Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) for adolescents (15-18 years) from grades 10-12: item response theory analysis of the content knowledge questionnaire

João Mota (joaorodrigues@fmh.ulisboa.pt)

Centro de Estudos de Educação, Faculdade de Motricidade Humana, Universidade de Lisboa, Estrada da Costa, Cruz-Quebrada-Dafundo, Oeiras, Portugal; UIDEF, Instituto de Educação, Universidade de Lisboa, Alameda da Universidade, Lisbon, Portugal

João Martins

Centro de Estudos de Educação, Faculdade de Motricidade Humana, Universidade de Lisboa, Estrada da Costa, Cruz-Quebrada-Dafundo, Oeiras, Portugal; UIDEF, Instituto de Educação, Universidade de Lisboa, Alameda da Universidade, Lisbon, Portugal

Marcos Onofre

Centro de Estudos de Educação, Faculdade de Motricidade Humana, Universidade de Lisboa, Estrada da Costa, Cruz-Quebrada-Dafundo, Oeiras, Portugal; UIDEF, Instituto de Educação, Universidade de Lisboa, Alameda da Universidade, Lisbon, Portugal

Research Article

Keywords: physical literacy, assessment, physical education, construct validity, reliability, high-school, adolescence

Posted Date: March 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1458688/v2

Abstract

Background: Aims of this study were to assess construct validity (dimensionality and measurement invariance) and reliability of the previously developed Cognitive module of the Portuguese Physical Literacy Assessment – Questionnaire (PPLA-Q). Secondary aims were to assess whether using distractor information was useful for higher precision, and whether a total sum-score has enough precision for applied PE settings.

Methods: Parametric Item Response Theory (IRT) models were estimated using a final sample of 508 Portuguese adolescents (M_{age}= 16, SD = 1 years) studying in public schools in Lisbon. A retest subsample of 73 students, collected 15 days after baseline, was used to calculate Intraclass Correlation Coefficient (ICC) and Svenson's ordinal paired agreement.

Results: A mixed 2-parameter nested logit + graded response model provided the best fit to the data, C2 (21) = 23.92, p = .21; CFI = .98; $RMSEA_{C2}$= .017 [0,.043] with no misfitting items. Modelling distractor information provided an increase in available information and thus, reliability. There was evidence of differential item functioning in one item in favor of male students, however it did not translate in statistically significant differences at test level (sDTF = -0.06; sDTF% = -0.14). Average score reliability was low (marginal reliability= .60); while adequate reliability was attained in the -2 to -1 θ range. ICC results suggest poor to moderate test-retest reliability (ICC = .56, [.38, .70]); while Svenson's method resulted in 6 out of 10 items with acceptable agreement (>.70), and 4 remaining items revealing a small individual variability across time points. We found a high correlation (r = .91 [.90,.93]) among sum-score and scores derived from calibrated mixed model.

Conclusions: Evidence supports the construct validity of the cognitive module of the PPLA-Q to assess *Content Knowledge* in the Portuguese PE context for grade 10-12 (15-18 years) adolescents. This test attainted acceptable reliability for distinguishing student with transitional knowledge (between *Foundation* and *Mastery*), with further revisions needed to target full spectrum of θ. Its sum-score might be used in applied settings to get a quick overview of student's knowledge; for precision IRT score is recommended. Further scrutiny of test-retest reliability is warranted in future research, along with the use of 3-parameter logistic models.

Background

Physical literacy corresponds to the skills and attributes that individuals demonstrate through physical activity (PA) and movement throughout their lives, enabling them to lead healthy and meaningful lifestyles (Physical Literacy for Life, 2021). It is a key competence that can be developed during quality physical education (PE) (UNESCO, 2015). Despite the wide array of specific definitions available in the literature (Edwards et al., 2017; Martins et al., 2020), all integrate a reference to some form of knowledge and understanding (Whitehead, 2010) or content knowledge (Sport Australia, 2019). Similarly, learning outcomes related to knowledge pertaining to PA and movement contexts are imbued into the PE

curriculum of many countries (e.g., Society of Health and Physical Educators (SHAPE) America, 2014), including Portugal (Ministério da Educação [Ministry of Education], 2001, 2018).

Relevancy of this knowledge is backed by evidence of positive association of knowledge of PA guidelines (World Health Organization, 2010, 2020) and health benefits, with PA participation (Abula et al., 2018; Haase et al., 2004), and physical fitness (Vaara et al., 2019). Similarly, awareness of health risks related to inactivity might predict PA participation in adults (Fredriksson et al., 2018) and adolescentes (Xu et al., 2017). Despite the posited benefits and inclusion in the PE syllabus, knowledge of these contents in Portugal is suggested to be low: both in school-age students (Marques et al., 2015) and young adults (Martins et al., 2019).

Few options exist to assess this type of knowledge in an integrated manner within a PL framework (Essiet et al., 2021; Shearer et al., 2021), and to our knowledge none exists for its direct measurement in adolescents. For this purpose, we previously developed the cognitive module of the Portuguese Physical Literacy Assessment – Questionnaire (PPLA-Q; Mota et al., 2021), part of the larger PPLA framework designed to assess PL in Portuguese PE for adolescents aged 15-18 (grade 10-12). This module is a test inspired by the Australian PL Framework (Sport Australia, 2019) element of *Content Knowledge* and is directly tied to the outcomes of the Portuguese PE syllabus. Previous work has been done on its preliminary validity and reliability testing (Mota et al., 2021), but gathering of robust evidence supporting its construct validity and reliability is needed before its scores are used for its intended use of informing teacher's practice, and provide feedback to students (American Educational Research Association et al., 2014). Within the educational arena, two main theories of assessment can be used for this purpose: Classical Test Theory (CTT) and Item Response Theory (IRT).

Among the differences, widely documented elsewhere (DeMars, 2010; Embretson & Reise, 2000; Hambleton et al., 1991), IRT: a) explicitly models the interaction between the item characteristics (e.g., difficulty, discrimination and guessing) and a person's latent variable (denoted by the Greek letter theta; θ) (Meijer & Tendeiro, 2018); allowing the estimation of essentially sample-independent parameters to evaluate item quality; b) opens the possibility to analyze test information (and thus reliability) at different θ ranges (Hambleton et al., 2010), which differs from the usual CTT view of a general summary of reliability (e.g., Cronbach's alpha, or McDonald's omega) for all levels across θ; c) allows the use of mixed-format tests (e.g., tests composed of both single and multiple selection items), with no unbalanced impact upon tests scores (Embretson & Reise, 2000); d) offers models to perform robust analysis of distractors in tests (e.g., R. Bock, 1972).

Recent studies have also used this family of procedures to model information contained in incorrect responses to increase the precision of measurement (Smith et al., 2020; Storme et al., 2019) by using models that were explicitly designed to acknowledge a cognitive response process with two-stages: nested logit models (NLM; Suh & Bolt, 2010).

This study is part of a series of studies to gather evidence in support of validity and reliability of the PPLA (Mota et al., 2021, 2022) – a tool composed of two instruments, the PPLA-Q, and the PPLA – Observation

tool (in development). In it, we sought to gather evidence to support construct validity (internal structure and measurement invariance across sexes) and reliability (score reliability and test-retest) of the cognitive module of the PPLA-Q (*Content Knowledge* test) through the lens of IRT. Secondary aims of this study were to assess a) whether modelling data from distractors posed an advantage in locating students in the latent continuum; b) whether the sum-score possessed enough accuracy for practical-oriented settings.

Methods

Since this study was part of a larger validation project for the PPLA, it used the same baseline and retest samples, and data collection procedures as those detailed in PPLA-Q previous study (Mota et al., 2022). As such, in the interest of parsimony, we describe only the essential details.

Participants

Main study (baseline)

The main study used a convenience sample of 521 grade 10-12 students from 6 public schools in the metropolitan Lisbon area. Recruitment was stratified by grade, and major; diversity of socioeconomic backgrounds was used as a secondary criterion. 13 students missed class on data collection day and were excluded from this analysis. The final sample (N=508) was 59% composed of female students (M_{age}= 16, SD = 1 years).

Retest study phase

A subsample of 73 students was used for retest application, 56% female, mean age 16 (1) years. This number was based on a minimum sample size of 64 participants derived from power analysis (Arifin, 2020) for an expected Intraclass Correlation Coefficient (ICC) of .80, with .10 precision in its 95% confidence interval (Bonett, 2002), adding in a 20% margin for attrition. Given the time frame of the project and COVID-19 constraints, all participants were from grade 11 of the same school and major (*Science, Technology, Engineering and Math*).

Measures

The cognitive module of the PPLA-Q is part of a questionnaire developed to assess the psychological, social and part of the cognitive domains of Physical Literacy, inspired by the Australian Physical Literacy Framework (Sport Australia, 2019) and the Portuguese PE syllabus. A previous content validity study resulted in Scale-Content Validity Indexes of .87 (average) and .60 (universal agreement), as evaluated by experts; and highlighted adequate cognitive elicitation of students (Mota et al., 2021). This module measures content knowledge in physical activity and movement settings and was comprised of 10 items: 7 single selection questions, 2 multiple selection questions, and a close-type question; these items are

subdivided into 5 different themes, with 2 items each (one designed to assess lower, foundational knowledge, the other designed to assess deeper knowledge application; Table 1).

Procedures

Main study (baseline)

PPLA-Q (version 0.6; available in Mota et al., 2021) was self-administered during PE classes from January to March 2021. Due to COVID-19 lockdown, two different data collection formats were used: initially, 3 classes (n=60) filled out the questionnaire in paper format – using two different mirrored versions of the test to reduce cheating – while the remaining 22 classes completed an online version in LimeSurvey (LimeSurvey GmbH, 2021). We used a standard initial instruction to inform participants about the questionnaire's goals, its anonymity and encourage them to provide their best effort. Mean completion time (n= 452) for the cognitive module was 8.6 (2.8) minutes.

Retest study phase

The two application of PPLA-Q were spaced 15 days apart to reduce carryover effects (Nunnaly & Bernstein, 1994). Data collection procedures were equal. Mean completion time (n=73) for the cognitive module was 6 (1.7) minutes.

Analysis

All statistical analysis used Rstudio 1.4.1106 (RStudio Team, 2020), with R 4.0.1(R Core Team, 2020). We had marginal missing data (< 1%) in each item (Table 1, column, 3). Prior to parametric IRT analysis, two datasets were derived: 1) answers coded as dichotomous (i.e., correct, or incorrect); 2) answer coded as polytomous (ordinal for items 5,6 and 10, and nominal for the remainder). Ordinal coding of multiple selection items used a penalization for each incorrect choice (i.e., -1 point) to reduce chances of students obtaining maximum score through selection of all options. The four blocks of item 6 (cloze-type) were collapsed into a single variable to suppress the possibility of local dependency between blocks; students who selected two options in each block were coded as missing (2 cases).

esponses and missing data for the PPLA-Q content knowledge test (N=508)

eme	Item (intended difficulty)	Missing (%)	Polytomous						Dichotomous		
			A	B	C	D	E	F	Incorrect	Correct[3]	Discrimination
	Item 1 (F)	0	0.4	**97.2**	2.2	0.2			2.8	97.2	.05
	Item 2 (M)	0	8.7	3.9	31.9	**55.5**			44.5	55.5	.53
d	Item 3 (F)	1 (0.2%)	6.9	**74.6**	11.2	7.3			25.4	74.6	.47
	Item 4 (M)	4 (0.8)	**49.0**	5.2	33.7	12.1			51.0	49.0	.45
d risk	Item 5 (F)[1]	0	**96.9**	60.7	**93.4**	65.2	**93.8**	**96.0**	49.4	50.6	.44
	Item 6 (M)[2]	3 (0.6%)							44.0	56.0	.56
	Block I		2.6	**91.5**	5.9						
	Block II		17.0	**74.3**	8.7						
	Block III		**73.7**	15.6	10.7						
	Block IV		**89.1**	4.0	6.9						
	Item 7 (F)	0	34.4	8.1	18.7	38.8			65.6	34.4	.40
	Item 8 (M)	0	9.6	6.5	5.1	**78.7**			21.3	78.7	.42
	Item 9 (F)	1 (0.2%)	**38.3**	20.7	32.3	8.7			61.7	38.3	.46
on											
	Item 10 (M)[1]	0	95.1	63.3	97.6	60.0	92.8		74.0	26.0	.39

ion level; M – Mastery level

t responses are bolded

election items; [2]Close-type item; [3] Equivalent to difficulty index (*p*) x 100

Model estimation

All IRT models were estimated using Maximum Marginal Likelihood and the Expectation-Maximization algorithm with the package *mirt* (Chalmers, 2012) in R 4.0.1 (R Core Team, 2020). Dichotomous models (1 and 2-parameter logistic; 1PL, 2PL) used the dichotomously coded dataset, while the polytomous models (nominal response, 2-parameter nester logit and mixed graded response; NRM, 2PNL, 2PNL + GRM) used the polytomous dataset. All models converged properly. Plots were extracted using *mirt* in conjunction with the *lattice* package (Sarkar, 2008).

Model fit and selection

Limited-information statistic C2 (Cai & Monroe, 2014) and corresponding p-value were used to assess absolute fit of each estimated model to the data. A .05 significance level was used, with non-significant p-values indicating good absolute fit. The Root Mean Square Error of Approximation based on C2 ($RMSEA_{C2}$) with a threshold of .06 (Li, 2019) was used as indicative of adequate approximate fit. Comparative Fit Index (CFI) was used as an auxiliary indicator of model fit with a tentative threshold of .95, since to our knowledge, no research has established the adequacy of this index based on the C2 statistic. Comparison between the nested 1PL and 2PL models used the likelihood-ratio test (LRT; e.g., Finch & French, 2015) based on the -2LL statistic for each model, with a significance level of .05, to

assess whether adding parameters significantly improved the fit of the model. Comparison between non-nested models (NRM, 2PNL, 2PNL + GRM) used both the Akaike informatic criterion (AIC; Akaike, 1998) and Bayesian information criterion (BIC; Schwarz, 1978), with lower values indicating better model fit.

Relative efficiency, calculated as the ratio between the total information available in a more complex model versus that of a less complex model (Finch & French, 2015; Lord, 1980), was used to assess the trade-off between information and model complexity. Item fit was assessed through the significance (p-value <.05) of the *S.X2* statistic (Orlando & Thissen, 2000, 2003) and its accompanying RMSEA. For concision's sake, item parameter estimates, and item/option characteristic curves plot are presented only for the best fitting model. Person fit was assessed through the *Zh* statistic (Drasgow et al., 1985), with a threshold of |1.96| for the final model, using *mirt*.

Score Reliability

Marginal reliability (Green et al., 1984) was used to quantify average score reliability across the θ continuum. For comparison purposes, Cronbach's alpha (α) and McDonald's omega total (ω) were computed using the *psych* package (Revelle, 2021). Thresholds of .70 and .80 were used for acceptable (Nunnaly & Bernstein, 1994), and good reliability, respectively (Price, 2017).

Assumptions

Unidimensionality was tested via estimation of a two-factor exploratory IRT model using *mirt,* and its comparison via LRT with the correspondent one-factor model (Finch & French, 2015); both its *p-value* (significant at .05, indicative of a significantly better model-data fit) and BIC (lower values representing a more parsimonious factorial structure) were used. Local independence was assessed through Q3 (Yen, 1984), using a threshold of |.20| to identify large pairwise residual correlations after accounting for θ (Chen & Thissen, 1997).

Score correlations

Estimated θ for all IRT models were computed using *expected a posteriori* (EAP; Embretson & Reise, 2000) in *mirt*. Sum-score was computed as the sum of correct responses in dichotomous format. Pearson correlation coefficient and corresponding 95%CI were estimated using the *rstatix* package (Kassambara, 2021). For comparison purposes with previous study, CTT difficulty (p) and discrimination (gULI) indexes were computed with the *ShinyItemAnalysis* package (Martinková & Drabinová, 2019).

Differential Item and Test Functioning

We analyzed differential functioning at item (DIF) and test level (DTF). DIF analysis was performed between sexes using a two-stage approach. First, a mixed-format multiple-group IRT model with no equality constraints across-groups was used as reference to run the DIF function in *mirt* – which adds, and tests via LRT, equality constraints for one item at a time, returning multiplicity-controlled (Benjamini & Hochberg, 1995) p-values. Three items with highest p-values were selected as anchors (i.e., assumed

invariant) and a final addictive sequential analysis was run on the anchored model, with freely estimated means and variances. DTF analysis were performed on the final anchored model via the sDTF statistic with 1000 draws (Chalmers et al., 2016) using the females as reference group – before this could happen, a case had to be removed to equalize the number of categories used in item 1 across groups. sDTF represents the number of points on the test, on average, that the reference group will score higher (Chalmers et al., 2016)

Test-retest reliability

To further examine test and item quality, we analyzed test-retest reliability at both levels. At test level, we computed a single rater, absolute agreement, two-way mixed effect model (formula 2.1 in Koo & Li, 2016) ICC and its 95%CI through the *irr* package (Gamer et al., 2019), using estimated θ scores derived from the final model. ICC values of .90, .75, .50 were used, respectively, as thresholds for excellent, good, and moderate scale level test-retest reliability (Koo & Li, 2016). At item level, we used Svensson's (2012) method for ordinal paired data to calculate proportions of agreement (threshold of .70), systematic variability and individual variability, using the dichotomous-scored dataset.

Results

Model fit

The 2PNL+GRM model showed the best absolute fit (C2(21) = 23.92, p = .30) and approximate fit ($RMSEA_{C2}$ = .017 [0, .043]) to the data, out of all the models (Table 2). All models displayed both adequate absolute fit (p-value > .05), and approximate fit ($RMSEA_{C2} \leq .06$, with 95%CI below this threshold). According to the LRT based on the -2LL statistic, the 2PL model fits the data better than the 1PL model ($\Delta \chi^2$ = 28.80, Δ df = 9, p = .001), and the 2PNL model fits the data better than the NRM model ($\Delta \chi^2$ = 28.79, Δ df = 0, p = <.001). Similarly, using information-based indices (AIC and BIC), the 2PNL + GRM model presents a more parsimonious fit to the data than its 2PNL counterpart (ΔAIC = - 0.091, ΔBIC= -22.03).

As expected, the amount of information offered by each of the models increased as the number of estimated parameters increased: the addition of the discrimination parameter in 2PL model offered a 9% increase in information against the 1PL, modelling data as nominal (NRM) increased the information further by 67%, a value which increase by 6% with the nested 2PNL model. With the mixed-format model (three items estimated using the GRM), the total information decreased by 3% versus all items estimated using the 2PLN (Table 2) – this decrease happens in the lower range of θ, approaching similar levels of information around -1.5 (Figure 4).

Model fit for Item Response Theory models

	C2	*df*	*p-value*	$RMSEA_{C2}$ [95% CI]	CFI	AIC / BIC	Total Information	Relative Efficiency[2]	Number of misfitting items	Marginal reliability
ious										
ter Logistic[a]	67.20	44	.01	.032 [.015, .047]	0.87	-	7.27	-	4	.49
ter Logistic[a]	45.10	35	.12	.024 [0, .042]	0.95	-	7.91	1.09	1	.54
us										
Response	21.99	15	.11	.031 [0, .056]	0.96	9853.72/ 10107.55	13.17	1.67	0	.58
ter Nested	23.84	15	.07	.034 [0, .059]	0.95	9824.93 / 10078.76	13.98	1.06	0	.61
ter Nested RM	23.92	21	.30	.017 [0, .043]	0.98	9828.28 / 10056.73	13.51	0.97	0[1]	.60

led Response Model; RMSEA – Root Mean Square Error of Approximation; CFI – Comparative Fit Index; AIC – ormation Criteria; BIC – Bayesian Information Criteria

od ratio test 1PL – 2PL model: $\Delta \chi^2 = 28.80$, $\Delta \chi^2$ df = 9, p = .001

item with p = .05, and $RMSEA_{x2}$= .03

n previous model / information of current model (Lord, 1980)

According to the S.X2 statistic, item fit was sequentially improved from the 1PL model to the 2PNL (Table 2). In the 2PNL + GRM, item 7 displayed a borderline p-value (.05), albeit with a low RMSEA value (.03). We identified 9 students whose *Zh* statistic was higher that |1.96| in the final mixed model with values ranging from -4.33 to -2.10. Analysis of their response pattern revealed that their removal would invalidate re-estimation of the mixed model (which requires 3 unique categories per item) to assess their impact on item parameters, and so we chose to keep them given their likely low impact.

Score Reliability

Parallel to the increase in information from the 1PL model to the 2PNL model, marginal reliability showed an increase throughout these models, decreasing slightly in the 2PNL + GRM model (Table 2). Conditional reliability analysis throughout different values of θ, revealed that the 2PNL model attained acceptable levels of reliability (i.e., r_{xx} = .70) from -3 to around -1, while the mixed-format model only did so from around -2 to -1 (Figure 5); the NRM model got closer to the threshold in the -3 to -2 range, providing equivalent reliability to the mixed model until θ≥0, where it underperforms comparatively to both aforementioned models. For comparison purposes, CTT reliability coefficients were estimated as Cronbach's α= .48, and McDonald's ω= .49, and were like the marginal reliability of the 1PL model.

Assumptions

Preceding item parameter interpretation, we tested the unidimensionality, and local independence assumptions of the best fitting model (2PNL+GRM). During unidimensionality assessment, an exploratory two-factor model fitted better than its one-factor counterpart (significant LRT test), at the cost of parsimony (higher BIC statistic; Table 3). Analysis of item loadings on factors revealed no interpretable pattern. Implications of this finding for interpretation of this test will be further discussed in the Discussion section. We used Yen's Q3 to check for any large violations of local independence. After controlling for θ, no items showed a pairwise residual correlation higher than |.20|. Absolute values ranged from .06 to .16 (not shown), with no discernable pattern of residual correlation among content duplets (e.g., item 1 and 2).

Table 3. Unidimensionality assumption testing for the 2-parameter nested logit model+graded response model

	AIC	BIC	$\Delta \chi^2$	Δ df	p-value
One-factor model	9828.28	10056.73	-	-	-
Two-factor model	9825.66	10092.18	20.63	9	0.01

AIC – Akaike's Information Criteria; BIC – Bayesian Information Criteria

Item parameters

For concision's sake, we display only item parameters (Table 4) and option characteristic curves (Figure 1 and Figure 2) for the 2PNL+GRM model. According to the CTT difficulty index (*p*), some items designed to be harder for the same content were found to be easier instead (item 5 and 6, and 7 and 8; Table 1); IRT parameters estimate the same relative pattern for these duplets, however, suggest that item 5 is only more difficult than item 6 at maximum score (Table 4). IRT model difficulty parameters also propose a different relative ordering of item's difficulty, with item 7 being the hardest in the test (*b*= 1.805), instead of item 10 (*p* = .26). Discrimination parameters for the correct response ranged from 0.368 (item 7) to 1.332 (item 3); as such, items with lower discrimination parameters, also display flat information trace lines (Figure 1), providing low amounts of information across the whole range of θ.

Some items modelled as 2PNL displayed flat distractor trace lines: item 7's B and C; item 1's C and D, item 2's B, item 3's A distractors were not very discriminative nor very popular (i.e., low a and γ, respectively; Figure 1 and Table 4). Distractors' order, according to their *a* parameter (De Ayala, 2009), was coherent with the theoretical correctness (i.e., based on item's content) of each distractor.

Scores correlation

The scores estimated using the 1PL model correlate perfectly with the sum-score (Table 5). Scores estimated using other models display strong, albeit decreasing, correlation with the sum-score according

to the degree of parameterization of the model (with the 2PNL being the most parameterized model, and lowest correlated, *r*= .89, [.87, .91]). There was a close to perfect correlation between the scores estimated using the 2PNL and the mixed-format 2PNL+GRM – different estimates mostly in the -1 to -2 θ range (Figure 6).

Table 4. Item parameters for the 2-parameter nested logit + graded response model

Item	2-Parameter Nested Logit									Graded Response				CTT	
	Correct response		Distractors						Distractor orrectness order[2]						
	a	b[1]	a_1	a_2	a_3	γ_1	γ_2	γ_3		b_1	b_2	b_3	b_4	Difficulty (order)	Discrimination
Item 1	0.81 (B)	-4.77 (1)	-1.53 (A)	-0.83 I	2.36 (D)	-0.74	1.69	-0.95	D > C > A					97.2 (1)	.05
Item 2	0.77 (D)	-0.33 (4)	-0.25 (A)	-0.15 (B)	0.39 (C)	-0.30	-1.02	1.32	C > B > A					55.5 (5)	.53
Item 3	1.33 (B)	-1.07 (3)	0.51 (A)	0.10 (C)	-0.61 (D)	0.20	0.44	-0.64	A > C > D					74.6 (3)	.47
Item 4	0.49 (A)	0.09 (7)	-0.59 (B)	0.50 (C)	0.09 (D)	-1.24	1.18	0.06	C > D > B					49.0 (7)	.45
Item 5	0.54									-8.85	-5.51	-3.28	-0.03 (6)[1]	50.6 (6)	.44
Item 6[3]	1.00									-2.16	-0.30 (5)[1]			56.0 (4)	.56
Item 7	0.37 (A)	1.81 (10)	0.21 (B)	-0.02 (C)	-0.19 (D)	-0.80	0.05	0.76	B > C > D					34.4 (9)	.40
Item 8	1.24 (D)	-1.36 (2)	0.83 (A)	-0.11 (B)	-0.72 (C)	1.00	-0.04	-0.95	A > B > C					78.7 (2)	.42
Item 9	0.61 (A)	0.86 (8)	-0.01 (B)	0.32 (C)	-0.31 (D)	0.17	0.67	-0.84	C > B > D					38.3 (8)	.46
Item 10	0.96									-2.37	-0.66	1.31 (9)		26.0 (10)	.39

CTT – Classical Test Theory

[1] Difficulty order

[2] Empirically implied by *a* parameters of each distractor

[3] Observed response pattern limited to 0,2 and 4 points

Note: letters in parentheses indicate the category's label

Table 5. Pearson correlations [95%CI] between estimated scores of each model

	1PL	2PL	NRM	2PNL	2PLN + GRM
2PL	.95 [.94, .96]				
NRM	.89 [.87, .90]	.93 [.92, .94]			
2PNL	.89 [.87, .91]	.92 [.90, .93]	.98 [.98, .99]		
2PLN + GRM	.91 [.90, .93]	.94 [.93, .95]	.98 [.97, .99]	.99 [.99, .99]	
Sum-score	1.00	.95 [.94, .96]	.89 [.86, .90]	.89 [.87, .91]	.91 [.90, .93]

1PL – 1-parameter logistic model; 2PL – 2-parameter logistic model; NRM – nominal response model; 2PNL – 2-parameter nested logit model; GRM – graded response model

Differential Item and Test Functioning

We found evidence of DIF in item 1 (p = 0.018, X2 (2) = 8.04). Analysis of the item parameters and OCC (Figure 7) highlighted the existence of non-uniform DIF (Finch & French, 2019) – item is easier and has lower discrimination for boys than for girls (*b* = -3.865 *versus b* = -3.028, and *a* = 0.875 *versus a* = 1.952). Also, distractor parameters suggest different functioning at distractor-level.

To analyze whether the detected DIF would translate in DTF, we calculated *sDTF* statistic (using females as the reference group). Results suggests non-existence of significant DTF (*sDTF* = -0.06 [-0.65, 0.58]; *sDTF%* = -0.14%, [-1.67, 1.49], p = .86). This would mean that, on average, boys would score an estimated 0.14% (0.06 points) higher than girls. Graphical analysis (Figure 8) shows that this is the case mostly around the -2 to -1 θ range.

Test-retest reliability

To assess the test-retest reliability of the scores estimated using the 2PNL + GRM model, we computed the Intraclass Correlation Coefficient (ICC) for two applications spaced 15 days in 73 students. There was poor to moderate/good test-retest reliability in these scores (ICC = .56, [.38, .70]; not shown) (Koo & Li, 2016). Follow-up analysis at item-level using Svensson's method scoring on dichotomously scored items, suggested that 6 items showed an acceptable percentage of agreement (>.70; Table 6). All other 4 items displayed signs of small individual variability (significant RV ranging from .04, to .11); additionally, item 4 displayed a small downwards systematic disagreement (RP < 0), while item 10 displayed a small upwards systematic disagreement (RP > 0).

Table 6. Sven'on's agreement based on ordinal paired data and Classical Test Theory difficulty at both time points (N=73)

	PA	RP [95%CI]	RV [95%CI]	Students scoring tendency			Baseline Difficulty	Retest Difficulty
				Same (n)	Down (n)	Up (n)		
Item 1	.99	-.01 [-.04, .01]	< .01 [0, 0]	72	1	0	1.00	.99
Item 2	.74	.01 [-.07, .09]	.03 [0, 0.05]	54	9	10	.55	.56
Item 3	.92	**.03 [.03, .03]**	< .01 [0, 0]	67	2	4	.92	.95
Item 4	.58	**-.01 [-.01, -.01]**	**.11 [.11, .11]**	42	16	15	.51	.49
Item 5	.73	.05 [-.06, .17]	.03 [0, .06]	53	8	12	.56	.62
Item 6	.74	**.07 [.07, .07]**	**.02 [.02, .02]**	**54**	**7**	**12**	.60	.67
Item 7	.59	.08 [-.06, .23]	**.10 [.02, .18]**	**43**	12	18	.30	.38
Item 8	.73	**-.03 [-.03, -.03]**	**.03 [.03, .03]**	**53**	11	9	.81	.78
Item 9	.60	.07 [-.07, .21]	**.09 [.02, .16]**	**44**	12	17	.44	.51
Item 10	.68	**.10 [.10, .10]**	**.04 [.04, .04]**	**50**	8	15	.32	.41

PA – Proportion of agreement; RP – Relative position; RV – Relative rank variance

Note: values in which the 95%CI does not include 0 are bolded

Discussion

We sought to gather evidence to support construct validity (internal structure and measurement invariance) and reliability (score reliability and test-retest) of the cognitive module of the PPLA-Q (content knowledge test) through the lens of IRT. Secondary aims of this study were to assess a) whether modelling data from distractors posed an advantage in locating students in the latent continuum; b) whether the sum-score possessed enough accuracy for practical-oriented settings.

Model fit

Overall, the mixed-format (2PNL+GRM) model provided the best trade-off between model fit, total information of the test, and parsimony. This model also provides more readily interpretable item parameters than the pure 2PNL for the ordinal items (De Ayala, 2009; Desjardins & Bulut, 2018) since under the latter, different discrimination parameters (category slopes) are estimated for each scoring level of the item (assumed as unordered nominal categories), which could be, in practice, constituted by different combinations of responses, and not a single discrete distractor.

Dimensionality analysis under this model suggested the existence of a possible second factor. Given the complexity and number of cognitive and personality factors at play during item response, it is usually the case that tests are not strictly unidimensional (Hambleton et al., 1991), and that the substantive consequences of a violation of this assumption must be analyzed according to the intended application of the test (Wells & Faulkner-Bond, 2016): in practice, small degrees of multidimensionality might not distort item parameters and score estimates as long as essential unidimensionality is assured (Harrison, 1986). Analysis of the residual correlation between items does not suggest any significant clustering pattern (> |.20|) between content duplets which could happen due to sampling from the same specific

subdomain (i.e., content theme). Some residual correlation (|.10 - .16|) did happen between item 2, 8 and 10 which, we surmise, could be due to similarity of the cognitive processes involved in response (i.e., analysis), or due to closer relationship between content domain for these items (energy balance, health benefits of different types of training, and body composition and its effect on health). As such, these results seem compatible with a parsimonious stance: that a single essential latent trait is being measured in grade 10 to 12 students – general content knowledge in the context of PA. Nonetheless, further studies should test this idea using other methods (e.g., bifactorial IRT modelling), as well as different stances on measurement – assuming that content knowledge could be surmised under a composite-formative model (Stadler et al., 2021).

Score reliability & correlations

Regarding reliability of the test score, both the 2PNL and mixed-format models outperformed the dichotomous models (1PL and 2PL), the nominal model (NRM) and the CTT-based estimates, as consequence of providing more information across the latent continuum. These results show that modelling information present in distractors is advantageous for estimating θ and increasing the reliability of scores, and are coherent with similar research (Storme et al. , 2019). A similar inference can be drawn from the correlation between different models.

There was a perfect correlation between 1PL-derived scores and a simple-sum score, as expected, since in 1PL model, the scores are a simple transformation of raw scores, without weights assigned to different items (Wu et al., 2016). As the parameterization increases, the correlation with sum-score is attenuated and results in differences in estimated scores, especially for students with lower knowledge.

Marginal reliability for the mixed-format model did not achieve the general acceptable threshold of .70 (Nunnaly & Bernstein, 1994), indicating that the test is still lacking on the capability to score students with desired precision across the whole range of ability. However, conditional analysis at different ranges of θ reveal that this single estimate seems to be underrepresenting reliability around the peak of test information - -2 to -1 θ – while overrepresenting the reliability in $\theta \geq 0$ (De Ayala, 2009). Taken together, this data leads to different implications regarding the intended uses of the test score (American Educational Research Association et al., 2014; Lane et al., 2015).

The sum-score might serve a purpose when a quick diagnosis and feedback to students is the chief concern since students can score their own test and detect areas of improvement with little, to no intervention from teachers. From a teacher's perspective it might also be useful to consider the raw score by content theme, allowing for specific changes to the curriculum to promote learning in these areas.

The scores derived from the 2PNL+GRM model would be better used to obtain a fined-tuned score including distractor information and measure student's knowledge around the transition point from structural knowledge (foundation level) to relational knowledge (mastery level) as the test might provide precise enough information in this range – a hypothetical student scoring all foundational items (odd numbered items) correctly would have an estimated θ of -1.21. This is specifically useful for creating class groups based on these general levels and provide appropriate learning tasks. To facilitate interpretation, we suggest a transformation so that these scores provide a 0 to 100 interpretation – like other scores in PPLA. For this transformation, the maximum obtainable θ in the test (1.591; not shown) can be used as the upper bound, and the estimated θ score for a student with the least informative response pattern (in all *least correct* distractors) as a lower bound (θ =-3.510, not shown). As such,

$$X = \frac{\theta + 3.510}{(1.591 + 3.510)} x100$$

with X being the new 0-100 score, and θ the estimated θ score.

For specific research in content knowledge about PA and healthy lifestyles, or high-stakes applications (summative assessment), the test needs further improvements so that items provide enough information across the whole spectrum of development.

One option for this would be to increase the number of items in the test, targeting higher θ ranges, as test length is related with the accuracy of its estimates (DeMars, 2010; Harrison, 1986). Some care should be taken however, as one of the emphasis of all PPLA measures during development was feasibility without compromising validity or reliability, to maximize application of the tool in PE contexts.

Another option would be to review both the plausibility and wording of flat curved distractors in items providing low amounts of information / low discrimination (items 5,7, and 9). This could lead to improved discrimination – approaching the guideline of 0.8 (De Ayala, 2009; Green et al., 1984) – by reducing guessing and confusion, and thus higher information and reliability especially for measuring higher ability students (θ>0). These choices can be further substantiated by estimating a guessing parameter (in a 3PNL model) to identify which items and distractors are more prone to guessing, and remove parameter confounding. This will, however, require a larger sample (De Ayala, 2009).

Item parameters

Regarding item's estimated difficulty versus their intended difficulty, 3 out of 5 duplets behaved as expected (i.e., item evoking higher-order cognitive abilities as harder, than their lower-order counterparts) with item pairs 5 and 6, and 7 and 8 not adhering to this. In the first case, both are scored as multiple selection items, and our data suggests that item 5 is only more difficult than item 6 at maximum score (Table 4), while it is easier for intermediate scores (i.e., scoring points in the latter requires higher ability,

than in the former, except for maximum score). This could be result of higher plausibility of distractors in item 5 (selected by ~60 to 65% of respondents; Table 1), and scoring penalization to wrong selection inflating the difficulty of achieving maximum score. It is also plausible that our decisions regarding coding of multiple selection items (5 and 6) might have introduced a degree of bias in the results by restricting the range of possible combinations (i.e., 2 points in item 5 could be obtained by multiple combinations of right and wrong answers). In the future, different coding schemes might be considered and compared.

In the second's duplet case, multiple factors might be at play: a) the ability to recall information which is not used daily (i.e., recommendations for physical activity, in item 7) might be confounding the intended difficulty as students were not aware that they were going to be tested; b) despite being based on a lower-order cognitive ability (memorization), these guidelines require a specific knowledge that cannot be inferred using an understanding of biology, or general health literacy, and as such, need to be taught explicitly during PE classes. This data is in accordance with previous research (Marques et al., 2015) that suggests that Portuguese students do not know the PA guidelines for health promotion. Nonetheless, a careful look at the distractor's popularity (δ; Table 4) suggests that they seem to be aware of the guidelines for children and adolescents, while not knowing the specific ones for adults (distractor D). This implies that more attention should be dedicated to explicitly teaching these guidelines, with more emphasis on those for adults, since arguably, they will be of most importance in the near-future of high-school students.

DIF and DTF

We found evidence of non-uniform DIF according to sex in item 1, however this did not result in significant DTF. Despite the possibility that actual differences in interpretation of the item exist between sexes, there is also a possibility that this might be due to parameter inaccuracy due to sampling variability (as suggested by the magnitude of the standard errors of distractor parameters, ranging from 62.866 to 94.708; not shown in tables), as there were no students with estimated θ in the difficulty range of this item (around -3). Similarly, the differential distractor functioning in this item might stem from a sparse selection of distractors – due to it being a very easy item – resulting in difficulties at estimation of distractors thresholds (Ostini et al., 2015). As such, if total score is of chief interest, the bias in scores will likely be negligible, as the sDTF statistics imply; whether if any specific inference is required at item-level, methods that account for DIF should be used, so that the suggested sex bias is minimized. Furthermore, other methods specifically designed for exploring differential distractor functioning could be used (Suh & Bolt, 2011), along with a larger sample.

Test -retest reliability

Test-retest reliability of estimated θ scores was poor to moderate (ICC = .51, [.32, .66]) (Koo & Li, 2016) over a 15 days interval. This might stem from a violation of the assumption of stability of the assessed trait that precludes the calculation of test-retest reliability (Polit, 2014), as learning between applications – either due to teacher intervention, or due to student's curiosity – is plausible; Longmuir et al. (2018) suggested as much in their assessment of a similar tool. Results from item-level analysis of agreement between the two time points lend some support to this idea. Out of the four items not achieving acceptable agreement (.70), one (item 10) was mostly due to an increase in correct responses in the second instance; despite achieving the threshold for agreement, items 3 and 6 also display a similar pattern. As for the remaining three items (4, 7 and 9), disagreement was mostly due to individual variability which could be indicative either of guessing, carelessness or low-quality of the items resulting in different understanding of the item across time points. In the future, a 3PNL model (accounting for guessing) could further improve this assertion and clarify the role of individual variability.

Strengths and limitations

To our knowledge, our study is the first to apply IRT to content knowledge in PA and healthy lifestyles. It exemplifies how applying nested logit models provides an increase in precision for estimating latent trait scores versus both a sum-score or dichotomous IRT models (1PL and 2PL), through the modelling of distractor information. It also provides an example of how to use these models to identify functional distractors. As such, use of IRT benefits the test in the short term, but also in the long term, as it opens the possibility of comparison between different versions of the cognitive module of the PPLA-Q by test-linking and equating; and adaptative testing.

Despite the pandemic context imposed by COVID-19, we recruited a diverse sample, mimicking the relative composition of grade 10 to 12 students' population in Portugal according to both grade and course major. Nonetheless, given its convenience nature, we advise caution before generalizing any findings of validity or reliability outside of this population, without further testing. A similar cautionary note should be made regarding the sample size used. Given the relative paucity of research using IRT nested logit models, no consensus on guidelines regarding sample exist. Even when referring to common-place models like the 2PL, NRM or GR, sample size recommendations vary widely across sources and seem to be dependent on various complex interaction between test length , number of response categories per item, number of parameters to estimate (De Ayala, 2009) and estimation method (Şahin & Anıl, 2017). Another factor in determining the sample size is the intended level of precision in the estimated parameters: while high-stakes testing will require larger sample sizes to attain small standard errors on estimated scores, other less demanding contexts might require smaller ones (De Ayala, 2009; Nguyen et al., 2014). As such, further testing using a larger, more representative sample should try to replicate, and improve upon our findings using a 3-parameters logistic model (3PL; Birnbaum, 1968) which accounts for the possibility of guessing. The same applies for DIF and DTF testing

Another limitation pertains to the use of test-retest reliability. This type of reliability is essentially a CTT concept, conceptualizing measurement error as a single statistic, whether IRT permits a detailed analysis of reliability at each θ point, as shown. Usage of IRT to model growth over time, or invariance over two time points would be better suited to the general framework of this study and allow for better inferences regarding adequacy of scores over time; this, however, was currently impossible to achieve due to sample size requirements.

Finally, concurrent with modification of items to improve the information available across higher ranges of knowledge, a second round of content validity with an expert panel might provide further support to the adequacy of these items to the pretended knowledge domain in grade 10 to 12 of Portuguese PE.

Conclusion

Overall, this study provides evidence for the construct validity of the cognitive module of the PPLA-Q, through the lens of a model combining a 2-parameter nested logit model and a graded response model. The test assesses content knowledge of themes related to PA and healthy lifestyles in grade 10 to 12 students (15-18 years). We have discussed the implications of different scoring models to each intended use. It has shown acceptable reliability in measuring students transitioning from foundational knowledge – based on recall and descriptive knowledge of facts – to mastery knowledge – based on analysis and relational understanding of concepts; however, its reliability to measure higher knowledge students still needs to be improved. This is a highly feasible test (9 minutes), useful to diagnose initial levels of content knowledge at the beginning of a school year and adapt learning tasks.

Improvements to the test could in practice be achieved via multiple paths: a) increasing number of items; b) improving the discrimination of items; c) review items to target different ranges of θ more appropriately. Evidence of DIF across sex groups was found in an item, with no significant effect at test level (DTF), as such, test scores can be compared across sexes. Further test-retest reliability evidence is warranted before test scores are used to assess change over time.

Declarations

Ethics approval and consent to participate

All the work was done in Portugal, as part of the doctoral project of the lead author, approved by the Ethics Council of Faculty of Human Kinetics, as well as the Portuguese Directorate-General of Education.

Before participation, a signed informed consent was required of all students, and their legal guardians (when students were minors).

Consent for publication

Use of anonymized data for scientific publication was disclosed and agreed upon in the consent form signed by all relevant parties.

Availability of data and materials

Participants of this study did not explicitly agree for their data to be shared publicly, so supporting data is not available.

Competing interests

The authors state no conflict of interest. No financial interest or benefit has arisen from the direct applications of this research.

Funding

This research work was funded by a PhD Scholarship from the University of Lisbon PhD Scholarship Program 2017, credited to the lead author.

Authors' contribution

João Mota wrote the main manuscript and prepared figures and tables as part of his PhD thesis. João Martins and Marcos Onofre actively supported the definition of the project and participated in the questionnaire development and revision along all phases (as PhD supervisors of João Mota). All authors reviewed the manuscript.

Acknowledgements

We would like to acknowledge the contribution of Filomena Araújo, invaluable to the development and refinement of this module. We reinforce our debt of gratitude to the entire R community for their selflessness and professionalism, especially Philip Chalmers, developer of the *mirt* package. The lead author would also like to thank his co-authors for their ever-present guidance and support during his PhD project.

References

Abula, K., Gröpel, P., Chen, K., & Beckmann, J. (2018). Does knowledge of physical activity recommendations increase physical activity among Chinese college students? Empirical investigations based on the transtheoretical model. *Journal of Sport and Health Science*, *7*(1), 77–82. https://doi.org/10.1016/j.jshs.2016.10.010

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

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

Arifin, W. N. (2020). *Sample size calculator*. http://wnarifin.github.io

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, Frederic M. & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–422). Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, *21*(9), 1331–1335. https://doi.org/10.1002/sim.1108

Cai, L., & Monroe, S. (2014). *A New Statistic for Evaluating Item Response Theory Models for Ordinal Data* [Technical Report]. National Center for Research on Evaluation, Standards, and Student Testing. https://files.eric.ed.gov/fulltext/ED555726.pdf

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It Might Not Make a Big DIF: Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement*, *76*(1), 114–140. https://doi.org/10.1177/0013164415584576

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285

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

DeMars, C. E. (2010). *Item response theory*. Oxford University Press.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of Educational Measurement and Psychometrics Using R.* CRC PRESS.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Edwards, L., Bryant, A., Keegan, R., Morgan, K., & Jones, A. (2017). Definitions, Foundations and Associations of Physical Literacy: A Systematic Review. *Sports Medicine*, *47*(1), 113–126. https://doi.org/10.1007/s40279-016-0560-7

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. L. Erlbaum Associates.

Essiet, I. A., Lander, N. J., Salmon, J., Duncan, M. J., Eyre, E. L. J., Ma, J., & Barnett, L. M. (2021). A systematic review of tools designed for teacher proxy-report of children's physical literacy or constituting elements. *International Journal of Behavioral Nutrition and Physical Activity*, *18*(1), 131. https://doi.org/10.1186/s12966-021-01162-3

Finch, W. H., & French, B. F. (2015). *Latent Variable Modeling with R* (0 ed.). Routledge. https://doi.org/10.4324/9781315869797

Finch, W. H., & French, B. F. (2019). *Educational and psychological measurement*. Routledge.

Fredriksson, S. V., Alley, S. J., Rebar, A. L., Hayman, M., Vandelanotte, C., & Schoeppe, S. (2018). How are different levels of knowledge about physical activity associated with physical activity behaviour in Australian adults? *PLoS ONE*, *13*(11). https://doi.org/10.1371/journal.pone.0207003

Gamer, M., Lemon, J., & Singh, I. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. (R package version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement*, *21*(4), 347–360.

Haase, A., Steptoe, A., Sallis, J. F., & Wardle, J. (2004). Leisure-time physical activity in university students from 23 countries: Associations with health beliefs, risk awareness, and national economic development. *Preventive Medicine*, *39*(1), 182–190. https://doi.org/10.1016/j.ypmed.2004.01.028

Hambleton, R. K., Linden, W. J. van der, & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21–42). Routledge, Taylor & Francis Group. https://research.utwente.nl/en/publications/irt-models-for-the-analysis-of-polytomously-scored-data-brief-and

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (pp. x, 174). Sage Publications, Inc.

Harrison, D. A. (1986). Robustness of IRT Parameter Estimation to Violations of the Unidimensionality Assumption. *Journal of Educational Statistics*, *11*(2), 91–115. https://doi.org/10.2307/1164972

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (0.7.0) [Computer software]. https://CRAN.R-project.org/package=rstatix

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2015). *Handbook of test development* (Second edition). Routledge, is an imprint of the Taylor & Francis Group, an Informa business.

Li, C. R. (2019). *Assessing the Model Fit of Multidimensional Item Response Theory Models with Polytomous Responses Using Limited-Information Statistics*. https://doi.org/10.13023/ETD.2019.006

LimeSurvey GmbH. (2021). *LimeSurvey: An Open Source survey tool*. LimeSurvey GmbH. http://www.limesurvey.org

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 (Suppl 2)*, 19–29. https://doi.org/10.1186/s12889-018-5890-y

Lord, F. M. (1980). *Applications of Item Response Theory To Practical Testing Problems*. Routledge. https://doi.org/10.4324/9780203056615

Marques, A., Martins, J., Sarmento, H., Rocha, L., & Costa, F. C. da. (2015). Do Students Know the Physical Activity Recommendations for Health Promotion? *Journal of Physical Activity and Health*, *12*(2), 253–256. https://doi.org/10.1123/jpah.2013-0228

Martinková, P., & Drabinová, A. (2019). ShinyItemAnalysis for Teaching Psychometrics and to Enforce Routine Analysis of Educational Tests. *The R Journal*, *10*(2), 503. https://doi.org/10.32614/RJ-2018-074

Martins, J., Cabral, M., Elias, C., Nelas, R., Sarmento, H., Marques, A., & Nicola, P. (2019). Physical activity recommendations for health: Knowledge and perceptions among college students. *Retos: Nuevas Tendencias En Educación Física, Deporte y Recreación*, *36*, 290–296.

Martins, J., Onofre, M., Mota, J., Murphy, C., Repond, R.-M., Vost, H., Cremosini, B., Svrdlim, A., Markovic, M., & Dudley, D. (2020). International approaches to the definition, philosophical tenets, and core elements of physical literacy: A scoping review. *PROSPECTS*. https://doi.org/10.1007/s11125-020-09466-1

Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. In *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, Vols. 1-2* (pp. 413–443). Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch15

Ministério da Educação. (2001). *Programa Nacional Educação Física: Ensino Secundário*. DES.

Ministério da Educação. (2018). *Aprendizagens Essenciais: Educação Física*. Ministério da Educação. https://www.dge.mec.pt/educacao-fisica

Mota, J., Martins, J., & Onofre, M. (2021). Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) for adolescents (15–18 years) from grades 10–12: Development, content validation and pilot testing. *BMC Public Health*, *21*(1), 2183. https://doi.org/10.1186/s12889-021-12230-5

Mota, J., Martins, J., & Onofre, M. (2022). *Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) for adolescents (15-18 years) from grades 10-12: Validity and reliability evidence of the Psychological and Social modules using Mokken Scale Analysis*. Research Square. https://doi.org/10.21203/rs.3.rs-1458709/v3

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An Introduction to Item Response Theory for Patient-Reported Outcome Measurement. *The Patient*, *7*(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Nunnaly, J., & Bernstein, I. (1994). *Psychometric Theory*. McGraw-Hill.

Orlando, M., & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Ostini, R., Finkelman, M., & Nering, M. (2015). Selecting among polytomous IRT models. In *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 285–304). Routledge/Taylor & Francis Group.

Physical Literacy for Life. (2021). *What is Physical Literacy*. https://physical-literacy.isca.org/update/36/what-is-physical-literacy-infographic

Polit, D. F. (2014). Getting serious about test–retest reliability: A critique of retest research and some recommendations. *Quality of Life Research*, *23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

Price, L. R. (2017). *Psychometric Methods Theory into Practice*. The Guilford Press.

R Core Team. (2020). *R: A language and environment for statistical compution*. R Foundation for Statistical Computing. http://www.R-project.org/

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research* (2.1.9) [Computer software]. https://CRAN.R-project.org/package=psych

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC. http://www.rstudio.com/

Şahin, A., & Anıl, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, *17*, 321–335. https://doi.org/10.12738/estp.2017.1.0270

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. Springer.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shearer, C., Goss, H. R., Boddy, L. M., Knowles, Z. R., Durden-Myers, E. J., & Foweather, L. (2021). Assessments Related to the Physical, Affective and Cognitive Domains of Physical Literacy Amongst Children Aged 7–11.9 Years: A Systematic Review. *Sports Medicine - Open*, *7*(1), 37. https://doi.org/10.1186/s40798-021-00324-8

Smith, T. I., Louis, K. J., Ricci, B. J., & Bendjilali, N. (2020). Quantitatively ranking incorrect responses to multiple-choice questions using item response theory. *Physical Review Physics Education Research*, *16*(1), 010107. https://doi.org/10.1103/PhysRevPhysEducRes.16.010107

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

Sport Australia. (2019). *Australian Physical Literacy Framework.* https://nla.gov.au/nla.obj-2341259417

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, *60*, 100832. https://doi.org/10.1016/j.newideapsych.2020.100832

Storme, M., Myszkowski, N., Baron, S., & Bernard, D. (2019). Same Test, Better Scores: Boosting the Reliability of Short Online Intelligence Recruitment Tests with Nested Logit Item Response Theory Models. *Journal of Intelligence*, *7*(3), 17. https://doi.org/10.3390/jintelligence7030017

Suh, Y., & Bolt, D. M. (2010). Nested Logit Models for Multiple-Choice Item Response Data. *Psychometrika*, *75*(3), 454–473. https://doi.org/10.1007/s11336-010-9163-7

Suh, Y., & Bolt, D. M. (2011). A Nested Logit Approach for Investigating Distractors as Causes of Differential Item Functioning: Differential Distractor Functioning. *Journal of Educational Measurement*, *48*(2), 188–205. https://doi.org/10.1111/j.1745-3984.2011.00139.x

Svensson, E. (2012). Different ranking approaches defining association and agreement measures of paired ordinal data. *Statistics in Medicine*, *31*(26), 3104–3117. https://doi.org/10.1002/sim.5382

UNESCO. (2015). *Quality Physical Education (QPE): Guidelines for policy makers*. UNESCO Publishing.

Vaara, J. P., Vasankari, T., Koski, H. J., & Kyröläinen, H. (2019). Awareness and Knowledge of Physical Activity Recommendations in Young Adult Men. *Frontiers in Public Health*, *7*. https://doi.org/10.3389/fpubh.2019.00310

Wells, C., & Faulkner-Bond, M. (Eds.). (2016). *Educational measurement: From foundations to future*. GP, Guilford Press.

Whitehead, M. (Ed.). (2010). *Physical literacy: Throughout the lifecourse* (1st ed). Routledge.

World Health Organization. (2010). *Global recommendations on physical activity for health*. WHO Press.

World Health Organization. (2020). *WHO guidelines on physical activity and sedentary behaviour*. World Health Organization. https://apps.who.int/iris/handle/10665/336656

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5

Xu, F., Wang, X., Xiang, D., Wang, Z., Ye, Q., & Ware, R. S. (2017). Awareness of knowledge and practice regarding physical activity: A population-based prospective, observational study among students in Nanjing, China. *PLoS ONE*, *12*(6). https://doi.org/10.1371/journal.pone.0179518

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Figures

Figure 1

Option Characteristic Curves for items 1-4, 7-9 (2-Parameter Nested Logit + Graded Response Model); A-D – response options

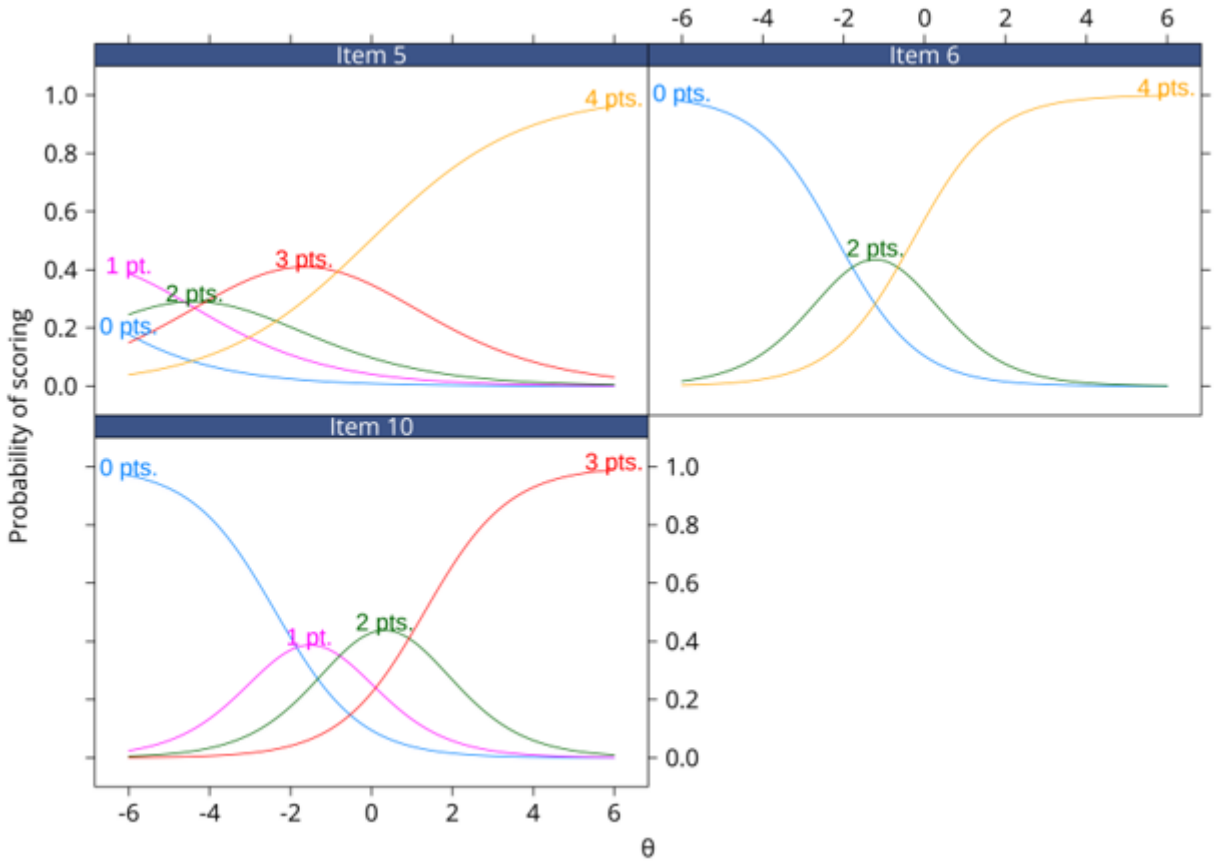

Figure 2

Option Characteristic Curves for items 5,6, 10 (2-parameter Nested Logit + Graded Response Model

Figure 3

Item information curves for the 2-Parameter Nested Logit + Graded Response Model

Figure 4

Test Information comparison between the Nominal Response Model (NRM), 2-parameter nested logit (2PLN) and 2-parameter nested logit + graded response model (2PLN+GRM)

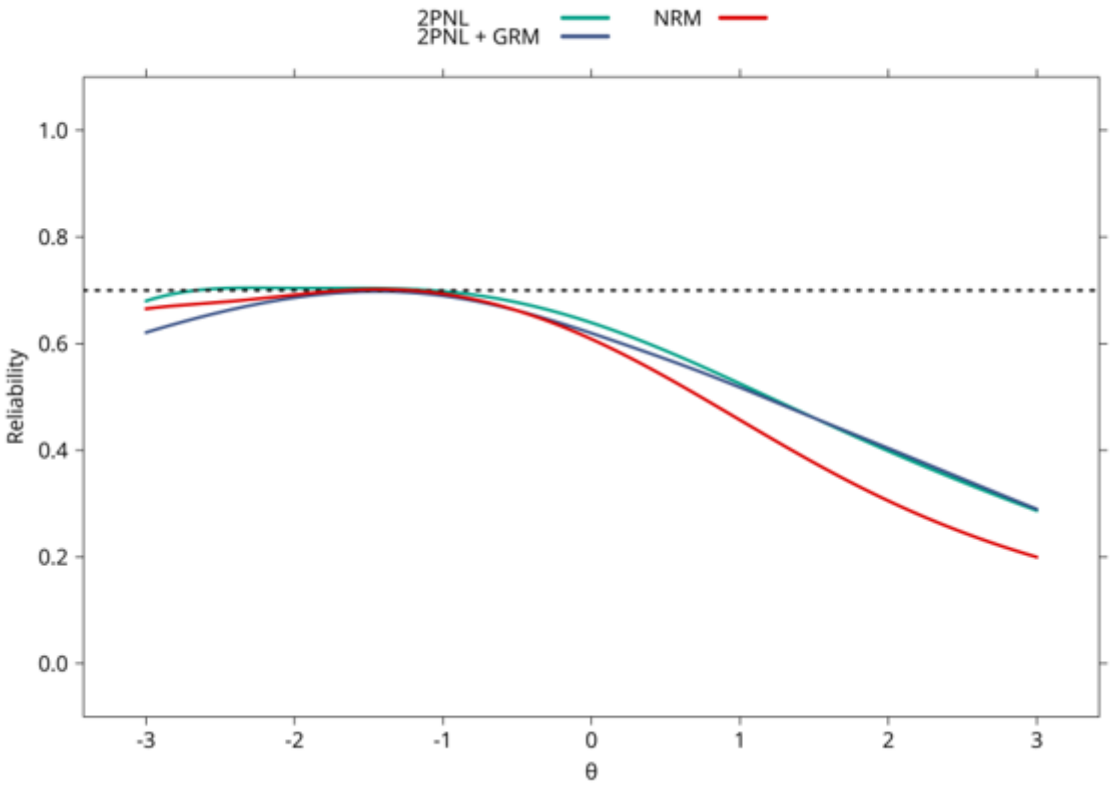

Figure 5

Conditional reliability plot comparison between the Nominal Response Model (NRM), 2-parameter nested logit (2PLN) and 2-parameter nested logit + graded response model (2PLN+GRM)

Figure 6

Scatter plot of estimated scores using 2-parameter logit model (2PNL) and 2-parameter nested logit + graded response model (2PLN+GRM)

Figure 7

Option Characteristic Curve stratified by sex for Item 1 (Differential Item Functioning)

Figure 8

Signed Differential Test Functioning pl–t - Reference group (red) Female; (black) Male