# Pilot study of a didactic test in physical education

**Rostislav Havel[1], Martin Komarc[2]**

**[1]Univerzita Karlova, Fakulta tělesné výchovy a sportu, Střední odborná škola sociální u Matky Boží Jihlava**

**[2]Univerzita Karlova, Fakulta tělesné výchovy a sportu**

**Abstract:**

**The study focuses on the pilot phase of didactic test development in physical education. The knowledge test contains 28 closed questions (one correct and 3 distractors) and is based on the requirements of the Standard for Primary Education. The results of the analysis are presented on two pilot forms of the knowledge test. 162 primary school students in the 9th grade of primary school participated in the pilot testing. A three-parametric 3PL model was used to analyse the IRT (Item Response Theory) to obtain the difficulty, discrimination, and pseudo-guessability values for each item. After eliminating problematic items for version, A and version B of the test, satisfactory fits for the single-factor model were obtained through exploratory factor analysis. Both tested versions show the greatest information gain for students with average ability. Internal consistency was assessed by Cronbach's alpha and iterative item reliability analysis (Cronbach-Mesbach curve). The internal consistency of Cronbach's alpha was 0.835 in the test of version A and 0.755 in the test of version B.**

*Keywords: task, piloting, three-parameter model, knowledge*

The current standards of primary education in the Czech Republic specify the expected outcomes of the relevant educational field. These standards are called minimum target evaluation standards. Several diagnostic tools can be used in the Czech Republic for the verification of the Standard for Primary Education in Physical Education (PE), although they are not directly designed for this purpose. Various types of assessment scales are used to evaluate movement skills and activities in physical education (Jansa et al., 1990; Svoboda, 2007; Jansa et al., 2014). In the Czech Republic, there is currently a seemingly sufficient number of test batteries assessing physical fitness and motor performance. According to Rubín, Suchomel & Kuper (2014), the following five standardised test systems can be used in the Czech context: EUROFIT, FITNESSGRAM, INDARES, OVOV, and UNIFITTEST. The attitude questionnaire "DIPO" was developed at the Faculty of Physical Education and Sport at Charles University and is still used in physical education today. The questionnaire has been used in some research investigations presented by Kostka et al. (1987); Jansa & Perič (1994); Jansa & Dašková (2005) and Hruška (2005). Attitudes towards physical activity among boys of younger school age were investigated by Holický, Kaplan & Honsová (2014), who used the CATPA/Grade Year 3 questionnaire (Schultz et al., 1985) standardised for Czech conditions

by Kaplan (2001). The "DEMOR" questionnaire (Svoboda, 1998) in turn identifies emotional expressions in school physical education. Monitoring and analysing the structure of physical activity of primary school students is carried out using a pedometer with recording in a recording sheet (Sigmund, Lokvencová, & Mitáš, 2007; Sigmundová et al., 2014; Homolka, 2015).

The greatest absence is found in the case of a diagnostic tool for verifying knowledge of physical education, which could be used in the case of evaluation of PE standards at the end of compulsory schooling. A study by Vašíčková et al. (2009) investigated the relationship between knowledge about physical activity and physical activity performed, but in high school students. Vašíčková, Neuls & Frömel (2010) presented a health and physical activity knowledge test to students in 10 secondary schools in the Czech Republic. A comparison of the level of knowledge about health and physical activity among students studying physical education at four Czech universities was the subject of a pilot study by Vašíčková et al. (2010). The evaluation of the level of the curriculum acquired by primary school pupils in the subject Health Education was part of a larger research by Hřivnová (2018). Abroad, the assessment of the knowledge component is part of test systems such as PE Metrics (USA), CAPL (Canada), Športna vzgoja preizkus znanja (Slovenia), Sport und Gesundheit (Germany).

In the Czech Republic, there is no diagnostic tool applicable for the evaluation of the knowledge component of the proposed physical education standard, as there are no tests aimed at verifying knowledge in physical education at primary school. For this reason, the intention of the research project "Standardisation of the Evaluation Tool for Verification of the Primary Education Standard in Physical Education" is to develop a psychometrically sound test that will focus on the diagnosis of this issue.

Item Response Theory (IRT) is used in test development, which is currently the most preferred approach used in performance and pedagogical tests. "*Opportunities for improving the quality of national testing in the Czech Republic can be seen in particular in methodological shortcomings in test development, such as the limited application of psychometric methods (e.g., Item-Response Theory, Rasch model) and the lack of ability to track the development of quality in education over time*" (ČŠI, 2013, p. 14).

## Research objective

The pilot studies aimed to test the functioning of the tasks and the knowledge (didactic) test. To identify non-conforming items and generally set appropriate test

parameters. Simultaneously, the intention was to determine the final questions for the final (standardisation) version of the test.

## Methods

*Knowledge Test*

The conducted preliminary versions of the knowledge test were constructed based on the theoretical foundations of the project "Standardisation of the Evaluation Tool for Verification of the Primary Education Standard in Physical Education", which were described in the section dealing with the development of the test. Based on this background, a version of pilot test A and a version of pilot test B were designed.

*Research Organisation*

The administration of the pilot tests was carried out through the evaluation cluster "National Inspection Evaluation of the Educational System in the Czech Republic" (NIQES) – a module of the inspection information system (*InspIS SET* | system for school testing), whose author is the Czech School Inspectorate. Data collection during the administrations proceeded in a roughly similar manner. Before testing, students received login codes and passwords to generate the test at "testy.csicr.cz". After the evaluation of the test, the details allowed students to view the results at "vysledky.csicr.cz". The allotted time to complete the test was 45 minutes.

The pilot data collection took place in the first half of June of the 2018/2019 school year with a research sample of 162 students in Year 9 of primary school (version A = 75 students; version B = 87 students). A random selection of Prague primary schools was carried out. During the pre-launch pilots, the entire class was always given the same version to reduce the administrative burden.

*Data Analytics*

All data from the tests and questionnaires were checked for the accuracy of the data read and the correctness of the assignment. The data from the knowledge test were transcribed by the principal investigator into MS Excel, then they were transferred to other programs. Specifically, the mathematical software R with the packages tpm, psych, and CMC was used.

*Three-Parameter Logistics Model – Model 3PL*

We used an item response theory (IRT) approach to analyse the test tasks. It is based on the assumption that a respondent's performance on a test question is predictable by a set of factors called latent traits. Specifically, we worked with a three-parameter logistic model (3PL). This model was chosen because the diagnostic tool (knowledge test) is a *multiple-*

*choice test*. During the test, the student can only select an answer and is not required to think. It is therefore to be expected that guessing the correct answer is a factor influencing the final answers not insignificantly. The three-parameter model determines the difficulty $b_i$ and discrimination $a_i$ for each item, and the parameter$c_i$. The parameter$c_i$ is called the pseudo-guessability parameter. It determines the lower asymptote of the question characteristic curve and the probability of an individual with a very low level of ability to answer the item correctly (Urbánek, Denglerová & Širůček,2011). This IRT model is given by the equation (Urbánek, Denglerová & Širůček, 2011):

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{D a_i(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \text{ pro i} = 1,2,\dots,n$$

*Characteristic Curve (Function) of the Item*

The Item Characteristic Curve (ICC) expresses the relationship between a respondent's latency and the probability of getting an item correct. It can be defined as a logistic function that models the relationship between a student's response to an item and his or her level on the construct measured by the test (Jelínek, Květon & Vobořil, 2011)

Edelen, & Reeve (2007) add that a three-parameter logistic (3PL) model is often used for items with dichotomous response options. This model provides a trace line (curve) that is described by the location (b) and slope (a) parameters. Parameter b (also called the threshold parameter) is the point along the ICC at which the probability of a positive response to a dichotomous item is 50%.

*Information Curve (Function) of the Test*

The information curve expresses the relationship between the cumulative contribution of all test items and respondent latency. It equals the sum of the information functions of all test items. Mathematically, the sum of information functions of all items is expressed (Urbánek, Denglerová & Širůček, 2011):

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

*Reliability*

We used the internal consistency method using Cronbach's alpha and iterative item reliability analysis (Cronbach-Mesbah curve) to estimate reliability. The calculation of the internal consistency coefficient is based on the ratio of the sum of the variances of the items to the variance of the sum of the items. For the test items, the point biserial coefficient method was also used for the calculation.

*Factor Analysis*

Generally recommended procedures were followed. Hence, factor analysis was performed to ensure uni-dimensionality, which is one of the prerequisites for the use of IRT models. On this basis, several items were excluded from further testing. Subsequently, a summary analysis of the test items was performed using the above three-parameter logistic model.

## Results

The results of the analysis are presented on two pilot forms of the knowledge test (version A, version B), which were constructed after the pre-survey and the creation of the task bank. We first examine the reliability and factor analysis to test the suitability of the data for item analysis through the IRT model.

*Reliability*

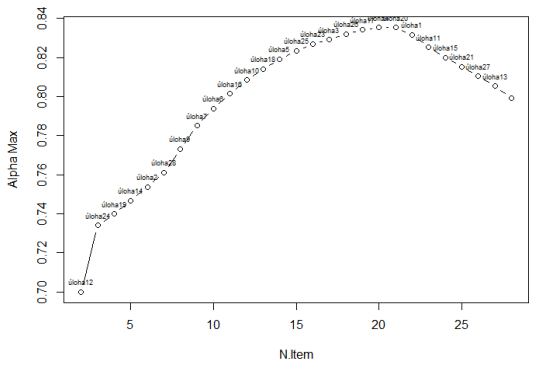


**Figure 1 Cronbach-Mesbach curve of test A**

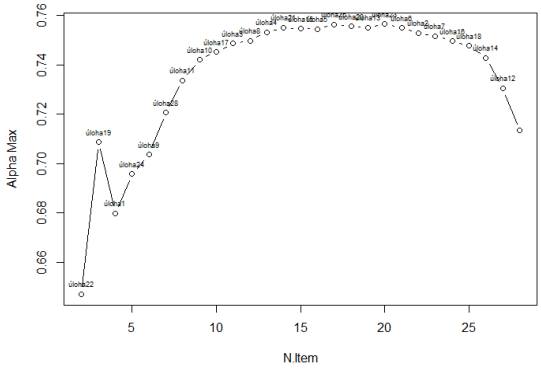


**Figure 2 Cronbach-Mesbach curve of test B**

In Figure 1 and 2 we can see that the tasks on the right – descending side of the curve reduce the reliability of the test. Excluding them would slightly increase the overall reliability. This fact obtained from the Cronbach-Mesbach curve is confirmed by the other procedures chosen to estimate reliability. Another key issue is to assess the contribution of each item to the internal reliability of the test. Here we take into account the Cronbach's alpha coefficient with the omission of one item at a time ($a_{-j}$ – Cronbach's alpha) and the point biserial correlation coefficient (*r*). It can be noted from Table 1 that some items do not contribute to internal reliability. The highlighted tasks of version A have low correlation with the rest of the scale; excluding them would slightly increase the overall reliability. Similarly, the highlighted version B tasks have a low correlation with the rest of the scale and excluding them would slightly increase the overall reliability. Items 12 and 14 have a negative correlation with the

rest of the test. Table 1 shows the test items that were discarded in the subsequent factor analysis.

**Table 1 Cronbach's alpha and point biserial correlation of test items**

| Version A | $a_{-j}$ | *r* | Version B | $a_{-j}$ | *r* |
|---|---|---|---|---|---|
| Task 1 | 0.802 | 0.140 | Task 1 | 0.697 | 0.355 |
| Task 2 | 0.793 | 0.324 | Task 2 | 0.711 | 0.176 |
| Task 3 | 0.795 | 0.277 | Task 3 | 0.706 | 0.239 |
| Task 4 | 0.794 | 0.310 | Task 4 | 0.704 | 0.273 |
| Task 5 | 0.796 | 0.268 | Task 5 | 0.712 | 0.147 |
| Task 6 | 0.790 | 0.392 | Task 6 | 0.712 | 0.165 |
| Task 7 | 0.789 | 0.445 | Task 7 | 0.713 | 0.108 |
| Task 8 | 0.785 | 0.586 | Task 8 | 0.707 | 0.232 |
| Task 9 | 0.788 | 0.431 | Task 9 | 0.689 | 0.447 |
| Task 10 | 0.792 | 0.362 | Task 10 | 0.707 | 0.234 |
| Task 11 | 0.802 | 0.118 | Task 11 | 0.702 | 0.303 |
| Task 12 | 0.786 | 0.476 | Task 12 | 0.731 | -0.122 |
| Task 13 | 0.805 | 0.027 | Task 13 | 0.709 | 0.196 |
| Task 14 | 0.781 | 0.583 | Task 14 | 0.726 | -0.019 |
| Task 15 | 0.802 | 0.127 | Task 15 | 0.703 | 0.278 |
| Task 16 | 0.791 | 0.367 | Task 16 | 0.715 | 0.127 |
| Task 17 | 0.794 | 0.291 | Task 17 | 0.705 | 0.309 |
| Task 18 | 0.798 | 0.204 | Task 18 | 0.719 | 0.038 |
| Task 19 | 0.788 | 0.427 | Task 19 | 0.691 | 0.418 |
| Task 20 | 0.800 | 0.175 | Task 20 | 0.707 | 0.233 |
| Task 21 | 0.802 | 0.120 | Task 21 | 0.706 | 0.245 |
| Task 22 | 0.787 | 0.535 | Task 22 | 0.693 | 0.427 |
| Task 23 | 0.791 | 0.376 | Task 23 | 0.711 | 0.152 |
| Task 24 | 0.788 | 0.454 | Task 24 | 0.689 | 0.493 |
| Task 25 | 0.789 | 0.415 | Task 25 | 0.708 | 0.220 |
| Task 26 | 0.796 | 0.263 | Task 26 | 0.695 | 0.412 |
| Task 27 | 0.803 | 0.102 | Task 27 | 0.690 | 0.447 |
| Task 28 | 0.786 | 0.473 | Task 28 | 0.700 | 0.330 |

For version A, 7 items with inappropriate values (Task 1, Task 11, Task 13, Task 15, Task 20, Task 21, Task 27) were discarded. In the initial reduction, there were also 7 failing items in version B (Task 5, Task 7, Task 12, Task 14, Task 16, Task 18 and Task 23). In the next step, two more items (Task 2 and Task 6) were removed from this version of the test.

The internal consistency of Cronbach's alpha of the whole test was 0.799 in the test of version A and 0.713 in the test of version B. Subsequently, we re-analysed the reliability estimate. For both test form A and test form B, the inappropriate test items mentioned above were eliminated. Option A had 21 items and Cronbach alpha = 0.835 after omitting the non-compliant items. After elimination, 19 items remained for version B, with reliability

increasing to Cronbach alpha = 0.755. Thus, in both cases, there was an increase in reliability. This is consistent with the assumption of univariate Cronbach's alpha as an estimate of internal consistency. Table 1 shows that most of the Cronbach alpha values of the test items are around the value of the full test or differ by a hundredth. For test A, the values range from 0.781 to 0.805. Test B generally has lower values and individual items range from 0.689 to 0.731.

*Exploratory Factor Analysis*

We used exploratory factor analysis with one to 3 factors. To obtain a good fit, seven items that were identifiable as problematic were excluded for version A. The tables also show the comparison of the models (Table 2, 3, 4, 5, 6).

**Table 2 Exploratory factor analysis values Version A – full scale**

| Alpha | Omega | | Chi | df | p | RMSEA | CFI | TLI |
|---|---|---|---|---|---|---|---|---|
| **0.799** | | 1 factor | 411.996 | 350 | 0.0125 | 0.049 | 0.831 | 0.817 |
| | | 2 factor | 372.037 | 323 | 0.0311 | 0.045 | 0.866 | 0.843 |
| | | 3 factor | 329.568 | 297 | 0.0938 | 0.038 | 0.911 | 0.887 |

**Table 3 Exploratory factor analysis values Version A – 7 items excluded**

| Alpha | Omega | | Chi | df | p | RMSEA | CFI | TLI |
|---|---|---|---|---|---|---|---|---|
| **0.835** | 0.93 | 1 factor | 208.966 | 189 | 0.1522 | 0.038 | 0.95 | 0.944 |
| | | 2 factor | 180.715 | 169 | 0.2549 | 0.03 | 0.97 | 0.963 |
| | | 3 factor | 152.973 | 150 | 0.4172 | 0.016 | 0.993 | 0.99 |

**Table 4 Exploratory factor analysis values Version B – full scale**

| Alpha | Omega | | Chi | df | p | RMSEA | CFI | TLI |
|---|---|---|---|---|---|---|---|---|
| **0.713** | | 1 factor | 389.812 | 350 | 0.0699 | 0.036 | 0.803 | 0.787 |
| | | 2 factor | 345.391 | 323 | 0.1874 | 0.028 | 0.889 | 0.87 |
| | | 3 factor | 304.957 | 297 | 0.3628 | 0.018 | 0.961 | 0.95 |

**Table 5 Exploratory factor analysis values Version B – items excluded 1**

| Alpha | Omega | | Chi | df | p | RMSEA | CFI | TLI |
|---|---|---|---|---|---|---|---|---|
| **0.75** | | 1 factor | 212.018 | 189 | 0.1204 | 0.037 | 0.895 | 0.883 |
| | | 2 factor | 175.886 | 169 | 0.3425 | 0.022 | 0.968 | 0.961 |
| | | 3 factor | 147.712 | 150 | 0.5375 | 0 | 1 | 1.015 |

**Table 6 Exploratory factor analysis values Version B – excluded items 2**

| Alpha | Omega | | Chi | df | p | RMSEA | CFI | TLI |
|---|---|---|---|---|---|---|---|---|
| **0.755** | 0.883 | 1 factor | 156.444 | 152 | 0.3857 | 0.018 | 0.979 | 0.976 |
| | | 2 factor | 129.906 | 134 | 0.5839 | 0 | 1 | 1.025 |
| | | 3 factor | 108.982 | 117 | 0.6889 | 0 | 1 | 1.055 |

For example, we ask whether a 3-factor model is better than a 2-factor model or whether a 2-factor model is better than a 1-factor model. The results of the factor analysis for version B

show that ideally 9 items would need to be excluded for an excellent fit. Seven test items were eliminated first and two more items were eliminated in the next stage. A p-value less than 0.05 means that statistically a more complex model is a better description of the data. Specifically, using all items, 2 factors are better than one and 3 are better than 2. However, after excluding problematic items, there is no relationship. Thus, the difference between the 2-factor model and the 1 factor model, and this is evidence that 1 factor is sufficient to describe the data. Both versions are therefore entered to be reasonably univariate and satisfactorily reliable based on exploratory factor analyses and after removing the worst discriminating items. (Omega version A 0.93 and Omega version B 0.883). This results in a satisfactory model for subsequent analysis by IRT.

*IRT Item Analysis*

Table 7 provides an analysis of the items in both forms of the test. The above mentioned problematic test items were excluded from the three-parameter IRT model analysis. The three-parameter IRT model worked with the following estimates: initial estimates for the Guess parameter of all items were ¼ (0.25), initial estimates for the Difficulty parameter of all items were 0, and initial estimates for the Discrimination parameter of all items were 1. The estimation of the Difficulty and Discrimination parameters was subsequently unconstrained.

The most difficult test item is task B13. In this task, students were asked to choose an answer for what the athlete should do with the length of his run if he has already stepped over the rebound line twice to a length of one foot. In contrast, the item with the lowest difficulty value is item B17, which asks students about the nature of fair play. In terms of discriminatory power (discrimination), none of the items analysed show unsatisfactory parameters. Since there are none with zero value or negative. Guessing parameter shows inappropriate values for some tasks. If the correct answer is evident even to a proband with a complete absence of the latent trait, the parameter c approaches a value of one. Task B 10 is identified as the easiest item to guess. A completely unfamiliar student will answer it correctly with a probability of 0.8. This is a task from the basic rules of cycling safety (On a public road, a maximum of? a) three cyclists can ride side by side; b) four cyclists; c) two cyclists; d) must not ride side by side, but behind each other.

**Table 7 IRT analysis of test items**

| Item | Guess | Difficulty | Discrimination |
|---|---|---|---|
| **A2** | 0.000 | -0.116 | 1.097 |
| **A3** | 0.000 | -0.751 | 0.653 |
| **A4** | 0.221 | 0.872 | 1.335 |
| **A5** | 0.000 | -1.034 | 0.810 |
| **A6** | 0.443 | -0.177 | 6.000 |
| **A7** | 0.000 | -1.586 | 1.359 |
| **A8** | 0.000 | -1.209 | 3.791 |
| **A9** | 0.000 | 0.233 | 1.212 |
| **A10** | 0.548 | 0.057 | 6.000 |
| **A12** | 0.177 | -0.470 | 2.598 |
| **A14** | 0.000 | -0.612 | 1.679 |
| **A16** | 0.000 | 0.009 | 1.226 |
| **A17** | 0.000 | -1.810 | 0.871 |
| **A18** | 0.000 | -1.953 | 1.084 |
| **A19** | 0.000 | 0.163 | 1.560 |
| **A22** | 0.157 | -1.136 | 2.882 |
| **A23** | 0.000 | -0.641 | 0.783 |
| **A24** | 0.000 | -1.040 | 1.859 |
| **A25** | 0.183 | 0.486 | 2.744 |
| **A26** | 0.000 | 0.368 | 0.806 |
| **A28** | 0.000 | -0.669 | 1.123 |
| **B1** | 0.034 | -0.487 | 1.041 |
| **B3** | 0.223 | -0.370 | 1.012 |
| **B4** | 0.411 | -1.480 | 0.761 |
| **B8** | 0.040 | 0.227 | 0.669 |
| **B9** | 0.039 | 0.037 | 1.295 |
| **B10** | 0.799 | -0.280 | 4.384 |
| **B11** | 0.592 | 0.102 | 3.786 |
| **B13** | 0.019 | 1.507 | 0.561 |
| **B15** | 0.474 | 0.617 | 2.766 |
| **B17** | 0.350 | -1.977 | 1.518 |
| **B19** | 0.168 | 0.423 | 2.449 |
| **B20** | 0.112 | 1.297 | 0.841 |
| **B21** | 0.340 | 0.719 | 2.140 |
| **B22** | 0.007 | -1.160 | 1.677 |
| **B24** | 0.121 | -1.094 | 1.449 |
| **B25** | 0.126 | 0.618 | 0.626 |
| **B26** | 0.265 | -0.744 | 3.761 |
| **B27** | 0.102 | -0.546 | 2.077 |
| **B28** | 0.469 | -0.127 | 4.040 |

*Characteristic Curves of Items*

In general, the characteristic curves of the form A test have a better course and a more unified shape (Figure 3). The sparsity of the parameter c is not as large as in the case of the

curves for test form B (Figure 4). The graph indicates that in the case of item B10, the pseudo-randomness is almost 0.8. Thus, a student has approximately an 80% chance of succeeding in the item even though he or she has almost no knowledge of the subject. The choice of the correct answer is probably largely a matter of guesswork rather than assumed ability. The choice of the correct answer is probably largely a matter of guesswork rather than assumed ability. We see that other items also have y-intercepts greater than zero, so even at very low ability levels there is some chance of these items being correct (by guessing). For all test items, the probability of a correct answer increases with the student's ability. Consequently, no curve has negative discriminant power.

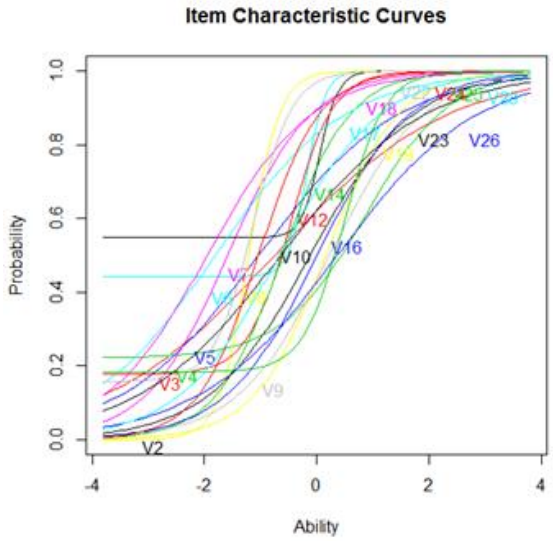


**Figure 3 Characteristic curve of version A tasks**

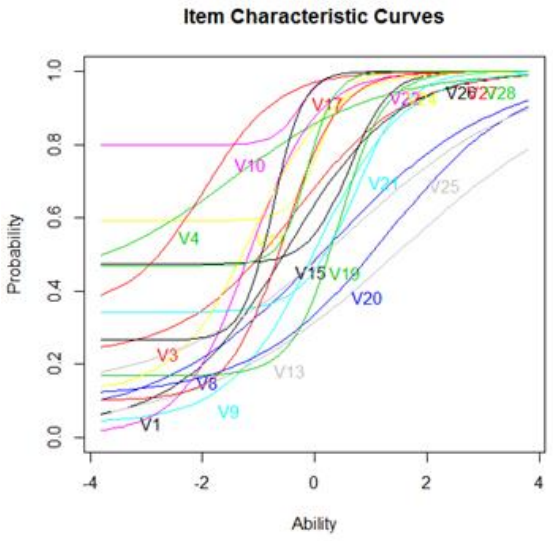


**Figure 4 Characteristic curve of version B tasks**

*Information Function of the Tests*

Figures 5 and 6 contain the information functions of both forms of the test. We see that version A has the highest information for students with average ability. Even version B is most informative for students with average ability. While for version A, there is a significant peak of higher information function even for students with lower levels of latency (ability). This phenomenon may to some extent be due to the fact that the tests are constructed and aimed at verifying a minimum standard of PE.

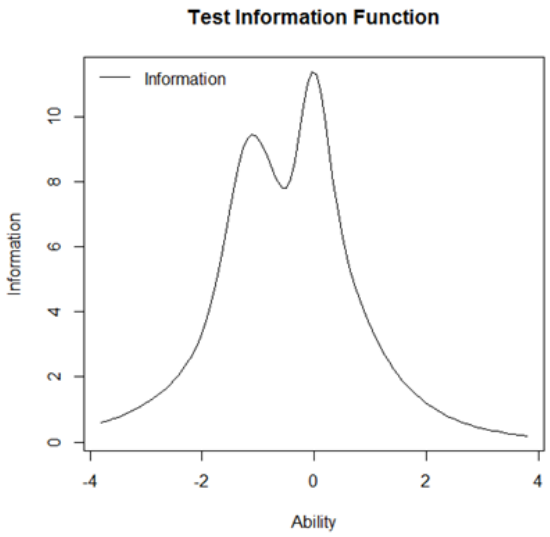


**Figure 5 Information function of version A**

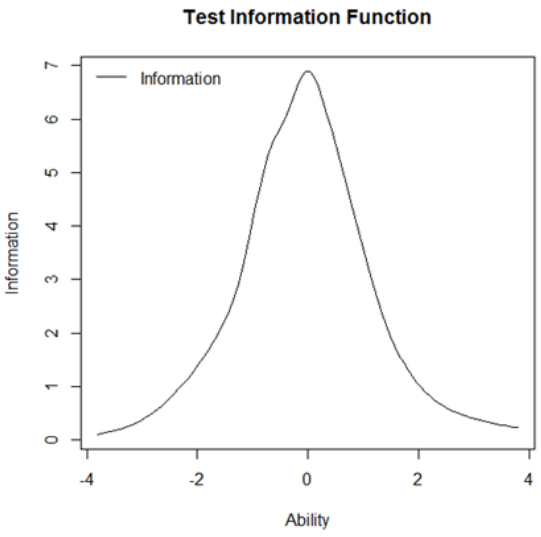


**Figure 6 Information function of version B**

## Discussion

During the pilot studies, reliability was estimated in several different ways. The reliability of the whole test was estimated through Cronbach's alpha. Cronbach's alpha is the most commonly used reliability estimation method for knowledge tests (Martínková & Vlčková, 2014). Green (2013) reports acceptable values for Cronbach's alpha above 0.7. The values of version A (0.835) and B (0.755) of our test can therefore be regarded as satisfactory. The reliability values of the overall Cronbach's alpha are appropriate to compare with the values of individual test items (Hrbáčková & Boháček, 2013). This type of analysis was also conducted in this research.

To estimate the sensitivity of an item based on its correlation with a candidate's overall test score, *a point-biserial correlation* can be used. This is a modification of the Pearson correlation coefficient for dichotomous items. The biserial correlation coefficient takes values in the interval from -1 to +1. A higher value also shares a higher proportion of variability with the overall test score (Charvát et al., 2014). Gnaldi et al. (2013) set 0.3 as the threshold. Chrástka (2016) reports a satisfactory value of the task *score biserial coefficient* in a didactic test of at least *brbis 0.2*. This criterion of whether to keep or exclude items was also chosen for our knowledge tasks.

The monotonically increasing CMC (Cronbach-Mesbah Curve) is interpreted as evidence of the assumed non-dimensional solution (Marko, 2016). By Cameletti and Caviezel (2012), this graphical tool is seen as a simple and effective method that can be used to check the uni-dimensionality of measurement scales adopted not only in psychological or social science research. Among other things, this method helped us to uncover items that reduce the

reliability of the test. Cígler et al. (2016) report that reliability estimation is usually underestimated for multidimensional data. In fact, the strict condition of uni-dimensionality for the use of the IRT model cannot always be met; therefore, according to Jelínek & Květoň, and Vobořil (2011), it is sufficient to ensure the presence of one dominant factor, and thus the test is considered unidimensional.

The chi-square statistic responds very sensitively to model complexity (the higher the number of model parameters, the better the fit regardless of model quality) and to ensemble size (the larger the ensemble, the more likely the model will be rejected) (Koudelková, 2007). It is recommended to assess the fit of the model in terms of multiple indices, as each of them evaluates the fit from a different angle (Štochl, 2005). As such, it is not appropriate to consider the values obtained in isolation. By testing the univariate model, the regression coefficients can also be considered as coefficients of factor validity (Koudelková, 2007). According to Browne et al. (1993) and McDonald (2002) RMSEA values less than 0.05 indicate very good fit, values between 0.05 and 0.08 indicate good fit, and values greater than 0.08 and approaching 0.1 indicate only average fit. The RMSEA values range from 0.018 to 0.037, indicating a good fit from this perspective. The CFI/TLI value should ideally be above 0.95. For well-functioning models, a p-value above 0.05 is desired. Both versions are therefore entered to be reasonably univariate and satisfactorily reliable based on exploratory factor analyses and after removing the worst discriminating items. (Omega version A 0.93 and Omega version B 0.883). This results in a satisfactory model for subsequent analysis by IRT.

The difficulty parameter of the IRT model usually takes values in the interval from -3 to 3. The more difficult the item, the higher the b*i* value. This suggests that the analysed test items are rather among the simpler ones. This corresponds to the focus of the tasks on the 3minimum level of the physical education standard. Hence, manageable for most students at the end of the primary stage. An item with zero discrimination, or close to it, is undesirable because it provides no information about latency and may unnecessarily prolong the test. If respondents with lower latency are more successful in the item, discriminative ability is negative. It is not important for the estimation of latency whether the discrimination is positive or negative. The latency estimation method in the three-parameter model can extract information even from items with negative discrimination. However, negatively discriminating items are discarded from performance tests, as there is clearly something wrong with them when the probability of a correct answer decreases with increasing ability.

Gnaldi et al. (2013) expect the discrimination parameters to be positive and greater than 0.7. In this respect, A3 (0.65), B8 (0.67), B13 (0.56) and B25 (0.63) are on the borderline

of acceptability. If the item cannot be guessed, then the pseudo-guessability parameter is 0. For fifteen test items, the parameter has a value of 0, which to some extent indicates the use of high-quality distractors. Such tasks provide more information about the proband's latent trait level. To show how the probability of answering a particular item correctly depends on the ability of the examinee, we construct an item characteristic curve (ICC). An ICC and a well-designed task should usually have the shape of a normal ogive (Jelínek, Květon & Vobořil, 2011; Urbánek, Denglerová & Širůček, 2011). Given these parameters, it will still be necessary to consider the use of some items in future test designs. Alternatively, an analysis and modification of the distractors of these items. The information curves demonstrate that one of the test parameters has been met. Since the knowledge test measures with the greatest degree of accuracy in the middle range and slightly below average. Form A measures more accurately in the upper zone.

## Conclusion

The study demonstrated that not all test items measure a given construct with sufficient validity. Nevertheless, the core of the tool is usable including possible further modification, for example adding new more suitable items. Based on factor analysis and item response theory, we have gained important insights into the test items that will allow us to select appropriate items in the further development of a physical education knowledge test. The pilot study helped to modify or eliminate some test items. It simultaneously reveals that the IRT-based approach is a useful tool in the development of a Physical Education knowledge test, as the focus is primarily on a specific item. It allows the diagnosis and evaluation of item characteristics at different ability levels of test subjects. In the next parts of the research of this type, it will be appropriate to use, for example, the IRT model for nominal categories (Nominal Categories Model - NCM), which is a complex and comprehensive apparatus for the analysis of distractors. Item response theory methods are by no means exhaustive and other methodological approaches and mathematical-statistical methods must also be used in constructing a knowledge test, some of which were applied in this pilot research. Based on these pilot analyses, the final (standardisation) version of the knowledge test was compiled. Items with unsatisfactory parameter values are discarded or their specifications are modified.

***The research was supported by the Grant Agency of Charles University (project no. 1056718 - Standardization of the evaluation tool for verifying the standard of basic education in the field of physical education)***

**References**

1. AMERICA, S.H.A.P.E. 2019. *Pe metrics Assessing student performance using the national standards & grade-level outcomes for K-12 physical education*. Human Kinetics: Illinois.
2. BROWNE, M. 1993. Alternative ways of assessing model fit. *Testing Structural Equation Models*.
3. CAMELETTI, M., & CAVIEZEL, V. 2012. *Package'CMC': Cronbach-Mesbach curve*. CRAN.
4. CÍGLER, H., 2016. *Měření matematických schopností*. Dizertační práce. Brno: Masarykova univerzita.
5. ČŠI, 2013. *Analýza současných systémů sledování a hodnocení kvality a efektivity ve vzdělávání.* Praha: ČŠI.
6. EDELEN, M. O., & REEVE, B. B. 2007. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1), 5.
7. GNALDI, M., MATTEUCCI, M., MIGNANI, S., & FALOCCI, N. 2013. Methods of item analysis in standardized student assessment: An application to an Italian case study. *The International Journal of Educational and Psychological Assessment*, 12, 78-92.
8. GREEN, R. 2013. *Statistical analyses for language testers*. Springer.
9. HOLICKÝ, J., KAPLAN, A., HONSOVÁ, Š. 2014. Postoje k pohybovým aktivitám u chlapců v mladším školním věku. *Česká kinantropologie*, 18(1), s. 53-62.
10. HRBÁČKOVÁ, J., & BOHÁČEK, M. 2013. Praktické postřehy k významu statistické analýzy při tvorbě jazykových testů. *ACC Journal*, 3, 37-45.
11. HRUŠKA, J. 2005. *Reliabilita dotazníku DIPO – J*. Univerzita Karlova v Praze, Fakulta tělesné výchovy a sportu. Diplomová práce.
12. HŘIVNOVÁ, M. 2018. *Analýza a evaluace kurikula vzdělávacího oboru Výchova ke zdraví.* Univerzita Palackého v Olomouci, Pedagogická fakulta. Habilitační práce

13. CHARVÁT, M., VIKTOROVÁ, L., VOBOŘIL, L., TOŠENOVSKÁ, M., & OPLETALOVÁ, V., 2014. *Tvorba, administrace a analýza testů studijních předpokladů*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-4415-4
14. CHRÁSTKA M., 2016. *Metody pedagogického výzkumu: Základy kvantitativního výzkumu, 2., aktualizované vydání*. Praha: Grada Publishing a.s.
15. JANSA, P. et al. 2014. *Pedagogika sportu*. Praha: Karolinum. ISBN 978-80-246-4015-0
16. JANSA, P. & DAŠKOVÁ, B. 2005. Názory, zájmy a postoje školní mládeže na sport a tělesnou výchovu (7-15 let). In Jansa, P. et al. *Sport a pohybové aktivity životě české populace*, Praha: Univerzita Karlova v Praze.
17. JANSA, P. & PERIČ, T., 1994. Vztah dětí k tělocviku a sportu. *Sport report*, 9, 108-109.
18. JANSA, P. et al., 1990. *Škálování v tělesné výchově*. Zpráva dílčího výzkumného úkolu Praha: UK FTVS.
19. JELÍNEK, M., KVĚTON, P. & VOBOŘIL, D., 2011. *Testování v psychologii*. Grada Publishing.
20. KAPLAN, A., 2001. *Identifikace pohybově indisponovaného žáka a studium jeho role v podmínkách školní tělesné výchovy*. Univerzita Karlova v Praze, Fakulta tělesné výchovy a sportu, Disertační práce.
21. KOSTKA, V. et al. 1987. *Tělesná výchova v systému výchovy a vzdělávání na školách všech stupňů*. Praha: FTVS UK.
22. KOUDELOVÁ, A., 2007. *Kvalita života ve vztahu k pohybovým aktivitám: mezikulturní převod a validizace profilu kvality života*. Univerzita Karlova v Praze, Fakulta tělesné výchovy a sportu, Disertační práce.
23. MARTINKOVÁ, P. & VLČKOVÁ, K. 2014. Hodnocení reliability znalostních a psychologických testů. *Informační bulletin České statistické společnosti*, 4, 1–15.
24. MARKO, M., 2016. Psychometrické zhodnotenie Mokkenového modelu pre Škálu na úlohu zameraných obáv. *Psychologie a její kontexty*, 7(1), 125-134.
25. McDONALD, I., 2013. Critical social research and political intervention: Moralistic versus radical approaches. In *Power Games* (pp. 100-116). Routledge.
26. RUBÍN, L., SUCHOMEL, A. & KUPR, J., 2014. Aktuální možnosti hodnocení tělesné zdatnosti u jedinců školního věku. *Česká kinantropologie*, 18(1), 11-22.
27. URBÁNEK, T., DENGLEROVÁ, D. & ŠIRŮČEK, J., 2011. *Psychometrika: měření v psychologii*. Praha: Portál.

28. SCHULTZ, R. W., SMOLL, F.L., CARRE, F.A. & MOSHER, R.E. 1985. Inventories and Norms for Children´s Attitudes Toward Physical Activity. *Research Quarterly for exercise and sport*, 56(3), 256-265.
29. SIGMUNDOVÁ, D., SIGMUND, E., HAMŘÍK, Z. & KALMAN, M. 2014. Trends of overweight and obesity, physicalactivity and sedentary behaviour in Czech school children: HBSC study. *The European Journal of Public Health*, 24(2),210–215.
30. SIGMUND, E., LOKVENCOVÁ, P. & MITÁŠ, J., 2007. Ověření možnosti celotýdenního monitorování pohybové aktivity dětí mladšího školního věku pomocí akcelerometru a pedometru pro tvorbu a kontrolu pohybových programů. *Česká kinanropologie*, 11(4), 9-20.
31. SVOBODA, B. 2007. *Pedagogika sportu*. Praha: Karolinum.
32. ŠTOCHL, J., 2005. *Structure of motor symptoms of Parkinson´s Disease.* Univerzita Karlova v Praze, Fakulta tělesné výchovy a sportu, Disertační práce.
33. VAŠÍČKOVÁ, J., CHMELÍK, F., FROMEL, K. & NEULS, F., 2009. Vztah mezi vědomostmi o problematice pohybové aktivity a realizovanou pohybovou aktivitou u středoškolských studentů. *Tělesná kultura*, 32(2), 33–44.
34. VAŠÍČKOVÁ, J., NEULS, F. & FROMEL, K., 2010. Comprehensive test in school physical education at secondary schools in the Czech Republic - Standardization and verification. *Acta Universitatis Palackianae Olomucensis. Gymnica*, 40(4), 7-14.

## Abstrakt

### Pilotní studie didaktického testu z tělesné výchovy

**Rostislav Havel & Martin Komarc**

*Studie je zaměřena na pilotážní fázi vývoje didaktického testu z tělesné výchovy. Vědomostní test obsahuje 28 uzavřených úloh (jedna správná a 3 distraktory) a vychází z požadavků Standardu pro základní vzdělávání. Výsledky analýzy jsou prezentovány na dvou pilotních formách vědomostního testu. Pilotního ověřování se účastnilo 162 žáků 9. ročníku základní školy. Na analýzu IRT (Item Response Theory) byl použit tříparametrický model 3PL, pomocí kterého byly získány hodnoty obtížnosti (difficulty), rozlišovací schopnosti (discrimination) a pseudouhádnutelnosti jednotlivých položek. Po vyloučení problematických položek u verze testu A a verze B bylo dosaženo vyhovujících fitů pro jednofaktorový model prostřednictvím explorativní faktorové analýzy. Obě pilotované verze vykazují největší informační přínos u žáků s průměrnou schopností. Vnitřní konzistence byla vyhodnocena pomocí Cronbachova alpha α iterativní analýzou reliability položek (Cronbach-Mesbahova křivka). Vnitřní konzistence Cronbachova alfa byla 0,835 v testu varianty A a 0,755 v testu varianty B.*

*Klíčová slova: úloha, pilotáž, tříparametrový model, vědomosti*.

**PhDr. ROSTISLAV HAVEL, Ph.D., MPA (*1984) – zabývá se vědomostní složkou u standardů tělesné výchovy.**

**Mgr. MARTIN KOMARC, Ph.D., (*1983) – jeho specializací a hlavním odborným zájmem je aplikace strukturálního modelování, položkové analýzy a adaptivního testování v oblasti psychologie, vzdělávání a mediciny.**