 2.0 technology	organizing learning and instruction
10	Tablet usage	capabilities pedagogical tool	Tablet technology diverse software applications (e.g. text processing, spreadsheets)	student-centered learning organizing learning and instruction
11	LMS examples	existence	Web-based software systems	organizing learning and instruction teaching methods: cooperative learning
12	Feedback	capabilities	e.g. Assessment software Intelligent Tutorial Systems	formative assessment, feedback
13	Communication	capabilities support pedagogical goals change of teaching	Communication tools; Web 2.0 technology	teaching methods: cooperative learning teaching methods: project-based learning

14	WebQuests	existence components	Internet	teaching methods: problem-based learning
15	Inclusion	existence capabilities support pedagogical goals	Tablet technology; Assistive technology (on tablets)	adaptive instruction
16	ePortfolios	capabilities	Content Management Systems Communication tools Assessment software	classroom assessment
17	IWB periphery devices	existence capabilities support pedagogical goals	IWB; Tablet technology; document camera; wireless slate, audience response systems	student-centered learning

Notes. *Keywords included in definitions on TPK (Cox, 2008): “Existence, components and capabilities of various technologies”; “Pedagogical affordances and constraints of a range of technological tools”; “How technology can support pedagogical goals”; “How teaching and learning change when particular technologies are used”; “Technology as a general pedagogical tool”; IWB = interactive whiteboard; LMS = learning management system; TPK = technological pedagogical knowledge. Note that the item names are derived from the respective item stem and do not contain the full theoretical mapping information.

APPENDIX B: Items 1-17 (Item Stem)

Item No.	Stem
1	*Internet search* You want your students to conduct an Internet search during class. You employ this method so that your students learn how to search, collect and systemize information online. Your students have to present their results afterwards. Which difficulties could possibly arise while your students search online?
2	*Adaptive Instruction* You as a teacher would like to adapt your instruction to the individual learning needs of your students. In what ways may digital media help you to achieve this aim?
3	*Serious Games* You as a teacher would like to employ Serious Games in your class. Serious Games are digital games that are used for educational purposes. Which potentials do Serious Games hold for your students' learning?
4	*Interactive Whiteboards* An interactive whiteboard is an electronic whiteboard that is connected to a projector via a computer. Which advantages (in terms of learning and instruction) do interactive whiteboards have over traditional blackboards?

5 *Functions of digital media in learning and instruction*

Which roles/functions may digital media play/fulfill in learning and instruction? An example of one such role/function is conducting an online search. Thus, digital media may be used to search and collect information on the Internet.

Please list further roles that digital media may play in learning and instruction.

6 *Learning Management Systems (LMS) – usage*

Please come up with ideas in which ways a Learning Management System may support your teaching and your students' learning.

7 *eAssessment*

You would like to conduct the next exam electronically with your students. An electronic exam is an exam that is conducted in specific computer labs in your school. Which advantages do electronic exams have over paper-pencil-based exams?

8 *Constructivism*

In what ways may digital media be employed to realize an instruction that is guided by constructivist principles?

9 *Homework delivery*

Please come up with ideas of how you could deliver the tasks for homework to your students electronically.

10 *Tablet usage*

You have a one-to-one model for tablet use in your classroom (each student has their own device). Please come up with ideas of how tablets may be used effectively for diverse learning tasks in your classroom.

11 *Learning Management Systems – Examples*

Learning Management Systems become increasingly important for individual student learning support. Please name as many examples of Learning Management Systems as you know of.

12 *Computer-based feedback*

Which potential does computer-based feedback hold as compared to feedback provided by you as a teacher for you and your students?

13 *Digital Communication*

You would like your students to work on their own projects in groups of four for several weeks. Your students are required to manage their group work independently and to create a presentation of their results together. How can web-based communication tools support your students' work on their projects?

14 *WebQuests*

	You would like to use a WebQuest in your classroom. An interested colleague asks you what a WebQuest is. Please describe the concept to your colleague.
15	*Inclusion of students with special educational needs* Please list possibilities of how tablets may be used as prostheses in an inclusive classroom.
16	*ePortfolios* You would like your students to submit a personal electronic portfolio at the end of the school year. Portfolios serve students to describe and reflect (on) their learning. What are advantages of electronic portfolios over traditional portfolios?
17	*Interactive Whiteboards – periphery devices* You as a teacher use an interactive whiteboard in your classroom. How can you use periphery devices in combination with the interactive whiteboard to foster student participation?

APPENDIX C: Coding Scheme for Item 5

Item 5: Functions of digital media in learning and instruction

Which roles/functions may digital media play/fulfill in learning and instruction? An example of one such role/function is conducting an online search. Thus, digital media may be used to search and collect information on the Internet.

Please list further roles that digital media may play in learning and instruction.

Answer categories	Code	Points
Not codable: vague/wrong answers: Organizing things eLearning Digital media may be employed so that student learn how to use them	998	0
Media/Technology to illustrate/present/depict information/data Show/Visualize/Present pictures, texts, videos, podcasts, music, animations, graphics (to students) → (teacher-centered)	**1** 1.1	**1** 0,5
Media/Technology as tools for students to create, edit, build, present their ideas/products: e.g. via text editing software, table calculation, presentation software, (technical) drawing, mind-mapping, (graphic) calculators, video production → (student-centered)	**2** 2.1	**1** 0,5
Communication/Cooperation/Collaboration: e.g. via E-Mail, Chat, Wikis, Learning Management Systems	**3** 3.1	**1** 0,5

Exchange of information, experiences, ideas; collective file sharing and editing		
Learning software, Instructional software, Web-based Trainings, Intelligent Tutoring Systems, Simulations, Drill-and-Practice-Software, (Serious) Games, Learning Management Systems (in order to practice, repeat, individualize instruction) e.g. vocabulary training, reading training, math training software	**4** 4.1	**1** 0,5
Documentation and Storage (of information/data) e.g. taking notes while listening to the teacher (on one's laptop/tablet); Taking pictures of/Videotaping experiments/results	**5** 5.1	**1** 0,5
Assessment/Evaluation/Test/Feedback e.g. Online-Quizzes, Surveys, eExams, Voting-Tools, ePortfolios	**6** 6.1	**1** 0,5
As prostheses/artificial replacement (for students with special educational needs) e.g. screenreader, voice control	**7** 7.1	**1** 0,5

Note: One point for the correct umbrella term (bold letters); Half a point for one of the examples of the respective function.

Appendix D: Overview of the Different Samples and Measurement Points

Sample	Name	Number of measurement points	Design	Sample size (total)	Sample size (pre and post data both available)	Gender (% female)	Semester *M* (*SD*)	Average grade *M* (*SD*)
Winter term of 2017-18	Group 1a	2 (pre-post)	EG: Pre-post-intervention	32	30	50	4.90 (1.84)	2.48 (0.55)
Winter term of 2017-18	Group 1b	2 (pre-post)	CG: No intervention	34	31	68	5.16 (2.66)	2.13 (0.62)
Summer term of 2018	Group 2	1 (pre)	Cross-sectional	102	--	74	3.60 (2.25)	2.39 (0.52)
Winter term of 2018-19	Group 3	2 (pre-post)	Pre-post-intervention	87	57	74	4.00 (1.72)	2.27 (0.50)

Notes. "Pre" refers to the beginning of each term; values for gender, semester and average grade in their Abitur (secondary school leaving examination in Germany; range: 1 = very good – 6 = insufficient) refer to participants for whom both pre- and post-test measures were available in Groups 1a/b and 3; CG = Control group, EG = Experimental group.