



Research paper

General pedagogical knowledge among pre-service vocational teachers: Measurement and development during teacher education

Stephan Abele^{a,*}, Diana Vogel-Blaschka^a, Andreas Leon^a, Marlen Beck^b, Bärbel Fürstenau^b

^a Vocational Education, Faculty of Education, Institute of Vocational Education and Vocational Didactics, Technische Universität Dresden, Dresden, 01062, Germany

^b Business Education and Management Training, Faculty of Business and Economics, Technische Universität Dresden, Dresden, 01062, Germany

ARTICLE INFO

Keywords:

Pedagogical knowledge
Educational knowledge
Psychological knowledge
Vocational teachers
Pre-service teachers
TEDS-M

ABSTRACT

The relevance of general pedagogical knowledge (GPK) for vocational teachers is internationally recognised but has not yet been systematically investigated. In this study, the GPK of pre-service vocational teachers is analysed using an established GPK test. Structural equation modelling was used to analyse the test's psychometric quality and GPK development using a cohort-sequential design ($N = 374$). The results indicate that the test measures GPK one-dimensionally and that there is substantial GPK development among pre-service teachers ($d = .68$ – 1.63). Our study suggests that previous studies may have been biased (e.g., because they assumed the test measures GPK multidimensionally).

1. Introduction

What individual characteristics do teachers need to succeed in their profession? Research suggests their success depends on their beliefs, motivational orientations, and self-regulation as well as various types of professional knowledge (e.g., Depaepe & König, 2018; Kunter et al., 2013; Nehls et al., 2020), such as general pedagogical knowledge (GPK). GPK represents explicit knowledge that encompasses general pedagogical principles, including pedagogical facts, concepts, models, and theories. GPK is a broad concept that includes, for example, the principles of effective teaching and learning (Leijen et al., 2022). It has been shown to predict teachers' professional development (Voss et al., 2015), professional success (Gindele & Voss, 2017), well-being (Klusmann et al., 2012), and teaching quality (König & Pflanzl, 2016). GPK is labelled 'general' because it is assumed to apply to all learning content and, therefore, all teachers (König & Blömeke, 2009), including (pre-service) vocational teachers. Vocational teachers prepare students in vocational education and training (VET) for the labour market in certain occupational areas and occupations that do not require a university education (e.g., occupations in business and administration or healthcare). Vocational teachers teach at the upper-secondary and post-secondary levels in vocational schools or other VET institutions. Although national education programmes for vocational teachers differ in the extent to which GPK is taught, it is internationally recognised that vocational teachers

should acquire and possess GPK (OECD, 2021). However, we do not know much about the level, development and measurement of GPK in vocational teacher education or among pre-service vocational teachers.

Measuring GPK and its development in pre-service vocational teachers is relevant for several reasons. First, we must measure GPK to test theories about its determinants and effects, that is, to develop empirically grounded theories about GPK (e.g., about its influence on the quality of lesson planning, teaching quality, and learning outcomes). Second, measuring the development of GPK provides information about the effectiveness of vocational teacher education programmes, enabling evidence-based decisions about which programmes should be refined. The extent of GPK development is one of several indicators of the effectiveness of vocational teacher education (others include, for example, the dropout rate or satisfaction with a given course of study).

This article presents a study measuring the level and development of GPK among pre-service vocational teachers. The aim is to measure GPK based on a conceptual framework and measurement instrument that are suitable for use in international comparative studies. In addition, the instrument should meet essential psychometric standards, including objectivity, validity, and reliability standards.

* Corresponding author.

E-mail addresses: stephan.abele@tu-dresden.de (S. Abele), diana.vogel-blaschka@tu-dresden.de (D. Vogel-Blaschka), andreas.leon@tu-dresden.de (A. Leon), marlen.beck@tu-dresden.de (M. Beck), baerbel.fuerstenau@tu-dresden.de (B. Fürstenau).

<https://doi.org/10.1016/j.tate.2025.105155>

Received 12 September 2024; Received in revised form 28 March 2025; Accepted 9 July 2025

Available online 12 July 2025

0742-051X/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

2. Theoretical framework

2.1. Conceptualisations of general pedagogical knowledge

GPK is one of the concepts introduced by Shulman (1987) to describe the professional knowledge of (pre-service) teachers. Instead of GPK, the terms ‘educational knowledge’ (e.g., Müser, Fleischer, Kunina-Habenicht, & Leutner, 2023) and ‘pedagogical/psychological knowledge’ (e.g. Voss et al., 2011) are sometimes used. GPK is structured across multiple dimensions, including specific principles (e.g., König & Blömeke, 2009; Voss et al., 2011). While the GPK dimensions represent various content areas of GPK (e.g., teaching and assessment), the principles represent the facts, concepts, models, and theories within these areas (e.g., lesson planning and psychometric criteria).

Leijen et al. (2022) presented an international review showing that GPK conceptualisations vary across studies and that there is no consensus regarding the dimensions it encompasses. Research by Kunina-Habenicht et al. (2019) and Müser, Fleischer, and Leutner (2023) was not included in this review but is part of our analysis. A comparison of different GPK conceptualisations reveals that they differ in terms of how broadly GPK is defined. While some studies focus on dimensions of effective teaching and learning, others take a broader perspective, incorporating additional dimensions into their definitions of GPK.

Kunina-Habenicht et al. (2019) defined GPK as a broad construct encompassing the following dimensions: teaching, learning and development, assessment, educational theory, school systems and policy, and the teaching profession. Müser, Fleischer, and Leutner (2023) provided an additional broad conceptualisation of GPK that includes the dimensions of teaching, assessing, educating and innovating. Compared to these studies, Voss et al. (2011) and König and Blömeke (2009) introduced narrower GPK concepts. They focused on the dimensions of effective teaching and learning while neglecting other teacher tasks (e.g., innovating) as well as the principles of the broader educational context (e.g., the school system).

Given these various GPK conceptualisations, those performing empirical studies must decide which they will use when measuring GPK. In our study, we chose to focus on GPK conceptualisations that emphasise the dimensions of effective teaching and learning as knowledge. The rationale for this is that these principles form the core of GPK (Leijen et al., 2022), while the other principles are nation specific (e.g., the school system and policy). Additionally, knowledge of effective teaching and learning principles has been shown to influence both teaching quality and student outcomes (Kunter et al., 2013).

Voss et al. (2011) argued that effective teaching and learning require knowledge of pedagogical principles across five dimensions: teaching methods, classroom management, learning processes, individual characteristics, and classroom assessment. König and Blömeke (2009) proposed five dimensions of GPK: lesson structuring, adaptivity, classroom management, motivation, and assessment. We follow the GPK conceptualisation of König and Blömeke (2009), as it was developed by an international research team and designed to guide the development of a GPK test for international comparative studies (König, Blömeke, Paine, et al., 2011). Though the concept introduced by Voss et al. (2011) overlaps with that of König and Blömeke (Weyers et al., 2024, p. 450), the GPK test associated with König and Blömeke (2009) has been evaluated in international studies, whereas the test linked to the concept created by Voss et al. (2011) has not.

The concept of König and Blömeke (2009) includes both declarative and procedural knowledge. While declarative knowledge refers to whether a (pre-service) teacher knows and understands specific principles, procedural knowledge refers to whether a person can apply these principles in making decisions (König & Blömeke, 2010). This concept includes the following dimensions and principles (König, Blömeke, Doll, et al., 2011, pp. 190–191): (1) *Structuring lessons* includes principles related to lesson planning, curricula, defining learning objectives, structuring the teaching process, and evaluating lessons. (2) *Adaptivity*

refers to the principles of teaching students with heterogeneous learning characteristics. This includes strategies of differentiation, the use of a wide range of teaching methods and the implementation of these methods according to the students’ individual characteristics. (3) The third dimension consists of knowledge regarding classroom management and motivation. *Classroom management* entails the principles one applies to prevent and effectively deal with classroom disruptions and how to establish and apply rules and routines to maximise effective learning time. (4) *Motivation* comprises the principles one applies to motivate individual students/groups in the classroom and basic concepts such as achievement motivation. (5) Finally, knowledge about *assessment* includes principles regarding the functions and types of assessments (e.g., types of summative assessments), assessment criteria (e.g., psychometric criteria, such as reliability), and teachers’ assessment errors (e.g., due to subjectivity).

A key assumption is that GPK represents principles of effective teaching and learning, which are relevant to all pre-service teachers and teachers overall (König & Blömeke, 2009). Empirical research supports this assumption, suggesting that effective teaching is grounded in general principles, which apply across content areas and types of teachers (e.g., Hattie, 2008; Merrienboer & Kirschner, 2018; Merrill, 2002; Praetorius et al., 2018). Therefore, it seems reasonable to assume that a core set of principles for effective teaching applies universally to all teacher education programmes and teachers and that the conceptualisation of GPK can be consistent across groups of teachers, including pre-service vocational teachers. However, this assumption does not imply that pre-service vocational teachers do not require specific principles. Rather, it highlights the fact that GPK reflects the principles common to various teaching contexts and teachers.

2.2. Measurement of general pedagogical knowledge in pre-service vocational teachers

2.2.1. Validity

GPK can be measured either through self-reports (Lyckander, 2021) or standardised tests (Voss et al., 2011). Self-reports primarily capture subjective perceptions of one’s own GPK and, therefore, lack objectivity and validity (Voss et al., 2015). For this reason, we focus on test-based measurements in which pre-service vocational teachers’ responses to GPK questions are evaluated using a coding manual.

Validity is the ‘most fundamental consideration’ in test evaluation and refers to the extent to which empirical evidence and theoretical reasoning support the interpretation of test scores as indicators of the underlying GPK concept (AERA, APA, & NCME, 2014, p. 11). Theoretical validation considers, for example, the extent to which the test content represents the GPK concept. Empirical validation involves analysing the test’s internal structure, the cognitive processes involved in completing GPK test items and the relationship between GPK test results and other measurement results. A comparison of research on various GPK tests indicates that the validity of the TEDS-M (Teacher Education and Development Study in Mathematics) GPK test has been examined extensively, as will be described below (Leijen et al., 2022; Voss et al., 2015).

The TEDS-M test items were developed by an international team of educational scientists and psychologists drawing on the GPK concept proposed by Blömeke and König as well as research findings from general didactics and empirical instructional science (König & Blömeke, 2009). Another international team of experts assessed the item pool for content validity and selected only the most suitable items for inclusion (König & Blömeke, 2009). The TEDS-M GPK test is available in both long and short versions. While the long test consists of 50 tasks and 136 items (König & Blömeke, 2009), the short test comprises 18 tasks and 58 items (König & Blömeke, 2010). Multiple-choice tasks were developed to measure declarative knowledge, while open-ended tasks were designed to measure procedural knowledge (König & Blömeke, 2009). Open-ended tasks present typical problem situations that teachers might

encounter and ask (pre-service) teachers to generate potential actions. In later studies, a reduced number of items has often been used for both the long test (e.g., König, Blömeke, Paine, et al., 2011) and the short test (e.g., Tachtsoglou & König, 2017).

The term ‘internal structure’ refers to the evaluation of whether the TEDS-M test measures distinct GPK dimensions. While measures within a single dimension should be homogeneous, measures of different dimensions should be empirically distinguishable (AERA APANCME, 2014, p. 16). Prior studies assumed that the TEDS-M test assesses five dimensions: structuring lessons, adaptivity, classroom management, motivation, and assessment (König & Blömeke, 2009). However, later studies combined motivation and classroom management into a single dimension, as evidence supported a corresponding four-dimensional structure (König, Blömeke, Paine, et al., 2011). König & Blömeke, 2010 argue that the TEDS-M short test is not suitable for measuring these four dimensions due to the limited number of items per dimension. However, Voss et al. (2011) tested the dimensionality of the test and found evidence for four empirically distinct GPK dimensions, despite each dimension comprising only eight to 12 items, which is comparable to or even fewer than the number of items per dimension on the TEDS-M short test (eight to 15, König & Blömeke, 2010). König and Rothland (2012) used the TEDS-M short test to analyse the effects of various GPK dimensions on lesson planning, implicitly assuming that the scale measures multidimensionally. However, the short test has not yet been explicitly tested to determine its ability to measure four empirically distinct GPK dimensions.

The relationships between the TEDS-M GPK test measures and other test measures have been analysed in several studies (Leijen et al., 2022; Voss et al., 2015). Studies on discriminant validity suggest that the TEDS-M test measures something distinct from tests assessing other aspects of teachers’ professional knowledge, such as content knowledge (König, Blömeke, & Doll, 2011) and pedagogical content knowledge (König et al., 2018b). The convergent validity of the TEDS-M long and short tests was examined by Lauermann and König (2016). They found a very high correlation between the results of the long and short tests ($r = .90$), indicating that interindividual differences in pre-service teachers’ GPK can be measured almost equally well using both versions. Finally, König and Pflanzl (2016) investigated the predictive validity of the TEDS-M test and found that test results predict the performance of VET teachers (e.g., students’ perceptions of effective classroom management).

Based on these validity-related findings, we assume that the TEDS-M test measures can be interpreted as indicators of König and Blömeke’s GPK concept, although the cognitive processes involved in completing the test items have not yet been investigated. In view of the results regarding the convergent validity of the short and long tests, we decided in favour of the short test. The short test requires only 20 min of testing time (König & Blömeke, 2010), which is advantageous in terms of reducing test fatigue and improving the participation rate. However, it remains unclear whether the short test is suitable for measuring the four GPK dimensions (structuring lessons, adaptivity, classroom management/motivation, and assessment). This is important for studies that investigate the dimension-specific development of GPK in vocational teacher education and the dimension-specific effects of GPK on the performance of vocational teachers (e.g., lesson planning, lesson quality, or student outcomes).

2.2.2. Reliability and item quality

Validity can be influenced by the reliability of the TEDS-M short test (AERA, AERAAPA and NCME, 2014, p. 34). Reliability refers to the test’s measurement precision. In the case of a very imprecise measurement, the test results would reflect measurement errors rather than accurately capturing the level and development of GPK. To quantify the measurement errors of the TEDS-M short test, previous studies have calculated the expected *a posteriori* (EAP) reliability (e.g., $re_{EAP} = .75$; Heine et al., 2023; $re_{EAP} = .88$; König, Doll, et al., 2018). EAP reliability is a measure

of internal consistency and is comparable to Cronbach’s alpha (Heine et al., 2023).

Both the validity and reliability depend on the quality of the items, including item difficulty and item discrimination. If the TEDS-M short test items are too easy or too difficult, they are not suitable for measuring the level and development of GPK. Similarly, if the items do not effectively discriminate between (pre-service vocational) teachers with low and high GPK levels, they cannot accurately measure interindividual differences and development in GPK. Studies have reported acceptable item difficulties for the TEDS-M short test across cohorts and measurement points (König & Blömeke, 2010). Tachtsoglou and König (2017) also found that item discrimination values were within an acceptable range, though their analysis focused on average values rather than the discrimination of individual items.

In studies on the reliability and item quality of the TEDS-M short test, either no (e.g., König, 2013; König & Blömeke, 2012; König & Klemenz, 2015) or only a few (König & Blömeke, 2009) pre-service vocational teachers participated. Therefore, not only whether the discrimination of individual items is acceptable but, more importantly, whether the findings regarding the short test’s reliability and item quality can be generalised to pre-service vocational teachers remain an open question.

2.3. Development (trend) of general pedagogical knowledge among pre-service vocational teachers

Following Weyers et al. (2024, p. 451), we hypothesise that GPK develops during vocational teacher education. Evidence for this hypothesis is derived from research using the TEDS-M short test as part of both longitudinal and trend studies. While longitudinal studies analyse GPK development using data obtained at various points in time within a single cohort of a vocational teacher education programme, trend studies examine data obtained from different cohorts at various time-points. In the following, we present evidence on GPK development, beginning with longitudinal studies and distinguishing between those that use the short test and those that use the long test. Evidence derived from studies using the long test is incorporated to provide a comprehensive view of the validity of the hypothesis on GPK development.

Weyers et al. (2024) conducted a longitudinal study using the TEDS-M short test and found that GPK developed significantly over the course of the programme. The long test was used in a longitudinal study to assess GPK development in students at the beginning of their studies and after two years of education: König and Klemenz (2015) found that the two-year difference in study time explained approximately 25 % of the variance in GPK. König (2012) reported an effect size of $d = 1.22$ and concluded that this indicates substantial GPK development.

The TEDS-M short test has been used in various trend studies (e.g., Heine et al., 2023; König, Doll, et al., 2018; Tachtsoglou & König, 2017). In these studies, GPK was measured in two cohorts. The first cohort consisted of pre-service teachers enrolled in a Bachelor’s degree programme for teaching, while the second cohort comprised Master’s students who had already completed a Bachelor’s degree in teacher education. Tachtsoglou and König (2017) and König, Doll, et al. (2018) found that study duration (Bachelor’s versus Master’s students) explained a substantial proportion of the variance in GPK but did not report effect sizes for the development trend. Heine et al. (2023) compared the mean values of Bachelor’s and Master’s students, reported an effect size (d) of .98, and concluded that Master’s students have substantially more GPK than Bachelor’s students. König (2013) used the TEDS-M long test to examine the development trend for GPK in two cohorts of pre-service teachers: one cohort at the beginning of their studies and another that had been studying for two years. He found an effect size (d) of 1.00 (own calculation based on the data provided in the relevant article).

The findings on GPK development are largely consistent regardless of whether measurements were conducted in trend or longitudinal studies or using the TEDS-M short or long test. The studies indicate substantial

GPK growth, with effect sizes (*d*) around 1.00 or greater, suggesting substantial development within just a few semesters. However, these studies have primarily been conducted with pre-service teachers from programmes other than vocational education or with mixed cohorts that include pre-service vocational teachers, alongside other teacher education students. This distinction is important, as research suggests that GPK development depends on the characteristics of the teacher education programme, specifically the learning opportunities offered within the programme, such as school internships (Weyers et al., 2024).

König et al. (2020a,b) examined the impact of school internships on GPK development by measuring GPK with the TEDS-M short test before and after a five-month internship. They assume that GPK would be activated and applied during school internships as pre-service teachers plan, conduct, observe, and reflect on lessons, thereby strengthening and deepening their GPK. Their study revealed a significant increase in GPK, with an effect size of Cohen's *d* = .21. Similarly, Tachtsoglou and König (2017) found that GPK increased during school internships. Their study also revealed that GPK development depends on pedagogical learning opportunities, such as specialised pedagogical courses in teacher education programmes. These findings suggest that school internships as well as the number and content of pedagogical courses play a crucial role in GPK development in vocational teacher education.

In their comparison of GPK development across teacher education programmes, Tachtsoglou and König (2017) found that GPK develops similarly across these programmes. However, König et al. (2008) observed that GPK growth varies depending on the teacher education programme teachers complete. This raises the question of whether the findings on GPK development during various teacher education programmes can be generalised to vocational teacher education programmes and underscores the need for research on GPK development in pre-service vocational teachers.

Scalar invariance is often seen as an important prerequisite for trend studies, particularly for the comparison of mean values between cohorts (Müser, Fleischer, Kunina-Habenicht, & Leutner, 2023; Putnick & Bornstein, 2016; Tachtsoglou & König, 2017; Voss et al., 2011). Scalar invariance can be tested with structural equation modelling and enables the comparison of cohorts based on latent mean values. In addition to the commonly used comparison of manifest mean values, a comparison of latent mean values allows us to compare the mean values of hypothetical constructs and, thereby, differentiate measurement errors. Scalar invariance is a specific form of measurement invariance. In this article, measurement invariance refers to the equivalence of the relevant psychometric properties of the GPK short test and, specifically, whether these properties of the test are equivalent in two cohorts of pre-service vocational teachers.

Tachtsoglou and König (2017) tested the GPK short test to ensure scalar invariance, meaning that the structure, loadings, and intercepts of the GPK indicators (items) are comparable between cohorts. Other types of measurement invariance are configural invariance (same structure across the two cohorts) and metric invariance (same structure and loadings). Tachtsoglou and König (2017) reported that the overall GPK means of the cohorts (novice versus two years of education) differed. They did not report the latent means, the statistical significance of the difference, Cohen's *d*, or other measures of effect size.

2.4. Research questions and hypotheses

In the previous sections, we presented arguments in favour of measuring the GPK of pre-service vocational teachers using the TEDS-M GPK short test (König & Blömeke, 2010). However, certain aspects of the test's psychometric quality, such as item discrimination and dimensionality, have not yet been investigated and reported on. Additionally, the short test has not been evaluated for use with pre-service vocational teachers, whose GPK may differ in psychometrically relevant ways from that of other pre-service teachers. Lastly, there is limited knowledge about the development (trend) of GPK in pre-service vocational

teachers. Gaining such insights would be valuable when evaluating the effectiveness of vocational teacher education and investigating the effects of GPK (e.g., on lesson planning or teaching quality). Against this background, we investigated the following research questions (RQs) and hypotheses:

RQ1: What psychometric quality does the TEDS-M short test achieve in measuring the level of and development trend in the GPK of pre-service vocational teachers? In terms of dimensionality, we assume (H01) that the TEDS-M short test is appropriate for measuring four dimensions (structuring lessons, adaptivity, classroom management/motivation, and assessment). Based on the current state of research, we also assume that its reliability, item discrimination, and item difficulty are acceptable (H02). In terms of measurement invariance, we follow the results of Tachtsoglou and König (2017) and hypothesise that the scalar invariance model represents the data satisfactorily (H03), which allows us to compare the latent mean values of cohorts and analyse the effect size of the GPK development trend.

RQ2: What is the development (trend) of GPK among these students? Based on previous research (e.g., König, 2013), we hypothesise a GPK development (trend) with an effect size (*d*) of approximately 1.00 or greater in our sample (H04).

3. Methods

3.1. Participants

A total of 374 pre-service vocational teachers from a German university took part in the study. These students were studying two vocational subjects related to occupational areas, such as business and administration or health and care, or one vocational subject and one non-vocational subject, such as mathematics or languages. All had taken pedagogical courses (e.g., Fundamentals of Didactics), and 68.4 %, 30.2 %, 0.3 %, and 1.1 % of the students were female, male, gender diverse, or had missing information, respectively. The students were 23.4 years old on average (SD = 4.3 years) and formed two cohorts. The first cohort consisted of students who had completed three or fewer semesters and had not yet completed a school internship, while the second cohort included students who had completed at least four semesters and a school internship (on the relevance of internships to GPK development, see above).

As Table 1 shows, the sample size of Cohort 1 was larger than that of Cohort 2. The differences between the cohorts in terms of age and

Table 1
Characteristics of the cohorts of pre-service vocational teachers.

Characteristics	Cohort 1	Cohort 2
Sample size	211	163
Semester	1.2 (.5)	6.2 (1.5)
Age	22.3 (3.9)	24.7 (4.5)
Gender		
Female	67.3 %	69.9 %
Male	31.3 %	28.8 %
Diverse	0.5 %	0.0 %
Missing	1.0 %	1.2 %
School type		
High school	67.8 %	64.4 %
Vocational high school	21.8 %	23.3 %
Other	10.0 %	9.8 %
Missing	0.5 %	2.5 %
Vocational qualification	37.4 %	31.9 %
Occupational area		
Business administration	11.4 %	39.3 %
Education	32.7 %	12.9 %
Food and housekeeping	9.0 %	11.7 %
Health and care	29.4 %	23.3 %
Technical specialisations	17.1 %	10.4 %
Other	0.4 %	1.2 %
Missing	0 %	1.2 %
Dropout tendency	29.5	28.9

semesters were as expected, with lower values for cohort 1 than cohort 2 ($M_{\text{age}} = 22.4$ and 24.7 ; $M_{\text{Semester}} = 1.2$ and 6.2 , respectively). The differences by gender, school type, and dropout tendency were small, with ‘school type’ referring to the type of school at which the higher education entrance qualification was obtained. None of these differences between the cohorts were significant. In terms of participation rates across occupational areas, the differences between Cohorts 1 and 2 were relatively small, except for business and administration and education, for which the proportions significantly differed between Cohorts 1 and 2.

3.2. Research design and procedure

The data were collected during various semesters using a questionnaire and the TEDS-M short test. The questionnaire contains questions on socio-demographics (e.g., age and gender), the tendency to drop out (Nagy, 2007), epistemic beliefs, beliefs about learning and teaching, and lesson planning. The TEDS-M test was used to measure GPK. The students had 45 min to answer the questions, including 20 min for the TEDS-M test. Initially, the data were collected in face-to-face sessions. Later, this was accomplished online due to social distancing regulations during the COVID-19 pandemic. The data collection began in winter 2021, spanned two years and followed a cohort-sequential design to collect both trend and longitudinal data. However, due to the pandemic and the transition to online data collection, the response rate was much lower than expected, and longitudinal data were available from only 19 students. Two parallel forms of the instrument were used, with different orderings of the questions across versions to prevent students from copying one another’s answers in the face-to-face setting.

3.3. Testing instrument

The TEDS-M short test comprises 18 tasks and 58 items (König & Blömeke, 2010). A task consists of a stimulus (e.g., What is a spiral curriculum?) and one or more items (e.g., A spiral curriculum is defined as ... a didactic concept for the organisation of the learning material [item 1], an interdisciplinary conception of the learning material [item 2], and so on). One task on the short test relates to the transition from primary to secondary school and was excluded because this topic is not relevant to pre-service vocational teachers. This task comprises five items. We also excluded the following task: Pupil Hans repeatedly reports to your school lessons. Is this a case of ‘operant conditioning’? The potential answers are as follows: (1) Yes, Hans displays a behaviour and experiences a positive consequence. (2) Yes, his behaviour is triggered by external influences. (3) No, his behaviour is the result of classical conditioning. (4) No, this behaviour is shown spontaneously and is not learned. We think the information provided in the task is not sufficient to solve it, and therefore, we excluded this task and its one item. The exclusion of the two tasks resulted in a set of 16 tasks (eight open ended and eight multiple choice) and 50 items measuring four dimensions. Multiple-choice items were scored automatically using a computer-based system, with responses being categorised as correct (1) or incorrect (0) based on the coding manual provided by König & Blömeke, 2010. Open-ended items were evaluated by two raters, who applied the same coding manual, resulting in binary codes (incorrect = 0; correct = 1). The interrater reliability for the open-response items was a Cohen’s κ of .90. As Table 2 shows, the four dimensions comprised several tasks, and the tasks consisted of several items. For example, structuring lessons comprised three tasks, and the first task in this dimension consisted of five items.

3.4. Data analysis

We applied classical test theory to analyse item statistics and reliability, while using structural equation modelling to examine dimensionality via confirmatory factor analysis and to examine measurement

Table 2

Dimensions of tasks and items on the TEDS-M short test.

Dimension	Numbering and content of tasks	Number of items
Structuring lessons	1 Structuring the lesson plan	5
	2 Definition of spiral curriculum	6
	3 Questions for analysing lessons	4
Adaptivity	4 Social forms of learning	4
	5 Learning stations	3
	6 Differentiation according to teaching goals	4
Classroom management/ motivation	7 Differentiating teaching methods	1
	8 Pace of teaching	1
	9 Preventing disruptions	3
	10 Learning time	1
	11 Intrinsic and extrinsic motivation	5
	12 Types of attribution	4
Assessment	13 Strategies for motivating students	3
	14 Intellectual giftedness	1
	15 Diagnostic quality criteria	2
	16 Assessment of learning	3

invariance via multi-group modelling (RQ1). Item statistics (i.e., missing data, item difficulty, and item discrimination) and reliability (i.e., Cronbach’s alpha) were calculated using R Statistical Software (V4.4.0; R Development CoreTeam, 2015) with the tidyverse packages (Wickham et al., 2019). Dimensionality and measurement invariance were analysed with MPlus Version 8.10 (Muthén & Muthén, 2017–2017).

To test for dimensionality, we used the full sample and conducted confirmatory factor analyses (CFAs). The fits of the four- and one-dimensional models were compared. The TEDS-M short test consists of tasks and items, so we distinguished between two one-dimensional models. In the first one-dimensional model, the 50 items were used as indicators; in the second one-dimensional model, the 16 tasks were used as indicators. The model parameters were estimated using the full information maximum likelihood (FIML) method. The overall fit of the models was assessed using conventional cut-off criteria (CFI $\geq .95$; RMSEA $\leq .06$; SRMR $\leq .08$; Hu & Bentler, 1999), the Akaike information criterion (AIC; Akaike, 1987), and the Bayesian information criterion (BIC; Schwarz, 1978).

Measurement invariance was tested using a multi-group model with two cohorts. In the two-cohort model, which assumes scalar invariance, the loadings and intercepts of the tasks were constrained to be equal across both cohorts. The model parameters were estimated using the weighted least squares mean and variance adjusted (WLSMV) method. Model fit was evaluated based on the cut-off criteria outlined above, while relative fit was assessed by examining changes in model fit. For example, a test was performed to establish whether the metric invariance model fit the data better than the scalar invariance model. Differences in fit exceeding the cut-off criteria ($\Delta\text{CFI} \geq .010$ and $\Delta\text{RMSEA} \geq .015$) were considered substantial (Chen, 2007). The best-fitting model was used for further analyses.

To assess the development trend for GPK (RQ2), the latent mean values of GPK in the two cohorts were compared. To test for significance, the latent means in both groups were set equal. We then investigated whether equating these parameters substantially worsened the model fit and calculated the size of the development trend (Cohen’s d). Finally, the development of GPK was examined using the longitudinal data obtained from 19 students.

4. Results

4.1. Psychometric quality (RQ1)

The psychometric quality of the short test was investigated in three steps: first, the dimensionality of the TEDS-M short test was tested.

Second, item statistics and reliability were determined. Third, the measurement invariance was analysed.

4.1.1. Dimensionality

To test whether the short test measures four dimensions or one dimension, we compared the fits of the models representing four-dimensional and one-dimensional structures. We began with the one-dimensional model. It can be constructed using the 16 tasks or the 50 items mentioned above. We found a better fit for the task-based model than the item-based model. Specifically, the values of the fit indices and the information criteria for the task-based model (CFI = .829; RMSEA = .04; SRMR = .055; AIC = 15,037; BIC = 15,225) were substantially better than those of the item-based model (CFI = .363; RMSEA = .093; SRMR = .106; AIC = 19,822; BIC = 20,411). It can be assumed that the task-based model fits the data better than the item-based model, as the items of a task share a common stimulus. This common stimulus increases the possibility that the items of the task correlate more strongly with one another than with other items, which impairs the model fit and violates the basic assumption of latent variable models, which is local statistical independence (see the Discussion section). However, it must also be noted that the task-based model still falls short of the target value of .95 for the CFI, while it achieves the target values for the RMSEA and SRMR indicators.

We compared the task-based one-dimensional model to the task-based four-dimensional model. The comparison of the fit indices and the information criteria of the one-dimensional model (CFI = .829; RMSEA = .04; SRMR = .055; AIC = 15,037; BIC = 15,225) with those of the four-dimensional model (CFI = .383; RMSEA = .091; SRMR = .121; AIC = 15,268; BIC = 15,464) showed that the one-dimensional model fit the data better. Furthermore, in the four-dimensional model, the estimated latent correlations between the four dimensions were very high ($r \approx .90$) or, in some cases, perfect ($r = 1.0$), which also indicated that the one-dimensional model is more appropriate than the four-dimensional model. In view of these results, the following analyses concentrate on one-dimensional task-based models.

4.1.2. Task statistics and reliability

The content of the 16 tasks and their factor loadings λ are shown in the two left-hand columns of Table 3. The loading of Task 8 ($\lambda = -.04$) was not significant in the respective models, but the other loadings were significant at the 5 % level. The other columns in the table show statistics for the tasks per cohort (i.e., relative frequency of missing data, task difficulty P , and task discrimination r_{it}).

Some tasks have more than 10 % of their data missing. These tasks were positioned at the end of the short test in one of the two test versions, indicating that some participants did not complete the test. For example, Task 1 was positioned at the end in one version (20 % of missing data in Cohort 1), while Task 15 was positioned at the end in the other version (19 % of missing data in Cohort 1).

The task difficulties (P) ranged from .05 to .68 in Cohort 1 and from .15 to .73 in Cohort 2. If the P values are multiplied by 100, the solution rates are obtained. For example, 8 % of the students in Cohort 1 completed Task 1 correctly. The difficulties of Tasks 13 and 14 did not differ substantially between the two cohorts. However, the comparison of the other difficulty levels showed that the tasks were more difficult for Cohort 1 than for Cohort 2, which corresponds to the assumption of a development trend. Task discrimination (r_{it}) was between $-.04$ and .50 in Cohort 1 and $-.13$ and .42 in Cohort 2. The task discrimination values for Tasks 8, 10, and 14 were less than .30 in both cohorts.

The reliability (α) of the TEDS-M short test was .72 when both cohorts were considered together (.69 for Cohort 1 and .61 for Cohort 2). Task 8 was also somewhat conspicuous here. Removing this task from the test would slightly increase the Cronbach's α value of the test in both cohorts and when analysing the cohorts together. However, we have retained this task, as the increase was only slight, and we had no substantive reason to delete it.

Table 3

Task statistics.

Tasks	λ	Cohort 1			Cohort 2		
		Miss.	P	r_{it}	Miss.	P	r_{it}
1. Structuring the lesson plan	.46	20 %	.08	.33	40 %	.34	.29
2. Definition of spiral curriculum	.35	5 %	.22	.39	3 %	.37	.08
3. Questions for analysing lessons	.55	7 %	.27	.39	10 %	.39	.23
4. Social forms of learning	.49	13 %	.15	.35	11 %	.67	.37
5. Learning stations	.42	0 %	.41	.35	1 %	.51	.08
6. Differentiation according to goals	.36	12 %	.06	.50	10 %	.42	.32
7. Differentiating teaching methods	.36	13 %	.05	.32	17 %	.15	.28
8. Pace of teaching	-.04	0 %	.35	-.04	0 %	.40	-.13
9. Preventing disruptions	.52	7 %	.35	.40	6 %	.43	.31
10. Learning time	.15	7 %	.63	.14	5 %	.71	-.06
11. Intrinsic and extrinsic motivation	.35	3 %	.61	.22	6 %	.71	.42
12. Types of attribution	.25	12 %	.43	.11	14 %	.48	.37
13. Strategies for motivating students	.47	1 %	.31	.34	1 %	.52	.26
14. Intellectual giftedness	.19	8 %	.45	.22	7 %	.44	.19
15. Diagnostic quality criteria	.37	19 %	.68	.37	24 %	.73	.36
16. Assessment of learning	.53	.5 %	.29	.35	1 %	.39	.30

Note. λ : factor loadings, Miss.: missing, P : item difficulty, r_{it} : item discrimination.

4.1.3. Measurement invariance

To assess measurement invariance, the fit of two-cohort models assuming either configural invariance, metric invariance, or scalar invariance was compared. In the model assuming configural invariance, the parameters of the cohorts were estimated freely. In the model assuming metric invariance, the task loadings of the cohorts were set equal. In the model assuming scalar invariance, the task loadings and task intercepts were set equal. Table 4 provides the fit indices for these models.

The fit indices for the model assuming configural invariance met the cut-off criterion for the RMSEA ($\leq .06$) but not that for the CFI ($\geq .95$). Surprisingly, the model assuming metric invariance fit the data better than the less restrictive model assuming configural invariance. The partial scalar invariance model fit the data almost as well as the metric invariance model. The fit of these models did not differ (relevantly) according to the cut-off criteria ($\Delta CFI \geq .010$; $\Delta RMSEA \geq .015$). In the partial scalar invariance model, the loadings of Task 1 (structuring the lesson plan), Task 2 (definition of spiral curriculum), and Task 4 (social forms of learning) were freely estimated. Thus, the model assuming partial scalar invariance is less restrictive than the model assuming scalar invariance. This is addressed in the discussion. The SRMR was slightly exceeded in all the models ($\leq .08$). Given these results, we accepted the model assuming partial scalar invariance and compared the latent means of the cohorts to determine the development trend for GPK.

Table 4

Measurement invariance: Fit indices and comparisons.

Model	χ^2	df	CFI	RMSEA	SRMR	ΔCFI	$\Delta RMSEA$
Configural	232.6	208	.84	.03	.10		
Metric	238.7	224	.90	.02	.11	.06	.01
Partial scalar	248.5	233	.90	.03	.11	.01	.00

4.2. Development trend (RQ2)

The latent GPK mean (M) in the model assuming partial scalar invariance was 0 in Cohort 1 and .68 in Cohort 2, with an average semester number of 1.20 ($SD = .5$) in Cohort 1 and 6.2 ($SD = 1.5$) in Cohort 2. When the latent means of the two cohorts were set equal to one another, the CFI dropped from .90 to .81, suggesting that the GPK level was not the same but, rather, different between the cohorts. Cohen's d for the latent mean difference was .68, and Cohen's d for the differences in the manifest means was .79 ($M_{\text{Cohort1}} = 16.3$, $M_{\text{Cohort2}} = 22.5$, $p < .001$, $t(372) = -7.76$).

We also analysed the development of GPK using longitudinal data obtained from 19 students and at two measurement points. The average number of semesters completed among these students was 3.4 ($SD = 1.9$) at the first measurement point and 5.5 ($SD = 1.8$) at the second measurement point. We found a significant difference in the manifest GPK means between the two measurement points ($M_{t1} = 18.4$, $M_{t2} = 27.3$, $t = -7.1$, $p < .001$, $df = 18$) and an effect size of 1.63 for Cohen's d .

5. Discussion

In this article, we intended to measure the level and development of GPK in pre-service vocational teachers using a conceptualisation and measurement instrument suitable for international comparative studies. Additionally, the instrument needed to meet acceptable psychometric standards, such as objectivity, validity, and reliability. Based on a review of conceptualisations, existing measurement instruments, and findings on GPK development, we selected the TEDS-M short test, created by König & Blömeke, 2010, and identified relevant research gaps. These gaps were addressed by examining the psychometric quality of the short test (RQ1) and analysing the development trend for GPK among pre-service vocational teachers (RQ2).

5.1. Psychometric quality (RQ1)

5.1.1. Dimensionality (H01)

We decided that the TEDS-M short test should be analysed based on tasks rather than items. From a theoretical perspective, the task-based model is preferable because the items of a task refer to the same stimulus, which may lead to the local dependence of items and, thus, biased estimates of psychometric measures (see below). From an empirical perspective, the task-based model demonstrated a better fit to the data than the item-based model, although its fit was not entirely satisfactory, as indicated by a CFI value of .829. Marsh et al. (2004) emphasise that the cut-off criteria proposed by Hu and Bentler (1999) should not be overgeneralised but, rather, supplemented with model comparisons. Based on our analyses, we conclude that the task-based model better fit the data as compared to the item-based model. However, further research is needed to examine the empirical adequacy of the task-based model among pre-service vocational teachers.

We expected the TEDS-M short test to be suitable for measuring four empirically distinct GPK dimensions: structuring lessons, adaptivity, classroom management/motivation, and assessment (H01). However, our findings indicate that the short test is not empirically appropriate for measuring the four GPK dimensions independently of each other. This is in line with König & Blömeke, 2010, who assumed that the TEDS-M short test is not suitable for multidimensionally measuring GPK but provided no evidence for this assumption. Our assumption was based on Voss et al. (2011), whose test measured GPK multidimensionally, despite containing an even smaller number of items than the TEDS-M short test. We argue that the test used by Voss et al. (2011) measured GPK multidimensionally because the items had no common stimulus; that is, they were independent of one another and yielded an appropriate number of indicators per dimension. On the TEDS-M short test, some items have a common stimulus, and tasks that integrate items must be used, which substantially reduces the number of indicators per GPK

dimension. Consequently, our results contradict H01 and led us to test the other hypotheses with task-based models that represented GPK one dimensionally.

5.1.2. Task statistics and reliability (H02)

We excluded two tasks from our analysis, as we found one was not suitable from a content perspective and that the other did not contain the relevant information required to complete the task correctly (see the Methods section). Whether the difficulty (P), discrimination (r_{it}), and test reliability (Cronbach's alpha α) of the remaining 16 tasks are acceptable (H02) can be clarified via a quantitative argument. In such an argument, cut-off criteria (e.g., Cronbach's $\alpha \geq .7$; $P = .05-.95$; $r_{it} \geq .3$) are applied to psychometric results. If the results meet the cutoff criteria, the psychometric quality is assessed as acceptable, and otherwise, it is assessed as unacceptable. However, the cutoff criteria used in such arguments are ultimately arbitrary and can vary considerably, as Taber (2018) shows using the example of Cronbach's alpha in the context of educational assessment. We use a qualitative argument in which relevant aspects (missing data, difficulty, loadings, discriminations, and Cronbach's alpha) are considered, weighed, and used to determine whether the psychometric quality of the short test is acceptable.

Some tasks are associated with a relatively high rate of *missing data*. We believe this is due to the position of these tasks at the end of the test and likely reflects test fatigue rather than poor task quality. The *difficulty* level of the tasks was in line with expectations. The tasks were more difficult for Cohort 1 and easier for Cohort 2, except for the two tasks for which the difficulty was almost the same in the two cohorts (Tasks 13 and 14). The *loadings* and *discriminations* of 13 tasks were inconspicuous. Tasks 8, 10, and 14 differed from the other tasks, as their discrimination was rather low in both cohorts. These tasks were the only ones that contained only a single dichotomous item and, thus, provided less diagnostic information than the other tasks.

The Cronbach's alpha of the short test was in line with expectations in both cohorts ($\alpha \approx .70$). GPK is considered a one-dimensional but heterogeneous construct, as it encompasses four areas of effective teaching and learning. Cronbach's alpha is a measure of internal consistency and is, therefore, lower for heterogeneous constructs than homogeneous constructs, which is why we consider a value of approximately .70 to be acceptable. The heterogeneity of GPK raises the question of whether future studies should examine the reliability of the test using reliability measures other than Cronbach's alpha (e.g., the reliability of parallel forms). It is noteworthy that the reliability levels of the short test in our study were lower than those found in previous studies (e.g., König, Doll, et al., 2018), as we will discuss below.

Overall, we conclude that the short test has acceptable psychometric quality in terms of measuring the overall level of GPK among pre-service vocational teachers. However, psychometric quality was only assessed based on selected criteria (item difficulty) and could be further improved if the three tasks (8, 10, and 14) that consist of only a single dichotomous item were supplemented with additional items. We also suggest using test versions with different item sequences in future studies, as this could prevent the accumulation of missing data in certain tasks.

5.1.3. Measurement invariance (H03)

We hypothesised (H03) that the test measures GPK in both cohorts with scalar invariance (i.e., identical structure of tasks, loadings, and intercepts). We found an acceptable fit for the partial scalar invariance model. This model fit the data as well as or even better than the competing models. However, the intercepts of three tasks had to be freely estimated in this model (Task 1: structuring the lesson plan, Task 2: definition of spiral curriculum, and Task 4: social forms of learning), which means that the intercept differences in these tasks could not be explained by differences in GPK. Based on the data, we cannot determine whether this differential task functioning between the cohorts is systematic or due to chance. However, we see no reason why the tasks are

irrelevant to the construct and believe that they should be included. In such cases, [Robitzsch and Lüdtke \(2023\)](#) recommend intentionally mis-specifying the model, that is, equating the two intercepts of the tasks and comparing the latent means. We conducted such an analysis but did not include it in the Results section, because it revealed a latent mean difference between the cohorts that was comparable to the difference found in the partial scalar invariance model. Our analyses corroborate H04 and provide evidence that the short test is suitable for measuring the GPK development trend in pre-service vocational teachers.

5.2. Development (trend) (RQ2)

Based on previous studies, we hypothesised a GPK development (trend) (d) of approximately 1.00 or greater in our study (H05). We found a development (d) of 1.63 as well as a development trend of .68 (latent means) and .79 (manifest means). We conclude that our results are partially in line with our expectations and the current state of research but stress the need for further studies. Such studies could, for example, investigate whether the differences in effect size (d) are due to methodological reasons (e.g., measurement errors due to the small number of longitudinal data) or substantive reasons (e.g., the teacher samples had completed more or fewer pedagogical courses compared to one another).

5.3. Implications

According to our study, the TEDS-M short test should be analysed on the basis of tasks rather than items, as the items of a task have the same stimulus. Items sharing the same stimulus often creates the problem of local item dependence (LID). 'LID is present if non-zero inter-correlations between items remain after the level of the latent trait or latent traits measured by the items has been controlled for' ([Frey et al., 2016](#), p. 2). In a supplementary analysis, which is not included in the Results section, we found correlations between the items of a task—although the GPK level was controlled for—and, therefore, evidence of LID. LID can lead to the overestimation of reliability ([Frey et al., 2016](#)). In this vein, we argue that the reliability values reported in previous studies (e.g., [Heine et al., 2023](#); [König, Doll, et al., 2018](#)) were higher than those reported in our study because the short test was analysed on the basis of items rather than tasks. EAP reliabilities were often determined in these previous studies. However, these reliabilities are comparable with Cronbach's alpha values ([Heine et al., 2023](#)). We found higher reliabilities when we calculated Cronbach's alpha using items instead of tasks (not reported in the Results section). It is therefore possible that the reliability values for the short test were overestimated in previous studies. However, as mentioned above, we would like to emphasise that the Cronbach's alpha was still acceptable for use in our task-based analysis, although its value was not as high as the values reported in other studies.

Our results also indicate that the short TEDS-M test is not suitable for measuring the four GPK dimensions listed above independently of one another. Thus, the differential effects of the GPK dimensions found by [König and Rothland \(2012\)](#) may be biased because they were measured using the short test. We suggest that the effect of GPK and the effectiveness of vocational teacher education should be analysed at the overall test level, not at the dimension level, when using the short test.

As the psychometric quality of the TEDS-M short test was found to be acceptable, we conclude that it is suitable for use in analysing the determinants and effects of GPK among pre-service vocational teachers. The use of the long test is not necessary for these purposes. [Lauermann and König \(2016\)](#) showed that the results of the TEDS-M short test and the TEDS-M long test correlate very strongly ($r = .90$). This indicates that the use of both tests yields very similar results in terms of inter-individual differences. The investigation of the determinants and effects of GPK is based on inter-individual differences. The TEDS-M short test seems to capture these differences just as well as the long test,

provided that GPK is measured one dimensionally.

We also argue that the short test is suitable for analysing the development trend for GPK one dimensionally and, thus, the effectiveness of vocational teacher education in a one-dimensional manner. Our findings indicate a substantial GPK development trend for pre-service vocational teachers. This development may have multiple explanations. Research suggests that both individual characteristics and learning opportunities, such as school internships and the amount and content of pedagogical courses, can affect GPK development ([Tachtsoglou & König, 2017](#); [Weyers et al., 2024](#)). Assuming that school internships are relevant, we divided our sample into two cohorts. The first cohort had not undertaken a school internship, while the second cohort had. Our findings are aligned with the assumption that school internships influence GPK development, and that GPK development can be interpreted as an indicator of the effectiveness of school internships and given vocational teacher education programmes. However, we did not test these assumptions or investigate whether GPK development is influenced by individual characteristics, study-related factors, other factors, or a combination thereof. Moreover, the results of studies that measure the GPK development trend to assess the effectiveness of vocational teacher education must be interpreted with caution, and it is important to avoid overgeneralisations. While such studies can provide insights into effectiveness in terms of GPK, it is important to recognise that GPK is only one of several indicators of the overall effectiveness of vocational teacher education programmes.

5.4. Limitations and future research

Our conclusions regarding dimensionality and measurement invariance are based on model comparisons. The selected models fit the data as well as or better than the alternative models. However, as in [Voss et al. \(2011\)](#), the models we selected did not fully meet the cutoff criteria proposed by [Hu and Bentler \(1999\)](#). The reasons for this and its implications for measuring the level and development trend of GPK using the TEDS-M short test should be explored in future research. Future studies could investigate whether model modifications improve the model fit and are also justifiable from a content perspective.

The database used for our study was small. GPK was measured in 374 pre-service vocational teachers at one German university. Nevertheless, we see no reason why our implications should not also apply to the measurement of GPK among pre-service vocational teachers at other German universities and even in other countries. The fact that the development and development trend we found were different but comparable to those found in earlier studies (e.g., [Heine et al., 2023](#)) indicates that our results may be representative of vocational teacher education programmes at other German universities. However, this must be investigated further, as we have not considered how the curricula of these programmes may differ. In addition, the transferability of the development trend results to other countries is uncertain, as the content of vocational teacher education programmes varies greatly from country to country. In Australia, for example, GPK is recognised as important but does not play a major role in vocational teacher education ([Smith & Tuck, 2023](#)).

The TEDS-M short test captures principles of effective teaching, including relevant aspects of learning. According to [Leijen et al. \(2022\)](#), these principles form the core of GPK, although GPK is often conceptualised more broadly. In our view, it would be more accurate to state that the short test measures the teaching knowledge of pre-service vocational teachers, rather than their general pedagogical knowledge.

Additionally, the test includes open-response items. If tests consisting solely of multiple-choice items, such as BilWiss 2.0 ([Kunina-Habenicht et al., 2020](#)), proved equally suitable for use in measuring teaching knowledge, this would represent a step towards the more empirically sound and time-efficient measurement of pre-service vocational teachers' teaching knowledge. Further studies are required to investigate this topic.

Building on previous research (König & Blömeke, 2009), we assumed that the conceptualisation of GPK and the corresponding test are consistent and applicable across various groups of teachers, including pre-service vocational teachers, as they encompass the fundamental principles of teaching and learning, which are relevant to all educators and teacher education programmes. This assumption is supported by our findings, as we identified only one test task, which was related to the transition from primary to secondary school, that does not align with vocational teacher education. However, this task could also be used to raise the broader question of which teaching and learning principles are universally relevant for all teachers and teacher education programmes and which are specific to pre-service vocational teachers.

While we did not address this question in our study, we believe the answer depends on the level of specificity one considers. On a general level, it seems appropriate to identify and measure principles that enhance effective, efficient, and engaging teaching and learning for all teachers (e.g., Hattie, 2008; Merriënboer & Kirschner, 2018; Merrill, 2002; Praetorius et al., 2018). On a more specific level, certain principles may be particularly relevant or irrelevant for pre-service vocational teachers (e.g., the principle of connecting learning to working life, as proposed by Lyckander (2021), or principles related to the transition from primary to secondary school). Further research is needed to explore this topic and its implications for the conceptualisation and measurement of GPK levels and development among pre-service vocational teachers.

5.5. Conclusion

Despite these limitations, we conclude that the level and development (trend) of GPK in pre-service vocational teachers can be measured with the TEDS-M short test with acceptable psychometric quality when the measurement is task based rather than item based and the goal is to measure GPK one dimensionally. We hope our promising results on the development (trend) of GPK will encourage researchers to further investigate the effectiveness of vocational teacher education programmes. Furthermore, we encourage researchers to use our findings to investigate the determinants and effects of GPK and thus develop empirically grounded theories of GPK. Such theories and evidence regarding effectiveness will enable vocational teacher training programmes to be improved on the basis of both sound understanding and evidence.

CRedit authorship contribution statement

Stephan Abele: Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Diana Vogel-Blaschka:** Writing – review & editing, Methodology, Formal analysis. **Andreas Leon:** Writing – review & editing, Conceptualization, Formal analysis. **Marlen Beck:** Writing – review & editing, Investigation. **Bärbel Fürstenau:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of generative AI in scientific writing

During the preparation of this work, the authors used DeepL and ChatGPT to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgments

We sincerely thank the pre-service teachers who participated in the study, as well as Stefan Hartmann and the reviewers for their valuable feedback on earlier versions of the manuscript. We also appreciate Johannes König's support in providing information on publications related to the TEDS-M GPK (short) test. This work was supported by Federal Ministry of Education and Research, Germany [grant number 01JA2022].

Data availability

Data will be made available on request.

References

- AERA, APA, NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3), 317–332. <https://doi.org/10.1007/BF02294359>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Depaepe, F., & König, J. (2018). General pedagogical knowledge, self-efficacy and instructional practice: Disentangling their relationship in pre-service teacher education. *Teaching and Teacher Education*, 69, 177–190. <https://doi.org/10.1016/j.tate.2017.10.003>
- Frey, A., Seitz, N.-N., & Brandt, S. (2016). Testlet-based multidimensional adaptive testing. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.01758>. Article 1758.
- Gindele, V., & Voss, T. (2017). Pädagogisch-psychologisches Wissen: Zusammenhänge mit Indikatoren des beruflichen Erfolgs angehender Lehrkräfte [Pedagogical-psychological knowledge: Correlations with indicators of prospective teachers' professional success]. *Zeitschrift für Bildungsforschung*, 7(3), 255–272. <https://doi.org/10.1007/s35834-017-0192-5>
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Taylor & Francis Group. <http://ebookcentral.proquest.com/lib/slub/detail.action?docID=367685>.
- Heine, S., Krepf, M., Jäger-Biela, D. J., Gerhard, K., Stollenwerk, R., & König, J. (2023). Preservice teachers' professional knowledge for ICT integration in the classroom: Analysing its structure and its link to teacher education. *Education and Information Technologies*, 29, 11043–11075. <https://doi.org/10.1007/s10639-023-12212-7>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Klusmann, U., Kunter, M., Voss, T., & Baumert, J. (2012). Berufliche Beanspruchung angehender Lehrkräfte: Die Effekte von Persönlichkeit, pädagogischer Vorerfahrung und professioneller Kompetenz [Occupational stress of prospective teachers: The effects of personality, previous teaching experience and professional competence]. *Zeitschrift für Pädagogische Psychologie*, 26(4), 275–290. <https://doi.org/10.1024/1010-0652/a000078>
- König, J. (2012). Die Entwicklung von pädagogischem Unterrichtswissen: Theoretischer Rahmen, Testinstrument, Skalierung und Ergebnisse [The development of pedagogical teaching knowledge: Theoretical framework, test instrument, scaling and results]. In J. König, & A. Seifert (Eds.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerbildung* (pp. 143–182). Waxmann. http://www.pedocs.de/volltexte/2020/21029/pdf/Koenig_Seifert_2012_Lehramtsstudierende_erwerben.pdf.
- König, J. (2013). First comes the theory, then the practice? On the acquisition of general pedagogical knowledge during initial teacher education. *International Journal of Science and Mathematics Education*, 11(4), 999–1028. <https://doi.org/10.1007/s10763-013-9420-1>
- König, J., & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften: Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerbildung [Pedagogical knowledge of future teachers. Capturing and structuring the results of cross-subject teacher education]. *Zeitschrift für Erziehungswissenschaft*, 12(3), 499–527. <https://doi.org/10.1007/s11618-009-0085-z>
- König, J., & Blömeke, S. (2010). *Pädagogisches Unterrichtswissen (PUW): Dokumentation der Kurzfassung des TEDS-M Testinstruments zur Kompetenzmessung in der ersten Phase der Lehrerbildung [Pedagogical teaching knowledge (PUW): Documentation of the short version of the TEDS-M test instrument for measuring competence in the first phase of teacher training]*. Humboldt-Universität zu Berlin.
- König, J., & Blömeke, S. (2012). Future teachers' general pedagogical knowledge from a comparative perspective: Does school experience matter? *ZDM*, 44(3), 341–354. <https://doi.org/10.1007/s11858-012-0394-1>
- König, J., Blömeke, S., & Doll, J. (2011a). Pädagogisches Wissen von Deutsch-, Englisch- und Mathematiklehramtsstudenten. In S. Blömeke, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold, K. Schwippert, & H. Willenberg (Eds.), *Kompetenzen von*

- Lehramtstudierenden in gering strukturierten Domänen. *Erste Ergebnisse aus TEDS-LT* (pp. 135–157). Waxmann.
- König, J., Blömeke, S., Paine, L., Schmidt, W. H., & Hsieh, F.-J. (2011b). General pedagogical knowledge of future middle school teachers: On the complex ecology of teacher education in the United States, Germany, and Taiwan. *Journal of Teacher Education*, 62(2), 188–201. <https://doi.org/10.1177/0022487110388664>
- König, J., Bremerich-Vos, A., Buchholtz, C., & Glutsch, N. (2020a). General pedagogical knowledge, pedagogical adaptivity in written lesson plans, and instructional practice among preservice teachers. *Journal of Curriculum Studies*, 52(6), 800–822. <https://doi.org/10.1080/00220272.2020.1752804>
- König, J., Darge, K., & Kramer, C. (2020b). Kompetenzentwicklung im Praxissemester: Zur Bedeutung schulpraktischer Lerngelegenheiten auf den Erwerb von pädagogischem Wissen bei Lehramtsstudierenden [Competence development during long-term practicum: On the significance of practical opportunities to learn on the acquisition of pedagogical knowledge among student teachers]. In I. Ulrich, & A. Gröschner (Eds.), *Praxissemester im Lehramtsstudium in Deutschland: Wirkungen auf Studierende* (Vol. 9, pp. 67–95). Springer VS. https://doi.org/10.1007/978-3-658-24209-1_2
- König, J., Doll, J., Buchholtz, N., Förster, S., Kaspar, K., Rühl, A.-M., Strauß, S., Bremerich-Vos, A., Fladung, I., & Kaiser, G. (2018a). Pädagogisches Wissen versus fachdidaktisches Wissen?: Struktur des professionellen Wissens bei angehenden Deutsch-, Englisch- und Mathematiklehrkräften im Studium [General pedagogical knowledge versus pedagogical content knowledge? The structure of professional knowledge in pre-service teachers of German, English, and Mathematics at university]. *Zeitschrift für Erziehungswissenschaft*, 21(3), 1–38. <https://doi.org/10.1007/s11618-017-0765-z>
- König, J., & Klemenz, S. (2015). Der Erwerb von pädagogischem Wissen bei angehenden Lehrkräften in unterschiedlichen Ausbildungskontexten: Zur Wirksamkeit der Lehrerausbildung in Deutschland und Österreich [The acquisition of pedagogical knowledge by future teachers from a comparative perspective: On the effectiveness of teacher education in Germany and Austria]. *Zeitschrift für Erziehungswissenschaft*, 18(2), 247–277. <https://doi.org/10.1007/s11618-015-0623-9>
- König, J., Peek, R., & Blömeke, S. (2008). Zum Erwerb von pädagogischem Wissen in der universitären Ausbildung: Unterscheiden sich Studierende verschiedener Lehramter und Kohorten? *Lehrerbildung auf dem Prüfstand*, 1(2), 664–682.
- König, J., Peek, R., & Blömeke, S. (2018b). Zum Erwerb von pädagogischem Wissen in der universitären Ausbildung: Unterscheiden sich Studierende verschiedener Lehramter und Kohorten? [Acquisition of pedagogical knowledge during the first phase of teacher education: Are there differences between students of different cohorts and in different teacher-education programs?]. *Lehrerbildung auf dem Prüfstand*, 1(2), 664–682. <https://doi.org/10.25656/01:14692>
- König, J., & Pflanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *European Journal of Teacher Education*, 39(4), 419–436. <https://doi.org/10.1080/02619768.2016.1214128>
- König, J., & Rothland, M. (2012). Motivations for choosing teaching as a career: Effects on general pedagogical knowledge during initial teacher education. *Asia-Pacific Journal of Teacher Education*, 40(3), 289–315. <https://doi.org/10.1080/1359866X.2012.700045>
- Kunina-Habenicht, O., Maurer, C., Schulze-Stocker, F., Wolf, K., Hein, N., Leutner, D., Seidel, T., & Kunter, M. (2019). Zur curricularen Validität des BilWiss 2.0-Tests zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften [On the curricular validity of the BilWiss-2.0-Test for the assessment of (pre-service) teachers' educational knowledge]. *Zeitschrift für Pädagogik*, 65(4), 542–556. <https://doi.org/10.25656/01:23994>
- Kunina-Habenicht, O., Maurer, C., Wolf, K., Holzberger, D., Schmidt, M., Dicke, T., Teuber, Z., Koc-Januchta, M., Lohse-Bossenz, H., Leutner, D., Seidel, T., & Kunter, M. (2020). Der BilWiss-2.0-Test: Ein revidierter Test zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften [The BilWiss 2.0 test: A revised test to assess the educational science knowledge of (prospective) teachers]. *Diagnostica*, 66(2), 80–92. <https://doi.org/10.1026/0012-1924/a000238>
- 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>
- Lauermann, F., & König, J. (2016). Teachers' professional competence and wellbeing: Understanding the links between general pedagogical knowledge, self-efficacy and burnout. *Learning and Instruction*, 45, 9–19. <https://doi.org/10.1016/j.learninstruc.2016.06.006>
- Leijen, Å., Malva, L., Pedaste, M., & Mikser, R. (2022). What constitutes teachers' general pedagogical knowledge and how it can be assessed: A literature review. *Teachers and Teaching*, 28(2), 206–225. <https://doi.org/10.1080/13540602.2022.2062710>
- Lyckander, R. H. (2021). Exploring vocational teacher preparation in Norway: A study of dimensions and differences in vocational teacher learning. *Journal of Vocational Education and Training*, 1–23. <https://doi.org/10.1080/13636820.2021.2007985>
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2
- Merriënboer, J. J. G. van, & Kirschner, P. A. (2018). Ten steps to complex learning: A systematic approach to four-component instructional design. *WEB 137a* (3rd ed.). Routledge, Taylor & Francis Group. vor Ort <https://katalog.slub-dresden.de/id/0-894372882>
- Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research & Development*, 50(3), 43–59. <https://doi.org/10.1007/BF02505024>
- Müser, S., Fleischer, J., Kunina-Habenicht, O., & Leutner, D. (2023a). The ESBW short scale: A test for assessing teachers' standards-based educational knowledge. *European Journal of Psychological Assessment*, 40. <https://doi.org/10.1027/1015-5759/a000763>
- Müser, S., Fleischer, J., & Leutner, D. (2023b). Entwicklung und Validierung des Essener Tests zur Erfassung des standardorientierten bildungswissenschaftlichen Wissens (ESBW) [Development and validation of the Essen Test for the Assessment of Standardised Educational Science Knowledge (ESBW)]. *Diagnostica*, 69(2), 62–73. <https://doi.org/10.1026/0012-1924/a000303>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical analysis with latent variables: User's guide (Version 8)*. Muthén & Muthén.
- Nagy, G. (2007). *Berufliche Interessen, kognitive und fachgebundene Kompetenzen: Ihre Bedeutung für die Studienfachwahl und die Bewährung im Studium* [Professional interests, cognitive and subject-related skills: Their importance for choosing a subject and proving oneself during studies] [Freie Universität Berlin] <https://refubium.fu-berlin.de/handle/fub188/10012>
- Nehls, C., König, J., Kaiser, G., & Blömeke, S. (2020). Profiles of teachers' general pedagogical knowledge: Nature, causes and effects on beliefs and instructional quality. *ZDM*, 52(2), 343–357. <https://doi.org/10.1007/s11858-019-01102-3>
- OECD. (2021). *Teachers and leaders in vocational education and training*. OECD Publishing. https://read.oecd-ilibrary.org/education/teachers-and-leaders-in-vocational-education-and-training_59d4fbb1-en
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM*, 50(3), 407–426. <https://doi.org/10.1007/s11858-018-0918-4>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- R Development CoreTeam. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org>
- Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 30(6), 859–870. <https://doi.org/10.1080/10705511.2023.2191292>
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Shulman, L. S. (1987). *Knowledge and teaching. Foundations of the new reform*. *Harvard Educational Review*, 57(1), 1–22.
- Smith, E., & Tuck, J. (2023). Do the qualifications of vocational teachers make a difference to their teaching? *Research in Post-Compulsory Education*, 28(1), 1–25. <https://doi.org/10.1080/13596748.2023.2166690>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tachtsoglou, S., & König, J. (2017). Der Einfluss universitärer Lerngelegenheiten auf das pädagogische Wissen von Lehramtsstudierenden [The influence of university learning opportunities on the pedagogical knowledge of pre-service teachers]. *Zeitschrift für Bildungsforschung*, 7(3), 291–310. <https://doi.org/10.1007/s35834-017-0199-y>
- Voss, T., Kunina-Habenicht, O., Hoehne, V., & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde [Teachers' pedagogical knowledge: Empirical approaches and findings]. *Zeitschrift für Erziehungswissenschaft*, 18(2), 187–223. <https://doi.org/10.1007/s11618-015-0626-6>
- Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, 103(4), 952–969. <https://doi.org/10.1037/a0025125>
- Weyers, J., Ligtvoet, R., & König, J. (2024). How does pre-service teachers' general pedagogical knowledge develop during university teacher education? Examining the impact of learning opportunities and entry characteristics over five time points using longitudinal models. *Journal of Curriculum Studies*, 56(4), 448–467. <https://doi.org/10.1080/00220272.2024.2355923>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>