Ger J Exerc Sport Res 2022 · 52:11–23
https://doi.org/10.1007/s12662-021-00735-5
Received: 27 April 2020
Accepted: 5 July 2021
Published online: 10 August 2021

Carmen Volk[1] · **Stephanie Rosenstiel**[1] · **Yolanda Demetriou**[2] · **Gorden Sudeck**[1] · **Ansgar Thiel**[1] · **Wolfgang Wagner**[3] · **Oliver Höner**[1]

[1] Institute of Sports Science, University of Tübingen, Tübingen, Germany

[2] Department of Sport and Health Sciences, Technical University of Munich, Munich, Germany

[3] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

Health-related fitness knowledge in adolescence: evaluation of a new test considering different psychometric approaches (CTT and IRT)

Supplementary Information

The online version of this article (https://doi.org/10.1007/s12662-021-00735-5) contains supplementary material, which is available to authorized users.

Introduction

The promotion of pupils' health is an established goal of physical education (PE). Developing and maintaining a healthy, physically active lifestyle represents a main objective of PE curricula around the world (e.g., Society of Health and Physical Educators, 2014). In this context, the acquisition of specific knowledge is assumed to have a supportive effect both in competence-based PE curricula in Germany and in most Anglo-Saxon PE curricula, which are based on the concept of physical literacy (Cale & Harris, 2018; Wagner, 2016). Physical literacy is defined as "the *knowledge* [emphasis added], skills and confidence to enjoy a lifetime of healthful physical activity" (Society of Health and Physical Educators, 2014, p. 11) or, more broadly, the "motivation, confidence, physical competence, *knowledge and understanding* [emphasis added] to value and take responsibility for engagement in physical activities for life" (Tremblay et al., 2018, p. 16). Since competence in German PE curricula is not consistently defined, it can also be understood as a combination of knowledge, skills, abilities, and motivational aspects that enable individuals to meet the complex demands of a specific domain (Klieme et al., 2003; Kurz, 2008; Weinert, 2001).

As the acquisition of knowledge on how to lead a healthy, physically active lifestyle is a proclaimed learning objective of PE, it is of interest for PE research to assess—via cross-sectional and interventional studies—(1) pupils' level of knowledge and whether this knowledge level differs, (2) how pupils' knowledge level develops due to PE, and (3) the actual role of this knowledge with regard to physical activity (PA) behavior. For researchers to develop reliable and valid findings to answers these questions, the use of a sound knowledge test is mandatory. However, the current literature indicates three fundamental challenges regarding existing knowledge tests applied in PE research in the context of health (Demetriou, Sudeck, Thiel, & Höner, 2015; Keating et al., 2009).

First, there is no consistent, comprehensive definition of the kind of knowledge that is important for a healthy, physically active lifestyle. Edwards, Bryant, Keegan, Morgan, and Jones (2017) reviewed the description of the cognitive domain of physical literacy and identified two topics of knowledge: knowledge and understanding of activities (e.g., sports' rules, values, and traditions) and of a healthy and active lifestyle. Accordingly, in the context of physical literacy, the term *knowledge* is not exclusive to health-related aspects of PA (e.g., knowledge of sports' rules). In North American PE literature, health-related fitness knowledge (HRFK) is considered the foundation for a healthy, physically active lifestyle (Keating et al., 2009; Zhu, Safrit, & Cohen, 1999). This term is also suitable to the German PE context, as examples of HRFK can be seen in German PE curricula (e.g., knowledge of how to enhance health-related fitness; Ministry of Education and Cultural Affairs, Youth and Sports of Baden-Württemberg, 2016; Wagner, 2016). However, a generally accepted definition of HRFK is still missing. While some authors have incorporated knowledge on how to enhance health-related physical fitness or knowledge about physiological responses to PA (Kulinna & Zhu, 2001), others have contributed knowledge about the effects of PA on health or nutrition (Zhu et al., 1999). Moreover, different labels for HRFK are in use (e.g., exercise knowledge or knowledge of physical fitness; Demetriou et al., 2015; Keating et al., 2009). These observations are reflected in the HRFK tests used in research. In their review of the ef-

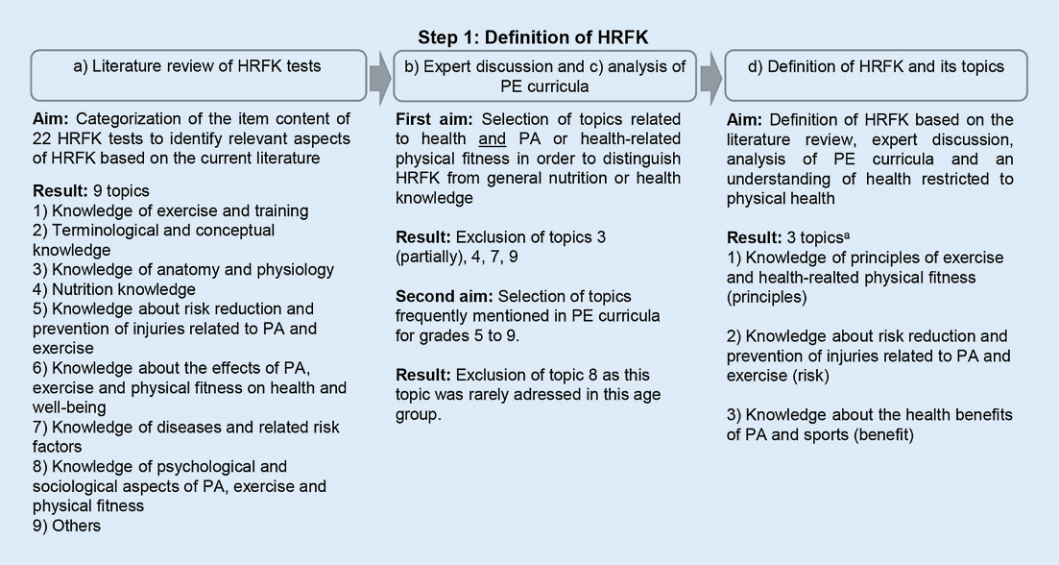


Fig. 1 ◀ Different phases to develop a comprehensive definition of health-related fitness knowledge. *PA* physical activity, *PE* physical education, *HRFK* health-related fitness knowledge. [a]Terminological and anatomical knowledge were not considered separate topics of the HRFK test but were included in the three newly defined topics to develop items that did not simply ask for the definition of a term

fects of PA intervention studies on HRFK in PE, Demetriou et al. (2015) concluded that a notable variety of HRFK tests are used in PE intervention studies. These tests differ in their assessment of HRFK (e.g., number of items or content of the test) and are often not based on an explicit definition of knowledge and HRFK (for similar observations, see Keating et al., 2009).

Second, reviews on HRFK research in PE have identified a lack of reliability and validity in terms of the psychometric properties of the tests applied (Demetriou et al., 2015; Keating et al., 2009). However, to date, there is no gold standard for the validation of HRFK tests, which poses a challenge for its evaluation (Demetriou et al., 2015). In addition, the validity of knowledge tests that examine the general relationship between knowledge and behavior is a topic of debate. According to Ajzen, Joyce, Sheikh, and Cote (2011), knowledge tests often assess a person's attitude rather than knowledge. Furthermore, items often deal with general concepts rather than a specific health behavior. Finally, items are often factual, and it remains unclear whether the assessed knowledge is actually supportive of behavior. These aspects have to be considered when developing a test to ensure the test's validity, depending on the understanding of the term *knowledge*.

Third, the majority of studies reviewed above refer solely to the criteria of classical test theory (CTT). This also applies to the physical literacy knowledge questionnaire that was recently developed and evaluated for children ages 8–12 years based on the Canadian PE and health education curricula (Longmuir, Woodruff, Boyer, Lloyd, & Tremblay, 2018). CTT is an established approach for test construction in sports science and psychology; however, from a methodological perspective, CTT has its shortcomings. For example, test and item statistics (e.g., item difficulty and reliability) are sample dependent, an individual's test score is influenced by the test's characteristics (Hambleton & Jones, 1993), and statistical analysis in the context of CTT requires continuous variables (Bühner, 2011). Thus, in educational research, item response theory (IRT) is often used in addition to CTT for test development and evaluation (e.g., Organisation for Economic Co-operation and Development [OECD], 2017). IRT models describe the relationship between an individual's response to an item, the individual's ability, and the characteristics of the item (e.g., item difficulty and item discrimination parameters). Moreover, these models are useful for dichotomous and polytomous variables. The major advantage of IRT is that model parameters are independent of the study sample, and standard errors can be calculated separately for each person's ability. However, IRT models rely on assumptions of dimensionality of the test and conditional independence of the test items (de Ayala, 2009). These assumptions can be investigated through various analyses that require large sample sizes (de Ayala, 2009; Hambleton & Jones, 1993). Regarding HRFK tests in North America, only the FitSmart Test developed for high school pupils has been evaluated using an IRT model (Zhu et al., 1999). In the European context, Töpfer (2019) systematically developed and scaled a test on sport-related health competence that included aspects of HRFK for German seventh to 10th graders based on IRT models. However, to our knowledge, no test for HRFK in the ongoing German discussion has been based on a substantiated test development and evaluation process using IRT.

Considering these three challenges — the need for a comprehensive definition, the lack of validity and reliability, and the applied test theory—we aimed to develop and evaluate an HRFK test for ninth graders attending secondary schools in Germany (Gymnasium, a type of school that provides learners with general university entry qualifications) in order to extend the options for a sound assessment of pupils' HRFK in the field of PE research on pupils' current level of HRFK (cross-sectional studies) and the development of HRFK due to PE (interventional studies). This article describes the development of a preliminary HRFK test based on a systematic definition of HRFK.

Subsequently, the article presents the results of two empirical studies that evaluate the psychometric properties of items and tests based on the established quality criteria of CTT and IRT. Study 1 aims to analyze item difficulty and discrimination, test reliability, and construct validity of the preliminary version of the HRFK test in a sample of ninth graders from different educational levels and types of PE in order to select appropriate items for the second study. Study 2 investigates the selected items' properties in the target population and examines the dimensionality and reliability of the test, aiming to scale the final HRFK test version for ninth graders of the same educational level.

Test development

The HRFK test was developed in three steps: (1) defining HRFK, (2) developing items, and (3) conducting pilot studies. In the first step (**Fig. 1**), HRFK tests were identified based on a keyword-based literature review by Demetriou et al. (2015). Subsequently, HRFK test items were analyzed and categorized with regard to their topics (Phase a of Step 1). The topics extracted from the literature review were discussed with experts (Phase b) and compared to German PE curricula (Phase c) to select topics and define HRFK for the knowledge test (Phase d; selection criteria shown in **Fig. 1**). As a result of Step 1, HRFK was defined as knowledge regarding principles of exercise and health-related physical fitness (principles), risk reduction and prevention of injuries related to PA and exercise (risk), and health benefits of PA and sports (benefit). The term *knowledge* was understood as an interaction of factual, conceptual, procedural, and metacognitive knowledge (Anderson & Krathwohl, 2001). Compared to general health knowledge, HRFK focuses on knowledge related to PA and health-related physical fitness. In line with an understanding of knowledge in terms of physical literacy, HRFK shares features with knowledge of healthy and active lifestyles but not with knowledge and understanding of activities (Edwards et al., 2017).

Abstract

Ger J Exerc Sport Res 2022 · 52:11–23 https://doi.org/10.1007/s12662-021-00735-5
© The Author(s) 2021

C. Volk · S. Rosenstiel · Y. Demetriou · G. Sudeck · A. Thiel · W. Wagner · O. Höner

Health-related fitness knowledge in adolescence: evaluation of a new test considering different psychometric approaches (CTT and IRT)

Abstract

Fostering health-related fitness knowledge is a common goal across physical education curricula. However, carefully developed knowledge tests that satisfy the psychometric criteria of educational assessment are lacking. Therefore, two studies were conducted to evaluate a newly developed health-related fitness knowledge test within the framework of classical test and item response theory regarding item quality, test reliability, construct validity, and dimensionality. Overall, 794 ninth graders (M_{age} = 14.3 years, 50.6% girls) took part in Study 1. They differed in the type of physical education classes (minor or major subject) and school (lower or higher educational level) they attended. Study 2 incorporated 834 ninth graders at the same educational level (M_{age} = 14.2 years, 52.5% girls). Item–test correlation, test reliability, and validity were examined. In addition, item and test quality were investigated using unidimensional two-parameter logistic item response models. In Study 1, pupils at the same educational level with physical education as a major achieved higher knowledge scores than pupils with physical education as a minor (t = –5.99, p < 0.001; d = 0.58), which confirmed the test's construct validity. In Study 2, the weighted likelihood estimate reliability of the final 27 items was 0.65, and the test–retest reliability reached r_{tt} = 0.70. The items satisfied the assumption of local independence. The final test fulfilled the psychometric criteria of reliability and construct validity to assess health-related fitness knowledge in cross-sectional and interventional studies. This test extends the possibilities of research on health-related fitness knowledge in physical education.

Keywords

Test development · Physical education · Psychometrics · Classical test theory · Item response theory

In Step 2, items were developed according to the definition of HRFK and in accordance with German PE curricula to ensure the curricular validity of the items. Moreover, items were designed to assess pupils' conceptual rather than factual knowledge, their understanding—not their reproduction—of HRFK (Anderson & Krathwohl, 2001), and their action knowledge. The number of developed items per topic varied due to the different range of topics, the weighting of the topics in PE curricula, and the extent to which knowledge was related to action (i.e., knowledge that can be used to perform PA). Therefore, the majority of the developed items was related to principles, whereas fewer items were related to risk or benefit.

In Step 3, the developed items were tested and revised through several pilot studies to identify comprehension problems in the question and answer options and any misconceptions of terms and to test different item response formats. A detailed description of the test development process and the pretests is provided in Supplementary Material S1. Overall, 30 items were chosen for the preliminary version of the HRFK test, which was empirically investigated in Study 1.

Study 1: evaluation of the preliminary HRFK test version

Sample and data collection

Altogether, 794 ninth graders (M_{age} = 14.3 ± 0.5 years, 50.6% girls) from 17 different secondary schools in the school district of Tübingen (Germany) participated in this study in the fall of 2015. In total, 171 ninth graders (21.5%) with PE as a minor subject attended a secondary school (Realschule), which enables them for example to participate in an apprenticeship after their examinations. In contrast, 623 pupils (78.5%) were enrolled in a secondary school (Gymnasium), which provides them with general higher education entry qualifications after examinations.

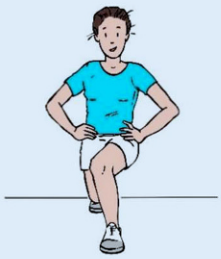
Matching (MA) item: Topic knowledge of principles of exercise and health-related physical fitness (principles)

Match the questions (a to d) with an arrow to the most appropriate answer (1 to 4)!

Subtask 1 →	a)	How can I lower my resting heart rate?	1)	Regularly balancing on a balance beam.
Subtask 2 →	b)	How can I get big muscles?	2)	Regularly running long distances.
Subtask 3 →	c)	How can I get my muscles to work over a longer time?	3)	Regularly lifting moderately heavy weights for many repetitions.
Subtask 4 →	d)	How can I improve my coordination?	4)	Regularly lifting heavy weights for a few repetitions.

Open-ended (OE) question: Topic knowledge about risk reduction and prevention of injuries related to PA and exercise (risk)

Conny and Tobi are doing strengthening activities in physical education for the first time. What are they not doing so well? Please put a check mark on that area.



Complex multiple choice (CMC) item: Topic knowledge about the health benefits of PA and sports (benefit)

Playing soccer and handball regularly…
Check right or wrong for each statement (a to f).

			True	False
Subtask 1 →	a)	…reduces hypertension.	☐	☐
Subtask 2 →	b)	…increases bone density.	☐	☐
Subtask 3 →	c)	…reduces the risk of a stroke.	☐	☐
Subtask 4 → and distractor	d)	…does not influence the resting heart rate.	☐	☐
Subtask 5 →	e)	…reduces percentage of body fat.	☐	☐
Subtask 6 →	f)	…reduces the risk of a myocardial infarction.	☐	☐

Fig. 2 ◀ Examples of health-related fitness knowledge test booklet items and subtasks on the three test topics

Of those pupils, 487 had PE as minor subject and 136 as a major. Compared to pupils with PE as a minor subject, pupils with PE as a major had not only a higher number of lessons per week but also explicitly covered theoretical aspects of sports and PA in their lessons.

Data were collected during regular classes. Time to fill out the paper-and-pencil HRFK test booklet was limited to 45 min. Trained testers conducted the study using a standardized test manual. Written informed consent to participate in this study was provided by all pupils and their parents. The study procedures were approved by the ethics committee at the Faculty of Economics and Social Sciences, University of Tübingen.

Measurement

Of the 30 items in the HRFK test booklet, 18 were related to principles (Principles 1–18), 10 to risk (Risks 1–10), and two to benefit (Benefits 1 and 2). Examples of each topic are provided in ◘ **Fig. 2**. The test comprised 18 complex multiple choice (CMC) items. These CMC items

contained three to six subtasks. For each subtask, pupils were given two answer options (true or false), of which only one was considered to be correct (**Fig. 2**). The applied matching items (MA; $n = 3$) required pupils to match questions or statements to the correct response option. In the case of the sorting item (SO; $n = 1$), pupils were asked to bring pictures in the correct order. For open-ended questions (OE; $n = 8$), pupils were asked to give a short explanation or to flag an area of a picture.

Data analysis

Data were analyzed using SPSS Version 26 (IBM, Armonk, NY, USA), Mplus Version 8.4 (Muthén & Muthén, 2017), and R Version 3.5.3. The TAM Package 3.3.-10 (Robitzsch, Kiefer, & Wu, 2019) was used to estimate the parameters of the logistic IRT models.

Item scoring

Initially, each subtask of the CMC, MA, and SO items was scored dichotomously (0 = incorrect, 1 = correct) to analyze how well the single subtasks worked. Subsequently, a dichotomous or polytomous (i.e., partial credit) item score was calculated for each item. The item score depended on the response format, with partial scoring for CMC and MA items and dichotomous scoring for OE questions and the SO item (Pohl & Carstensen, 2012). Items were also rated according to theoretical considerations: seven CMC items were only coded as correct if all true–false subtasks were correctly solved. These items required knowledge of a single concept that had to be fully understood (e.g., the meaning of intensity of endurance training); therefore, partial scoring was not sufficient. OE questions with written answers were coded independently by two sports scientists to investigate intercoder reliability. In 90.1% of the cases, the two raters agreed in their coding.

Missing responses

Different kinds of missing responses were distinguished in the coding procedure of Study 1 to examine how well pupils from different schools coped with the items (e.g., their understanding of item tasks and different item formats). There were missing responses due to invalid responses (e.g., pupil marked both the true and false answer option in a CMC item) and missing responses because of omitted items or subtasks. If an item subtask had a missing response, the whole item was scored as missing. We re-examined items with a missing response rate greater than 10% to identify any problem items (OECD, 2017). With regard to statistical analyses of CTT, we estimated corrected item–test correlations in a structural equation framework using the full information maximum likelihood (FIML) method to handle missing responses (with one model for each item). The correlation of a given item and the estimated sum score of the remaining items were modeled using the pseudo-indicator model (PIM; Rose, Wagner, Mayer, & Nagengast, 2019). Analogously, for the parameter estimation in the IRT models, missing responses were ignored. Thus, missing responses were treated as missing values instead of incorrect responses because this procedure has been shown to result in unbiased parameter estimates in IRT models and has been applied in large-scale studies (e.g., the National Panel Study [NEPS]; Pohl, Gräfe, & Rose, 2014; Pohl & Carstensen, 2012).

Analysis of item subtasks

Subtask discrimination values were analyzed within the framework of CTT and IRT to select subtasks that could be aggregated to CMC, MA, and SO items. We assumed that one latent variable representing the construct HRFK was essentially responsible for pupils' test answers. Therefore, the values had to be positive. A positive discrimination value indicated that the subtask was more likely to be solved by individuals with a higher level of HRFK than those with a lower level. With regard to CTT, corrected point-biserial correlations between each subtask, the item score, and the total test score were computed to analyze subtask discrimination. With regard to IRT, the subtask discrimination parameter for all subtasks was estimated with a two-parameter logistic (2PL; Birnbaum, 1968) model using marginal maximum likelihood (MML) estimation.

Furthermore, the selection rates of item distractors were investigated. Distractors are a subset of subtasks that represent the incorrect response option of an item (see **Fig. 1** for an example). Following the NEPS (Pohl & Carstensen, 2012), we classified a distractor as good if the correlation between the selection rate of an item's distractor and the total test score was negative (i.e., $r_{pb} < 0.00$ = good, $0.00 \leq r_{pb} \leq 0.05$ = acceptable, and $r_{pb} > 0.05$ = problematic). All defined criteria were decisive for the inclusion of the subtasks in the subsequent analyses.

Item analysis

Item difficulty and discrimination were analyzed within the framework of CTT and IRT to select appropriate items. Item difficulty had to vary between easy and difficult items. Analogous to subtask discrimination, item discrimination had to be positive in order to distinguish between pupils with different levels of HRFK.

As for item difficulty in the context of CTT, the percentage frequency distribution of the item score was computed (0 to a maximum of 2 or 3 points for polytomous items, 0 to 1 point for dichotomous items). Dichotomous items with less than 5% or greater than 95% correct responses were flagged as conspicuous. This indicated that hardly any or almost all pupils, respectively, answered the item correctly (i.e., scored 1 point). Subsequently, with regard to item discrimination in CTT, the corrected item–test correlation was evaluated. Corrected item–test correlations were rated according to the NEPS ($r_{it} > 0.30$ = good, $0.30 \geq r_{it} \geq 0.20$ = acceptable, and $r_{it} < 0.20$ = problematic; Pohl & Carstensen, 2012). Within the IRT framework, the generalized partial credit (GPC; Muraki, 1992) model was used to evaluate the quality of the items with regard to estimated item difficulty and item discrimination parameters (estimation method: MML). For polytomous items, the difficulty for each score category of an item can be described by a transition location parameter that "is the point where the probability of responding in

two adjacent categories is equal" (de Ayala, 2009, p. 167). Therefore, the mean of the transition location parameters of an item was used as the average item difficulty for polytomous items (Wu, Tam, & Jen, 2017). Item difficulty can theoretically vary between $\pm\infty$, but it usually ranges from +3 (difficult item) to –3 (easy item) in IRT models (de Ayala, 2009). Item discrimination parameter $\alpha \geq 0.53$ was chosen as the selection criterion ($\alpha = 0.53$ is comparable with a standardized factor loading of $\lambda = 0.30$ in a categorical confirmatory factor analysis model with an assumed underlying normally distributed variable for a categorical indicator).[1] The decision to eliminate individual items was based on the above-defined criteria as well as on the basis of content aspects.

Evaluation of the validity and reliability of the preliminary test

Selected items were included in the GPC model to estimate pupils' IRT-based HRFK scores (i.e., person abilities) and the reliability and validity of the preliminary test. Weighted maximum likelihood estimation (WLE; Warm, 1989) was used to estimate HRFK scores. In order to identify the GPC model and thus be able to unambiguously estimate HRFK scores based on the GPC model, the variance and mean of the person ability distribution were constrained (mean = 0, variance = 1). The WLE person separation reliability was calculated to describe the test reliability. This index is used in the context of IRT and is comparable to Cronbach's alpha in CTT (de Ayala, 2009). In addition, *t*-tests were conducted to examine the construct validity of the test by comparing the estimated HRFK scores of ninth graders from different educational levels and ninth graders from the same educational level with PE as a major or minor subject.

[1] Item discrimination parameter (a_j) in the 2PL IRT model can be transformed into standardized factor loadings following Wirth and Edwards (2007): $\lambda_j = \frac{a_j/1.7}{\sqrt{1+\left(\frac{a_j}{1.7}\right)^2}}$.

Results

Missing responses

The number of missing responses ranged from 3 (0.38%) to 107 (13.48%) per item, including all types of missing responses. Two OE questions showed a missing response rate of ≥ 10% (Risks 2 and 4). In total, 51.51% of the pupils had no missing response at all, and only 4.65% had five or more missing responses. On average, pupils had 1.03 ± 1.59 (Min = 0, Max = 16) missing values in a test of 30 items.

Analysis of the subtasks

Each single subtask of six CMC items (Principles 6, 8, 15, and 16, Benefit 1, Risk 10) was excluded due to negative point-biserial correlations and/or negative estimated discrimination parameters and positive correlation with regard to the distractor analysis ($r_{pb} > 0.05$). In addition, one item (Risk 9) was not included in the subsequent analysis because two of the four subtasks of the CMC item showed a low item and total score correlation and subtask discrimination parameter. All other subtasks were aggregated to dichotomous or polytomous items.

Item analysis

Table 1 shows the results of the item analysis and gives a short explanation of each item's content (knowledge area). The percentage of correct responses ($M = 57.23 \pm 23.53$) as an indicator of item difficulty in CTT ranged from 9.88% (Principle 4) to 96.38% (Risk 5) for dichotomous items (**Table 1**, % item score of 1 point in case of a dichotomous item). As for polytomous items, the percentage of pupils who answered the item completely correct (**Table 1**, % item score of 2 or 3 points) was between 27.50% and 79.21%. With regard to the results of the GPC IRT model, the estimated item difficulty varied between –3.52 (easy) and 5.34 (difficult). Overall, the 29 items were of medium difficulty, including easy (e.g., Principle 16) and very difficult (e.g., Principle 4) items. With regard to item discrimination, 15 items reached a corrected item–test correlation of at least $r_{it} \geq 0.20$ and/or an estimated item discrimination parameter ≥ 0.53.

Based on the aforementioned results, eight items were excluded from the final analysis because of their low corrected item–test score correlation (< 0.20) and item discrimination parameter (< 0.53). Despite falling below both of these psychometric cut-off criteria, a total of six items were retained because of their importance to the content validity of the HRFK test. Finally, 21 items were considered for the subsequent analyses of the reliability and validity of the preliminary test version. These items are marked in bold in **Table 1**.

Reliability and validity of the preliminary HRFK test

The reliability of the HRFK test (21 items) was moderate (WLE person separation reliability = 0.59). Ninth graders' estimated IRT-based HRFK scores ranged from –3.90 (low level of knowledge) to 7.25 (very high level of knowledge). On average, pupils' estimated HRFK scores were $M = 0.00 \pm 1.31$. Pupils with PE as a major subject ($n = 136$, $M = 0.75 \pm 1.31$) reached significantly higher levels of HRFK ($t = -5.99$, $p < 0.001$) than pupils with PE as a minor subject ($n = 487$, $M = 0.02 \pm 1.24$); the group effect was of medium size (Cohen's $d = 0.58$). In addition, estimated HRFK scores differed significantly and meaningfully depending on the pupils' educational level ($t = 5.94$, $p < 0.001$; $d = 0.53$). Pupils from a lower educational level ($n = 171$) had lower HRFK scores ($M = -0.63 \pm 1.16$) compared to pupils from a higher educational level with PE as a minor subject, indicating construct validity of the test.

Study 2: evaluation of the final HRFK test

Item revisions and purpose of study 2

As the results of Study 1 revealed that item–test correlations and test reliability were rather low for the selected items, we decided to add new items to the test booklet that were similar in content to the selected ones in order to increase the homogeneity of the test. For example, fur-

Table 1 Results of the item analysis (Study 1)

Item name	Knowledge area	Item analysis (CTT)				Item analysis (IRT)		
		Frequency distribution of item score (%)				r_{it} (SE)	Difficulty	Discrimination (SE)
		0	1	2	3			
Principle 1	Perceived exertion during PA	5.99	27.99	66.02	–	0.16 (0.04)	–3.42	0.37 (0.06)
Principle 2	Frequency, intensity, time, and type of PA to improve fitness	14.47	38.45	47.08	–	0.34 (0.03)	–0.95	0.79 (0.06)
Principle 3	Training principles	28.09	71.91	–	–	0.20 (0.04)	–1.56	0.66 (0.09)
Principle 4	Factors that influence heart rate	90.12	9.88	–	–	0.10 (0.04)	5.34	0.43 (0.12)
Principle 5	Type of activity to improve fitness	5.10	15.69	79.21	–	0.23 (0.03)	–2.58	0.64 (0.06)
Principle 6	Meaning of intensity (muscular fitness)	67.70	32.30	–	–	0.20 (0.04)	1.31	0.62 (0.08)
Principle 7	Exercise to improve flexibility (back, leg)	79.49	20.51			0.22 (0.04)	1.84	0.84 (0.09)
Principle 8	Frequency, intensity, time, and type of PA to improve muscular fitness	10.23	41.30	48.47		0.17 (0.04)	–2.40	0.34 (0.06)
Principle 9	Factors that influence heart rate	11.70	44.09	44.22		0.20 (0.04)	–1.74	0.41 (0.06)
Principle 10	Monitoring heart rate	11.43	48.49	40.08		0.24 (0.04)	–1.35	0.51 (0.06)
Principle 11	Training principles	80.15	19.85			0.18 (0.04)	2.29	0.67 (0.09)
Principle 12	Physiological responses (cardiovascular system) to PA	9.49	25.51	65.00		0.24 (0.03)	–2.36	0.45 (0.06)
Principle 13	Frequency, intensity, and time to improve fitness	44.76	55.24			0.18 (0.04)	–0.48	0.46 (0.08)
Principle 14	Training principles	28.15	71.85			0.14 (0.04)	–2.35	0.42 (0.08)
Principle 15	Training principles	39.05	60.95			0.11 (0.04)	–1.60	0.28 (0.08)
Principle 16	Exercise to improve cardiovascular fitness	8.67	91.33			0.15 (0.04)	–3.52	0.73 (0.12)
Principle 17	Exercise to improve muscular fitness (back, stomach, leg)	40.71	59.29			0.21 (0.03)	–0.64	0.64 (0.08)
Principle 18	Training principles	29.20	70.80			0.17 (0.04)	–2.16	0.43 (0.08)
Risk 1	Proper knee position (squat)	19.87	80.13			0.18 (0.04)	–2.47	0.61 (0.09)
Risk 2	Proper back position (quadruped arm/leg extension)	39.88	60.12			0.11 (0.04)	–1.17	0.36 (0.08)
Risk 3	Proper trunk/arm position (side crunches)	35.22	64.78			0.20 (0.04)	–1.15	0.56 (0.08)
Risk 4	Proper shoulder position (front raise arm extension)	57.22	42.78			0.19 (0.04)	0.56	0.55 (0.08)
Risk 5	Proper back and knee position (lift a box)	3.62	96.38			0.00 (0.04)	–34.30[a]	0.10 (0.19)
Risk 6	Proper arm position (carry a box)	30.19	69.81			0.01 (0.04)	142.03[a]	–0.01 (0.08)
Risk 7	Proper back position (carry a box)	39.76	60.24			0.15 (0.04)	–1.12	0.38 (0.08)
Risk 8	Proper strength training	50.77	49.23			0.16 (0.04)	0.08	0.38 (0.08)
Risk 9[b]	Effects of a warm-up for exercise	–	–	–	–	–	–	–
Risk 10	Proper warm-up for exercise	9.45	38.70	51.85		0.18 (0.04)	–2.70	0.33 (0.06)
Benefit 1	Effects of soccer on health	16.29	22.16	34.05	27.50	0.22 (0.04)	–0.70	0.28 (0.04)
Benefit 2	Effects of PA on health	17.76	43.24	39.00		0.18 (0.04)	–1.30	0.32 (0.05)
Mean (SD)						0.17 (0.07)	–0.97 (1.93)	0.47 (0.19)

Names of selected items as a result of the item analysis are marked in bold. The item score (CTT) can range from 0 to 1 point (correct answer) for dichotomous items and 0 to a maximum of 2 or 3 points for polytomous items. Cells of frequency distribution of item score are therefore left blank if the item score cannot be obtained for the specific item. Difficulty corresponds to the (average) estimated item difficulty parameter of the generalized partial credit (GPC) model. Discrimination is the estimated item discrimination parameter of the GPC model

PA physical activity, *CTT* classical test theory, *IRT* item response theory, r_{it} corrected item–test correlation using pseudo-indicator model (PIM) and full information maximum likelihood (FIML)

[a] The item difficulties were not considered in the summary statistics as their values were not plausible due to estimation problems. Less than 5% of the pupils answered the item incorrectly or negative item discrimination parameter

[b] Dashes (–) in Risk 9 represent data that are not reported because the item was not included in the item analysis

ther items were developed that required knowledge of the proper exercise to improve muscular fitness (e.g., Principle 17, **Table 1**) but asked for different exercises than the selected items of Study 1. Moreover, we added items that represented the benefit topic, as the HRFK booklet from Study 1 only included two items on this topic. All new items were based on our definition of HRFK and aligned with PE curricula.

Furthermore, we revised the selected items from Study 1 by adding subtasks in case one subtask was deleted as a result of Study 1 (Principle 6) so that we would still have a sufficient, comparable number of subtasks per item. Moreover, we adapted the item response format (e.g., an OE question instead of a CMC item; Principles 7, 16, and 17, Risks 1–4) as we assumed that this change would increase the possibility of measuring pupils' understanding rather than their reproduction of knowledge. The revised test booklet was finally tested with the target group in the context of Study 2. In addition to the analysis of Study 1, differential item functioning, conditional independence, and dimensionality of the test were examined.

Sample and design

Data gathered within the GEKOS CRCT study (Haible et al., 2019) between fall 2017 and spring 2019 were used for this study. Pupils' HRFK was measured before and after an intervention (time interval: $M = 11.0 \pm 1.6$ weeks). Baseline data of $N = 834$ ninth graders (52.5% girls, $M_{age} = 14.2$) were used to study item and test quality. Furthermore, the test–retest reliability of the HRFK test was calculated by utilizing pretest and posttest results for 325 pupils (47.1% girls) from the control group. Data collection followed the procedures of Study 1.

Measurement

The HRFK test booklet included 33 items (14 CMC, 7 MA, 1 SO, 9 OE, and 2 single-choice items). Twenty-one items were drawn from the selected item pool of Study 1, and 12 new items were developed (marked with the superscript lowercase letter "a" in **Table 2**). Altogether, principles were covered with 25 items, whereas risk and benefit were assessed with four items each. All new items were pretested with ninth graders.

Data analysis

In preparation for the main analyses, data processing and analyses of items followed the procedures outlined in Study 1. The nine OE questions were scored by the main researcher and three trained student research assistants using a standardized coding system. However, two questions (Principles 27 and 28) were excluded due to insufficient interrater agreement. Overall, the intercoder reliability of the included items ranged from 79.81% to 94.11%.

As all test items had to be equally applicable to girls and boys, differential item functioning was studied with the package ltm (Rizopoulos, 2018). Item difficulty and item discrimination parameters were estimated separately for boys and girls using the 2PL model (Birnbaum, 1968).[2] Regarding conditional independence (i.e., given the HRFK level, responses to an item are independent of the responses to any other item) and test dimensionality, correlations between the residuals for item pairs (Q_3 statistic; Yen, 1993) were calculated. Item pairs with $Q_3 > |0.20|$ were flagged as conspicuous using the cut-off value suggested by Yen (1993; see also Gnambs, 2017). At the same time, values below 0.20 suggested essential unidimensionality (Gnambs, 2017). As we intended to scale the final HRFK test based on a unidimensional IRT model, violations of a strict unidimensional assumption of the test were further investigated with exploratory factor analysis (EFA; rotation method: GEOMIN) and confirmatory factor analysis (CFA) for categorical data (estimator: WLSMV) to gain an in-depth understanding of the data structure. With regard to the CFA, dimensionality was tested by specifying both a one- and three-dimensional model based on the topics of the test, which could potentially reflect different subfacets. A Chi-squared (χ^2) difference test was conducted to compare the two models using the DIFFTEST command in Mplus. Finally, selected items were included in the GPC model (Muraki, 1992). Pupils' IRT-based HRFK scores (i.e., person abilities) were estimated using WLE (Warm, 1989). For the same reasons cited in Study 1, the mean and variance of the person ability distribution were constrained (mean = 0, variance = 1). The reliability of the test was calculated as WLE person separation reliability. Pearson's correlation between pupils' pre- and postintervention estimated HRFK scores was calculated to determine test–retest reliability.

[2] Since the number of students per category was relatively small for several polytomous items, and the parameter estimation per group resulted in a bisection of the sample size, we recoded the polytomous items into dichotomous items (i.e., 1 point = all subtasks are answered correctly).

Results

Missing responses

The share of missing responses per item was low ($M = 3.11\%$, Min = 0.60%, Max = 9.23%). In total, 56.35% of pupils had no missing response at all, and only 4.44% had five or more missing responses. On average, pupils had $M = 0.96 \pm 1.63$ missing values.

Analysis of the subtasks

Each single subtask of four CMC items (Principles 6 and 22, Benefits 3 and 4) was excluded from the subsequent item analysis due to negative correlation coefficients and/or negative estimated item discrimination parameters. All distractors of the CMC items were nonpositively correlated with the test score, apart from the two previously mentioned subtasks (Principles 6 and 22; $r_{pb} > 0.00$). All the excluded subtasks were newly developed and thus not part of Study 1.

Item analysis

The percentage of correct answers for the dichotomous items ($M = 40.02 \pm 22.27$) varied between 6.82% and 83.27%. The results of the GPC model yielded item difficulties between −5.78 and 12.54. Overall, these results indicate item variability with regard to their difficulty, whereas two items (Principles 24 and

Table 2 Results of the item analysis (Study 2)

Item name	Knowledge area	Item analysis (CTT)				Item analysis (IRT)		
		Frequency distribution of item score (%)				r_{it} (SE)	Difficulty	Discrimination (SE)
		0	1	2	3			
Principle 1	Perceived exertion during PA	7.42	33.99	58.59		0.14 (0.04)	–3.69	0.29 (0.06)
Principle 2	Frequency, intensity, time, and type of PA to improve fitness	12.80	40.58	46.62		0.30 (0.03)	–1.23	0.61 (0.06)
Principle 3	Training principles	41.62	58.38			0.18 (0.03)	–0.72	0.50 (0.08)
Principle 4	Factors that influence heart rate	93.18	6.82			0.09 (0.04)	5.59	0.49 (0.13)
Principle 5	Type of activity to improve fitness	16.73	83.27			0.18 (0.03)	–2.51	0.70 (0.09)
Principle 6	Meaning of intensity (muscular fitness)	76.24	23.76			0.14 (0.04)	2.40	0.52 (0.09)
Principle 7	Exercise to improve flexibility (back, leg)	14.98	41.07	43.95		0.34 (0.03)	–0.88	0.77 (0.06)
Principle 9	Factors that influence heart rate	6.28	32.02	34.36	27.34	0.12 (0.04)	–3.29	0.15 (0.04)
Principle 10	Monitoring heart rate	10.92	49.88	39.21		0.19 (0.03)	–1.73	0.40 (0.06)
Principle 11	Training principles	79.29	20.71			0.12 (0.04)	3.66	0.38 (0.09)
Principle 12	Physiological responses (cardiovascular system) to PA	9.02	24.39	66.59		0.22 (0.03)	–2.61	0.42 (0.06)
Principle 15	Training principles	32.69	67.31			0.11 (0.03)	–2.17	0.34 (0.08)
Principle 16	Exercise to improve cardiovascular fitness	66.55	33.45			0.14 (0.04)	2.00	0.36 (0.08)
Principle 17	Exercise to improve muscular fitness (back, stomach, leg)	11.71	37.73	50.55		0.19 (0.03)	–1.94	0.41 (0.05)
Principle 18	Training principles	28.85	71.15			0.19 (0.03)	–1.91	0.50 (0.08)
Principle 19[a]	Exercise to improve flexibility (back, leg)	17.81	23.87	25.94	32.39	0.32 (0.03)	–0.48	0.53 (0.04)
Principle 20[a]	Exercise to improve muscular fitness (back, stomach, leg)	17.55	30.30	34.22	17.93	0.18 (0.04)	–0.03	0.27 (0.04)
Principle 21[a]	Physiological responses (muscle) to PA	9.77	21.37	68.86		0.10 (0.04)	–5.78	0.17 (0.05)
Principle 22[a]	Meaning of intensity (cardiovascular fitness)	63.45	36.55			0.27 (0.03)	0.83	0.75 (0.08)
Principle 23[a]	Factors influencing perceived exertion	16.65	36.82	46.54		0.27 (0.03)	–1.24	0.46 (0.05)
Principle 24[a]	Frequency, intensity, and time of PA to improve cardiovascular fitness	82.45	17.55			0.08 (0.04)	8.71	0.18 (0.09)
Principle 25[a]	Frequency, intensity, and time of PA to improve muscular fitness	71.25	28.75			0.09 (0.04)	3.13	0.30 (0.08)
Principle 26[a]	Exercise to improve muscular fitness	81.89	18.11			0.03 (0.04)	12.54	0.12 (0.09)
Principle 27[a,b]	Exercise to improve coordination	–	–	–	–	–	–	–
Principle 28[a,b]	Exercise to improve flexibility	–	–	–	–	–	–	–
Risk 1	Proper knee position (squat)	36.57	63.43			0.13 (0.04)	–1.84	0.30 (0.08)
Risk 2	Proper back position (quadruped arm/leg extension)	65.24	34.76			0.10 (0.04)	2.19	0.29 (0.08)
Risk 3	Proper trunk/arm position (side crunches)	58.18	41.82			0.16 (0.04)	0.78	0.45 (0.08)
Risk 4	Proper shoulder position (front raise arm extension)	65.52	34.48			0.24 (0.04)	1.01	0.71 (0.09)
Benefit 1	Effects of soccer on health	10.56	21.74	36.02	31.68	0.19 (0.03)	–1.60	0.24 (0.04)
Benefit 2	Effects of PA on health	17.76	37.47	44.77		0.25 (0.03)	–1.10	0.47 (0.05)
Benefit 3 [a]	Effects of PA on health	18.10	40.83	41.07		0.27 (0.03)	–0.93	0.50 (0.05)
Benefit 4 [a]	Effects of swimming on health	6.64	32.23	61.13		0.21 (0.03)	–2.90	0.42 (0.06)
Mean (SD)						0.18 (0.08)	0.14 (3.70)	0.42 (0.17)

Names of selected items as a result of the item analysis are marked in bold. The item score (CTT) can range from 0 to 1 point (correct answer) for dichotomous items and 0 to a maximum of 2 or 3 points for polytomous items. Cells of frequency distribution of item score are therefore left blank if the item score cannot be obtained for the specific item. Difficulty corresponds to the (average) estimated item difficulty parameter of the generalized partial credit (GPC) model. Discrimination is the estimated item discrimination parameter of the GPC model

PA physical activity, *CTT* classical test theory, *IRT* item response theory, r_{it} corrected item–test correlation using pseudo-indicator model (PIM) and full maximum likelihood estimation (FIML)

[a] Newly developed items for Study 2

[b] Dashes (–) in Principles 27 and 28 represent data that are not reported as the items were not included in the item analysis

26) were conspicuously difficult. With regard to item discrimination, 11 items had a corrected item–test correlation $r_{it} \geq 0.20$ and/or an estimated item discrimination parameter ≥ 0.53. The results of the item analysis are summarized in Table 2.

As a result of the item analysis, three items (Principles 21, 24, and 26) were excluded due to very low item discrimination and either very high or relatively low item difficulty. Moreover, 15 items were initially retained, despite not meeting the cut-off values (item discrimination parameters < 0.53 or $r_{it} < 0.20$) because they distinguished between pupils with and without PE as a major subject and between pupils from different school types in Study 1. The same decision was made for two newly developed items (Principle 25, Benefit 4) due to their importance regarding the content validity of the test.

The selected 28 items (Table 2, marked in bold) were scaled separately for boys ($n = 396$) and girls ($n = 438$) to investigate differential item functioning. As a result of this analysis (see Supplementary Material S2 for further detail), we did not exclude any items due to gender differences.

Conditional independence and dimensionality

The 28 items were further studied within a unidimensional GPC IRT model in order to investigate conditional independence and dimensionality of the test. The correlations between the residuals of item pairs (Q_3 statistics) were generally small ($M = -0.01 \pm 0.05$) except for two pairs (the correlations $Q_{3\_principle6,\ principle22} = 0.27$ and $Q_{3\_benefit2,\ benefit3} = 0.37$), which exhibit some amount of dependency. As the items Benefit 2 and Benefit 3 were almost identical with regard to content, and the Q_3 statistic indicated item dependency, the two items were combined into a single polytomous item (i.e., 27 items remained). Considering the dimensionality of the HRFK test, the generally low residual correlations of item pairs also indicated that an essentially unidimensional test could be assumed.

The results of the EFA suggested that 11 factors could be extracted from the data in case eigenvalues (≥ 1) were considered. Viewing two- and three-factor solutions of the EFA, items were clustered according to their item response format rather than their topic, indicating method-specific (response format) associations among items rather than underlying content-related factors.

The three-dimensional CFA model, based on the three topics, fit the data significantly better than the unidimensional model ($\chi^2 = 31.75$, $df = 3$, $p < 0.001$). The correlation between Factor 1 (items related to the topic principles) and Factor 2 (risk) was relatively high ($r_{12} = 0.68$, $p < 0.001$). Correlations between these two factors and Factor 3 (benefit) were lower ($r_{13} = 0.61$, $p < 0.001$; $r_{23} = 0.21$, $p = 0.07$). Thus, the results of the CFA indicated that items related to the topic benefit could describe an independent dimension. However, the number of items on this topic was rather small, and the estimation of a separate score showed no sufficient reliability. A detailed description of the EFA and CFA are shown in Supplementary Material S2.

Considering the results of the three different analyses (Q_3, EFA, and CFA) with regard to the test's dimensionality and to avoid construct underrepresentation by excluding all items related to the topic benefits, we chose the most parsimonious and conceptually considered model and estimated a unidimensional HRFK score, even though the assumption of strict unidimensionality might be slightly violated.

Reliability and distribution of the estimated HRFK score of the final HRFK test

Finally, 27 items were scaled using the GPC model. Item discrimination scored between 0.14 and 0.87 ($M = 0.44 \pm 0.18$). The mean item difficulty varied between −3.69 (easy item) and 5.38 (difficult item; $M = -0.42 \pm 2.31$). The distribution of the final HRFK score is shown in Fig. 3. Pupils' estimated HRFK scores ranged from −6.05 (low level of knowledge) to 4.87 (high level of knowledge). The middle 95% of the distribution of pupils' estimated HRFK scores was within a range of −2.35 to 2.76. On average, pupils' HRFK scores were $M = 0.00 \pm 1.26$. Estimation accuracy was the highest for an estimated HRFK score around −0.62 ($SE = 0.67$, $M_{SE} = 0.74 \pm 0.10$). The accuracy decreased in cases where estimated HRFK scores were high or very low. The WLE person separation reliability (WLE reliability = 0.65) and test–retest reliability using the pretest and posttest HRFK scores from $n = 325$ pupils were acceptable ($r_{tt} = 0.70$, $p < 0.001$).

Discussion and conclusion

The aim of this study was to develop an HRFK test for ninth graders based on a systematic definition of the construct and the evaluation of the test within two large samples using standards of educational assessment (CTT and IRT). The new HRFK test extends the options for researchers to obtain a reliable and valid assessment of pupils' levels of HRFK in cross-sectional and interventional studies with ninth graders in PE.

This study used analyses of CTT and IRT to evaluate *item quality*. The results were fairly similar: item difficulty varied satisfactorily, and the number of missing responses per item was very low, indicating that ninth graders have no specific problems coping with the items of the final HRFK test; however, the item discrimination (corrected point-biserial correlation, estimated item discrimination parameter) of some items of the final HRFK test was relatively low regarding the defined cut-off criteria, which might have influenced the findings of the final HRFK test reliability.

The *reliability* (WLE person reliability, test–retest reliability) of the final HRFK test was reasonable in order to study differences between HRFK scores at the group level in PE research (e.g., differences between the HRFK level of PE classes from different federal states in Germany; Höner & Roth, 2002; Lienert & Raatz, 1998). The WLE reliability of the final HRFK test was slightly lower compared to Töpfer's (2019) health-related sport competence test (WLE = 0.78), which includes aspects of knowledge as well; however, Töpfer (2019) examined

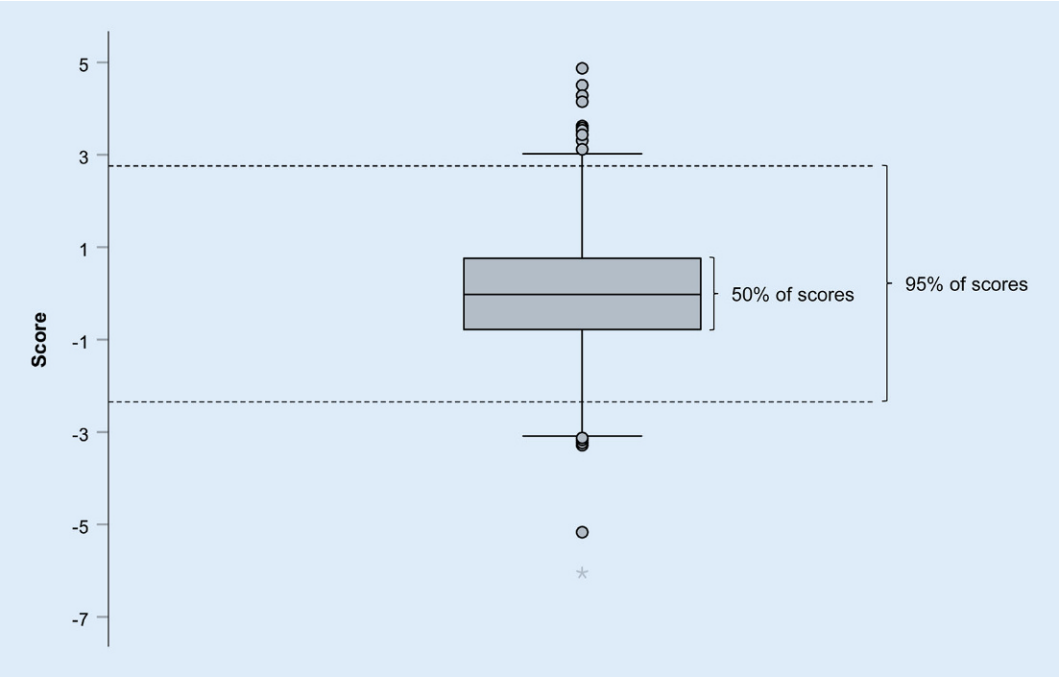


Fig. 3 ◀ Distribution of the weighted maximum likelihood estimated health-related fitness knowledge score

the test's quality in a more heterogeneous group (i.e., seventh to tenth graders from different types of schools) than that of the present article (ninth graders), which might explain the differences. The test–retest reliability was similar to the results of Longmuir et al. (2018), who investigated the reliability of a physical literacy knowledge questionnaire ($r_{tt} = 0.60$ and 0.69 after 2 and 7 days, respectively), with considerable shorter retest intervals in comparison to the present Study 2 ($M = 11$ weeks).

The reliability of the final HRFK test still seems insufficient for individual diagnostics (e.g., the comparison of HRFK scores between individuals) due to measurement error. The measurement error may have been caused by the low item discrimination parameters and the item–test correlations of some items, respectively, of the final HRFK test. As HRFK represents a heterogeneous construct (i.e., low item intercorrelations) and the content of the items contains a wide variety of topics (i.e., heterogeneous items), the item discrimination and test reliability—in terms of internal consistency—may have been negatively affected (Schermelleh-Engel & Werner, 2008). In addition, although the distribution of the IRT-based HRFK score of the final HRFK test indicated that the test can measure different levels of HRFK (low and high levels of knowledge), the majority of items are of medium difficulty. Therefore, the standard error was highest for pupils achieving a high or a relatively low estimated HRFK score. Considering test reliability, the addition of some easy and difficult items in particular could be an initial step in improving the reliability of the test.

This article provides preliminary evidence of the *validity* (content validity, construct validity) of the HRFK test. Similar to the approach of Zhu et al. (1999) and Longmuir et al. (2018), items were developed in accordance with the content of different PE curricula to ensure the content validity of the test. As descriptions of German PE curricula are often imprecise and lack detail, it was difficult to derive precise questions for the test items. In future research, it could therefore be worthwhile if PE teachers could classify and weigh the items with regard to the PE content to study the content validity in further detail.

The preliminary HRFK test (Study 1) successfully differentiated between pupils with PE as a major or minor subject, which confirms the test's construct validity. The results of the GEKOS intervention study (Volk et al., 2021) support the present findings with regard to construct validity for the final HRFK test as well and evince the test's sensitivity to measure change as the HRFK test successfully differentiated between pupils of the control and intervention groups who participated in a 6-week intervention program on HRFK. Thus, the final HRFK test can not only be applied by researchers to distinguish between different HRFK levels of PE classes but also to measure the intervention-related development of HRFK within groups.

Nevertheless, further studies on the validity of the final HRFK test booklet are still needed to confirm the current results. For example, it may be valuable to compare the knowledge level of ninth graders with different age groups (e.g., eighth and tenth graders in secondary school) or in contrast to university students enrolled in sports science programs.

While the reliability and validity of the final HRFK test was found to be sufficient for application in PE research investigating pupils' HRFK at the group level, some challenges with regard to the *dimensionality* of the test emerged.

As the EFA and CFA did not provide a consistent picture regarding test dimensionality, the unidimensional IRT model that was conceptually considered and supported by the results of the Q_3 statistics (indicating essential unidimen-

sionality of the HRFK test with negligible correlations among the residuals) was maintained to scale the final HRFK test.

The inconsistent results with regard to the final HRFK test's dimensionality could be explained by different factors. First, the number of items between the different topics of the final HRFK test varied considerably. Of the 27 items on the final HRFK test, 20 items were related to the topic principles, while four and three items were related to risk and benefit, respectively. Considering the results of the CFA, especially for the topic benefit, the development of additional items to further clarify whether this topic actually represents a separate dimension would therefore be desirable in the future.

Second, different item formats were used, which varied with regard to the assessed level of understanding and the possibility of guessing a correct answer (OE question or CMC item). Third, the HRFK test items were created based on a variety of German PE curricula—not just the PE curriculum of the studied sample. Therefore, pupils might not have had an in-depth understanding of HRFK for all topics as they may not have been exposed to certain content in their lessons.

A new PE curriculum that focuses more on the different topics of HRFK than the current PE curriculum will soon be implemented in the schools studied. Assuming that pupils would gain an in depth-understanding of all topics of HRFK after the new PE curriculum is implemented, it would therefore be desirable to investigate the final HRFK test again to obtain a further understanding of the test dimensionality in particular.

In addition to the results regarding HRFK test reliability, validity, and dimensionality, this article documents the first findings on the distribution of ninth graders' estimated HRFK score, which can serve as a reference for researchers who apply the HRFK test in future studies in PE.

This article focused on the development of a HRFK test for ninth graders in secondary schools with high educational levels. To increase the applicability of the test for cross-sectional and interventional studies on HRFK in PE research, future studies should investigate the extent to which psychometric properties are generalizable to pupils who attend other types of schools (lower educational levels) or who belong to different age groups. The present HRFK test could thus serve as a basis for age-specific adaptions related to the respective content of PE curricula up to the relevant age group. Furthermore, as the intention of this study was to develop an HRFK test that measures knowledge used to perform PA (action knowledge), future studies are required to examine whether HRFK actually supports pupils' PA behaviors or levels related to physical fitness. Finally, the present HRFK test was designed for application in PE research. As the development of the HRFK represents a learning objective in PE (Cale & Harris, 2018; Wagner, 2016), PE teachers also need a test to assess pupils' learning progress. Therefore, it could be worthwhile to consider how the current HRFK test can be adapted as an assessment for PE teachers that is easy to use.

Corresponding address

Carmen Volk
Institute of Sports Science,
University of Tübingen
72074 Tübingen, Germany
carmen.volk@uni-tuebingen.de

Acknowledgements. Study 2 was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). We would like to thank all the schools, teachers, and students who participated in our studies. Furthermore, we thank our research assistants and the regional council of Tübingen (Department 7, Sport) for their support.

Funding. Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest. C. Volk, S. Rosenstiel, Y. Demetriou, G. Sudeck, A. Thiel, W. Wagner and O. Höner declare that they have no competing interests.

All procedures performed in studies involving human participants or on human tissue were in accordance with the ethical standards of the institutional and/or national research committee and with the 1975 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

Ajzen, I., Joyce, N., Sheikh, S., & Cote, N. G. (2011). Knowledge and the prediction of behavior: the role of information accuracy in the theory of planned behavior. *Basic and Applied Social Psychology*, *33*, 101–117. https://doi.org/10.1080/01973533.2011.568834.

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

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3rd edn.) [Introduction to the development of tests and questionnaires]. Pearson.

Cale, L., & Harris, J. (2018). The role of knowledge and understanding in fostering physical literacy. *Journal of Teaching in Physical Education*, *37*, 280–287. https://doi.org/10.1123/jtpe.2018-0134.

Demetriou, Y., Sudeck, G., Thiel, A., & Höner, O. (2015). The effects of school-based physical activity interventions on students' health-related fitness knowledge: a systematic review. *Educational Research Review*, *16*, 19–40. https://doi.org/10.1016/j.edurev.2015.07.002.

Edwards, L. C., Bryant, A. S., Keegan, R. J., Morgan, K., & Jones, A. M. (2017). Definitions, foundations and associations of physical literacy: a systematic review. *Sports Medicine*, *47*, 113–126. https://doi.org/10.1007/s40279-016-0560-7.

Gnambs, T. (2017). NEPS technical report for English reading competence: scaling results of the starting cohort 4 for grade 10 (NEPS survey paper no. 26). Leibniz institute for educational trajectories, national educational panel study. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XXVI.pdf. Accessed 12 January 2020

Haible, S., Volk, C., Demetriou, Y., Höner, O., Thiel, A., Trautwein, U., & Sudeck, G. (2019). Promotion of physical activity-related health competence in physical education: study protocol for the GEKOS cluster randomized controlled trial. *BMC Public Health*, *19*, Article 396. https://doi.org/10.1186/s12889-019-6686-4.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development.

Educational Measurement: Issue and Practice, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x.

Höner, O., & Roth, K. (2002). Klassische Testtheorie: Die Gütekriterien sportwissenschaftlicher Erhebungsmethoden [Classical test theory: The psychometric properties of measurements in sports science]. In R. Singer & K. Willimczik (Eds.), *Sozialwissenschaftliche Forschungsmethoden in der Sportwissenschaft* (pp. 67–97). Czwalina.

Keating, X. D., Harrison, L., Chen, L., Xiang, P., Lambdin, D., Dauenhauer, B., Rotich, W., & Piñero, J. C. (2009). An analysis of research on student health-related fitness knowledge in K-16 physical education programs. *Journal of Teaching in Physical Education*, *28*(3), 333–349. https://doi.org/10.1123/jtpe.28.3.333.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise [On the Development of National Education Standards: An Expertise]*. BMBF.

Kulinna, P. H., & Zhu, W. (2001). Fitness portfolio calibration for first- through sixth-grade children. *Research Quarterly for Exercise and Sport*, *72*(4), 324–334. https://doi.org/10.1080/02701367.2001.10608969.

Kurz, D. (2008). Der Auftrag des Schulsports [Aims of physical education]. *Sportunterricht*, *57*(7), 1–8.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6th edn.) [Test construction and test analysis]. Betz.

Longmuir, P. E., Woodruff, S. J., Boyer, C., Lloyd, M., & Tremblay, M. S. (2018). Physical literacy knowledge questionnaire: feasibility, validity and reliability for Canadian children aged 8 to 12 years. *BMC Public Health*, *18*, Article 1035. https://doi.org/10.1186/s12889-018-5890-y.

Ministry of Education and Cultural Affairs, Youth and Sports of Baden-Württemberg (2016). *Bildungsplan des Gymnasiums – Sport [Physical education curriculum for secondary schools]*. Neckar-Verlag.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th edn.). Muthén & Muthén.

Organisation for Economic Co-operation and Development (2017). PISA 2015 technical report. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf. Accessed 20 January 2020

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*(3), 423–452. https://doi.org/10.1177/0013164413504926.

Pohl, S., & Carstensen, C. H. (2012). NEPS technical report – scaling the data of the competence tests (NEPS working paper no. 14). Otto-Friedrich-Universität, Nationales Bildungspanel. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf. Accessed 26 January 2020

Rizopolous, D. (2018). Package 'ltm'. Latent trait models under IRT version 1.1-1. https://cran.r-project.org/web/packages/ltm/ltm.pdf. Accessed 26 January 2020

Robitzsch, A., Kiefer, T., & Wu, M. (2019). TAM: Test analysis modules. R package version 2.7-56. https://cran.r-project.org/web/packages/TAM/index.html. Accessed 26 January 2020

Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-based manifest and latent composite scores in structural equation models. *Collabra: Psychology*, *5*(1), 9. https://doi.org/10.1525/collabra.143.

Schermelleh-Engel, K., & Werner, C. (2008). Methoden der Reliabilitätsbestimmung [Methods for assessing reliability]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 114–133). Springer. https://doi.org/10.1007/978-3-540-71635-8.

Society of Health and Physical Educators (SHAPE) (2014). *National standards & grade-level outcomes for K-12 physical education*. Human Kinetics.

Töpfer, C. (2019). *Sportbezogene Gesundheitskompetenz: Kompetenzmodellierung und Testentwicklung für den Sportunterricht [Sport-related health competence: A competence model and test development for physical education]*. Czwalina.

Tremblay, M. S., Costas-Bradstreet, C., Barnes, J. D., Bartlett, B., Dampier, D., Lalonde, C., Leidl, R., Longmuir, P., McKee, M., Rhonda, P., Way, R., & Yessis, J. (2018). Canada's physical literacy consensus statement: process and outcome. *BMC Public Health*, *18*, Article 1034. https://doi.org/10.1186/s12889-018-5903-x.

Volk, C., Rosenstiel, S., Demetriou, Y., Krustrup, P., Thiel, A., Trautwein, U., Wagner, W., Höner, O., & Sudeck, G. (2021). Effects of a physical education intervention programme for ninth-graders on physical activity-related health competence: findings from the GEKOS cluster randomised controlled trial. *Psychology of Sport and Exercise*, *50*, Article 101923. https://doi.org/10.1016/j.psychsport.2021.101923.

Wagner, I. (2016). Wissensvermittlung als Ziel des Sportunterrichts: eine Forderung in aktuellen Sportlehrplänen? [Knowledge transfer as a goal of physical education: a demand in current physical education curricula?]. In G. Stibbe (Ed.), *Lehrplanforschung: Analysen und Befunde* (pp. 170–186). Meyer & Meyer.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrica*, *54*, 427–450. https://doi.org/10.1007/BF02294627.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit [Comparative assessment in schools—a controversal matter of course. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17–31). Beltz.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58.

Wu, M., Tam, H. P., & Jen, T.-H. (2017). *Educational measurement for applied researchers: theory into practice*. Springer. https://doi.org/10.1007/978-981-10-3302-5.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of educational measurement*, *30*, 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x.

Zhu, W., Safrit, M. J., & Cohen, A. S. (1999). *Fitsmart test user manual: high school edition*. Human Kinetics.