



OPEN ACCESS

EDITED BY

Alexandre Garcia-Mas,
University of the Balearic Islands, Spain

REVIEWED BY

Emily Bremer,
Acadia University, Canada
Diogo Vicente Martinho,
University of Coimbra, Portugal

*CORRESPONDENCE

João Mota
✉ joao.mota@ucc.ie

RECEIVED 22 March 2023

ACCEPTED 29 May 2023

PUBLISHED 27 June 2023

CITATION

Mota J, Martins J, Onofre M and Dudley D
(2023) Portuguese Physical Literacy Assessment
for adolescents (15–18 years): validation using
confirmatory factor and composite analyses.
Front. Sports Act. Living 5:1192025.
doi: 10.3389/fspor.2023.1192025

COPYRIGHT

© 2023 Mota, Martins, Onofre and Dudley. This
is an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Portuguese Physical Literacy Assessment for adolescents (15–18 years): validation using confirmatory factor and composite analyses

João Mota^{1,2,3*}, João Martins^{2,3}, Marcos Onofre^{2,3}
and Dean Dudley^{4,5,6}

¹Sports Studies and Physical Education Programme, School of Education, College of Arts, Celtic Studies & Social Sciences, University College Cork, Cork, Ireland, ²Centro de Estudos em Educação, Faculdade de Motricidade Humana, Universidade de Lisboa, Estrada da Costa, Oeiras, Portugal, ³UIDEF, Instituto de Educação, Universidade de Lisboa, Alameda da Universidade, Lisbon, Portugal, ⁴Macquarie School of Education, Macquarie University, Sydney, NSW, Australia, ⁵School of Human Movement and Nutrition Sciences, University of Queensland, Brisbane, QLD, Australia, ⁶Centre of Educational Measurement and Assessment, University of Sydney, Sydney, NSW, Australia

Objective: The aim of this study was to assess the construct validity and reliability of the Portuguese Physical Literacy Assessment (PPLA) instruments (a questionnaire and a tool using teacher-reported data). We also investigated the conceptual and practical implications of reflective vs. formative measurement of Physical Literacy using the PPLA.

Methods: Multiple Confirmatory Factor Analysis (CFA) and Confirmatory Composite Analysis (CCA) models were used complementarily to assess construct validity in a sample of 521 grade 10–12 Portuguese students from Lisbon, Portugal. Bifactor model-based indices (ω), Explained Common Variance (ECV), and Percentage of Uncontaminated Correlations (PUC) were used to assess score reliability and adequacy.

Results: Using CFA, an asymmetrical bifactor model (S^*1-1) provided the best fit to the data [Robust Comparative Fit Index = 97, Robust Root Mean Square Error of Approximation = 0.05 (0.04–0.06), Standardized Root Mean Square Residual (SRMR) = 0.04], while CCA resulted in the best absolute fit for single first-order composite models (d_G , d_L , and SRMR below or borderline of their 95% critical value). Through a reflective paradigm, the total PL score should not be used in isolation (ECV = 0.49, ω_H = 0.71, lower than recommended 0.80). Subscales for the Physical, Psychological, and Social domains attained acceptable reliability scores (ω_s = 0.76, 0.82, 0.80, and 0.60).

Conclusions: A general trait of PL accounts for considerable variance in all indicators. We advise calculation of a total summed PL score and domain scores, which should be interpreted conjointly in applied settings. Despite both paradigms being tenable, future research efforts should use a bifactor measurement model, which permits disentanglement of the variance attributed to the general PL trait and its domains. Overall, evidence supported the construct validity and reliability of the PPLA for its intended use as an integrated tool to measure PL as a multidimensional construct in 15- to 18-year-old Portuguese students in a physical education setting.

KEYWORDS

physical education, construct validity, reliability, high school, physical literacy, confirmatory factor analysis, confirmatory composite analysis, adolescents

Introduction

Physical literacy (PL) is a holistic concept referring to the skills and attributes that individuals demonstrate through physical activity (PA) and movement throughout their lives, enabling them to lead healthy and fulfilling lifestyles (1). This multidimensional concept is argued as the foundation for many international physical education (PE), sport, and public health agendas (2–5). In recent years, multiple efforts have been made by diverse authors and research groups based around the globe (e.g., Australia, Canada, China, France, and Iran) to develop and refine measurement instruments that assess an individual's physical literacy journey (6–13).

The Portuguese Physical Literacy Assessment (PPLA) was developed as a tool composed of two instruments (a questionnaire, PPLA-Q, and an observational instrument, PPLA-O) to be used in PE to provide a feasible and holistic assessment of adolescent's PL in grades 10–12 (15–18 years) of high school. It was inspired by the Australian Physical Literacy Framework (APLF) (14), which is a conceptual model of PL learning composed of 30 elements across four learning domains (Physical, Psychological, Social, and Cognitive). The PPLA was also informed by the outcomes and didactic philosophy of the Portuguese PE syllabus (15, 16). Previous studies using Item Response Theory models have supported construct validity and reliability of both the PPLA-Q and PPLA-O at the elemental level (17–19). However, the higher-order dimensionality of these tools requires further investigation.

Assessment of their dimensionality can be assessed through structural equation modeling (SEM) whereby two main approaches are undertaken depending on the auxiliary theories assumed to underlie measurement (20–22). In this study, they were reflective measurement and formative measurement, whereas previous studies on PL measurement have always implicitly assumed a multidimensional reflective view, modeling PL as a (a) correlated factor (7, 13, 23) or (b) higher-order factor (10, 24). However, there are ontological and conceptual issues inherent to both approaches, which are reviewed in **Supplementary Material S1**.

Our initial model for the PPLA (25) hypothesized PL as a higher-order composite formed of domain-specific composites, based on the idea of non-exchangeability of domains and indicators, along with the assertion that variation in each of the domains would be plausibly independent of each other (e.g., one could conceive that an increase in cognitive-related skills would not be simultaneous with an increase in psychological-related attitudes). Despite this, given the recency of PL construct testing, it is cogent to test alternative competing models that could further provide practical and conceptual advantages.

As such, this study sought to establish evidence supporting construct validity and reliability of the PPLA by integrating measures derived from the PPLA-Q and PPLA-O. It then compared results drawn from factor-based (reflective) methods and composite-based (formative) methods. Based on these findings, it then assessed the adequacy of using a PL total score and respective subscales. As a secondary research aim, we

investigated the implications of the different methods for the wider conceptual understanding of PL.

Materials and methods

Participants

A sample of 521 (58% female) grade 10–12 students ($M_{\text{age}} = 16$, $SD = 1$ years) from six public schools in Lisbon's metropolitan area was used (25 classes, 22 different PE teachers). Sampling procedures and full sample characteristics are detailed in prior work (17, 18). Briefly, recruitment was stratified by grade and course major according to population percentage quotas. Schools from diverse socioeconomic backgrounds were chosen to increase sample representativeness. The participant recruitment and participation flow are shown in **Figure 1**.

A minimum sample size of 275 was initially chosen based on a power analysis conducted in *WarpPLS* software (26), using the Inverse Square Root Method (27), for a minimum absolute path coefficient of 0.15 and power of 0.80. Data collection occurred between January and March 2021. PPLA-Q was self-administered (students) both in a paper and online format in the presence of the lead investigator, and PPLA-O was self-administered (PE teachers) using an online spreadsheet.

Measures

PPLA measures

Physical domain

The Physical domain of the PPLA was assessed through the PPLA-O (18). It is composed of two modules: *Movement Competence, Rules, and Tactics (MCRT)* and *Health-related fitness (HRF)*. The MCRT includes two scores—Manipulative-based activities and Stability-based activities—and these were calculated through a two-factor Graded Response Model (GRM; an Item Response Theory model). These scores summarize the general movement competence (including tactical decision and rule knowledge) of the student in physical activities, which elicit mostly manipulative movement skills (e.g., team-sports), and which elicit mostly stability movement skills (e.g., gymnastics). To facilitate interpretation, factor scores derived with Expected *A Posteriori* (EAP) scoring were transformed into a 0–100 score.

Health-related physical fitness module included seven indicators, all assessed through existing FITescola[®] protocols (28), in three major subareas: (1) *Cardiorespiratory endurance* was assessed through the 20-m Progressive Aerobic Cardiorespiratory Endurance Run (PACER), using the number of laps completed; (2) *Muscular endurance* was assessed through the number of executions in the Curl-Up and the 90° Push-Ups protocols; (3) *Flexibility* was assessed through the Backsaver Sit-and-Reach (lower body) measured in centimeters for each leg and Shoulder Stretch (upper body)—with binary coding (unable/able) for each arm—protocols. All these protocols are routinely applied by PE teachers and are part of teachers' initial education curriculum.

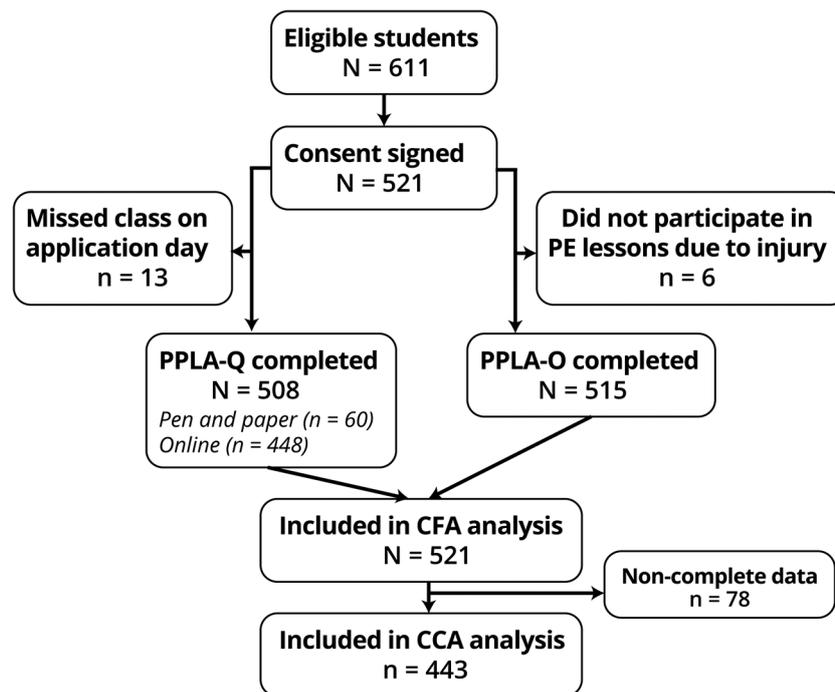


FIGURE 1

Portuguese physical literacy assessment validation participant flowchart. CFA, confirmatory factor analysis; CCA, confirmatory composite analysis.

Psychological domain

The Psychological domain included four indicators assessed through the PPLA-Q: *Motivation*, *Confidence*, *Emotional Regulation*, and *Physical Regulation*. All these indicators comprised the total summed score of responses in each respective scale (composed of seven, nine, seven, and eight items, respectively), and then transformed into a percentage of maximum points (0–100 score) to normalize the different number of items in each scale. All scales have been calibrated through Mokken Scale Analysis [e.g., (29)] using nonparametric Item Response Theory (IRT) models, and have shown evidence supporting good score reliability (Molenaar–Sijtsma's ρ of 0.83 to 0.94) (30) and construct validity in this sample (17): dimensionality (Loevinger's H of .47 to .66) (31), discriminant validity (deattenuated correlations between subscales of 0.27 to 0.73) and convergent validity.

Social domain

The Social domain included four indicators: *Culture*, *Ethics*, *Collaboration*, and *Relationships*. These indicators followed the same logic as those of the Psychological domain presented above, using a total summed score across the seven and six items (for the last three mentioned subscales), respectively. Previous validation using Mokken Scale Analysis (17) resulted in good score reliabilities (ρ of 0.86 to .0.91) and construct validity (H of 0.54 to 0.64; deattenuated correlations of 0.18 to 0.74).

Cognitive domain

The Cognitive domain was assessed through a single indicator: *Content Knowledge*. Its score was derived from calibration of an IRT model (mixed 2-parameters nested logit and graded response model) on 10 response items dealing with knowledge in 5 main content

themes (19). This calibration gathered evidence on construct validity and score reliability of the test (marginal reliability of 0.60) to distinguish students with descriptive (*Foundation*) knowledge from those with higher analytical knowledge (*Mastery*). Factor scores derived with EAP were transformed into a 0–100 score.

Self-reported physical activity

The short form of the *International Physical Activity Questionnaire* (IPAQ-SF) (32) was used to obtain weighted estimates of each intensity of physical activity per week (MET/min/week). No total summed score was used since this instrument has shown different validity across intensities (33, 34) and since it is tenable that different intensities might interact differently with the different domains of PL.

Statistical analysis

All statistical analyses used RStudio 1.4.1106 (35), with R 4.0.1 (36). Missing data and descriptive statistics (Table 1) were computed using the packages *naniar* (37) and *psych* (38). A statistic of χ^2 (593) = 791.65, $p < 0.001$ with 38 missing patterns, in Little's test (39) provided evidence against data Missing Completely at Random. Missing data most likely originated from students missing class on the day of measures' application and therefore is tenable to assume a Missing at Random (MAR) mechanism occurred.

Data were screened for univariate and multivariate normality through the *MVN* package (40); however, the shoulder stretch assessment had to be removed from the latter test to achieve

TABLE 1 Descriptive statistics for measures in the PPLA-questionnaire and PPLA-observation (N = 521).

Variable	n missing (%)	M (SD)	Median	Univariate normality	
				Shapiro-Wilk W	p-Value
Self-reported PA					
Vigorous ^a	22 (4.2)	2,071 (2,084.2)	1,440.0	0.85	<0.001
Moderate ^a	26 (5.0)	1,071 (1,122.2)	720.0	0.80	<0.001
Walking ^a	26 (5.0)	767.5 (950.1)	396.0	0.76	<0.001
PPLA-O measures					
PACER	22 (4.2)	49.5 (22)	44.0	0.93	<0.001
Push-ups	26 (5.0)	18.1 (9.6)	18.0	0.93	<0.001
Curl-ups	23 (4.4)	48.6 (21.7)	45.0	0.91	<0.001
Shoulder stretch (frequency of achievement)					
Right	83 (15.9)	95%		0.21	<0.001
Left	83 (15.9)	89%		0.37	<0.001
Sit and reach (cm)					
Right	85 (16.3)	30.7 (8.3)	31.0	0.99	0.009
Left	84 (16.1)	30.2 (8.2)	31.0	0.99	0.006
Manipulative-based activities ^b	6 (1.2)	54.9 (21.8)	55.4	0.99	0.076
Stability-based activities ^b	6 (1.2)	43.4 (16.4)	43.7	0.98	<0.001
PPLA-Q measures					
Content knowledge ^b	13 (2.5)	68.8 (15.2)	70.0	0.99	<0.001
Motivation ^b		74.9 (14.8)	77.1	0.97	<0.001
Confidence ^b		68.4 (16.8)	68.9	0.98	<0.001
Emotional regulation ^b		69.9 (14.6)	71.4	0.98	<0.001
Physical regulation ^b		75.1 (12.2)	75.0	0.98	<0.001
Culture ^b		58.7 (19.5)	57.1	0.98	<0.001
Ethics ^b		81.5 (11.9)	83.3	0.92	<0.001
Collaboration ^b		85.0 (11.4)	86.7	0.94	<0.001
Relationships ^b		77.7 (13.5)	80.0	0.97	<0.001

PACER, progressive aerobic cardiovascular endurance run; PA, physical activity; PPLA-O, PPLA-observation; PPLA-Q, PPLA-questionnaire.

^aMET/week.

^bMaximum score = 100.

convergence due to it being a binary indicator. Results of the univariate tests are presented in **Table 1**. Mardia's statistics (skewness = 2,739.39, $p < 0.001$; kurtosis = 13.33, $p < 0.001$) render any normality assumption untenable.

Data were screened for multivariate outliers using the Minimum Covariance Determinant approach (41) through the *Routliers* package (42) that highlighted 69 multivariate outliers. Sensitivity analysis revealed no differences in model fit or parameters in the main analysis and so outliers were kept in the analyses. Bivariate Pearson and polychoric correlations between measures were obtained in *polycor* (43) and reported in **Table 3**. For factor-based analysis, the *Push-ups* indicator was multiplied by a factor of 2 to rescale its variance.

Confirmatory factor analysis

Since PPLA is based on the APLF (44), a clear rationale for factorial structure has been previously defended (25). We employed CFA to test the previously hypothesized model structure against other tenable competing models presented in the literature (10, 24) by adding the assumption of a reflective measurement model. Six models were estimated (see **Figure 2**), which included unidimensional, correlated first-order factors, second-order, canonical (symmetric) bifactor, and bifactor S-I-1 models. All models were estimated in *lavaan* 0.6.9 (45)

with the included variables being specified as continuous. Given the violation of multivariate normality, robust maximum likelihood estimation (MLR) with robust "Huber-White" standard errors (46) and a scaled test statistic (equivalent to Yuan-Bentler T2*) (47) were also used. Based on the existence of missing data on various variables and the assumption of MAR, Full Information Maximum likelihood (FIML) (48) was used to estimate unbiased parameters (49).

In all models, the metric of latent factors was fixed by using the first indicator as a marker. Error covariances were constrained to zero unless otherwise specified. In all multiple factors models, the Cognitive domain factor was specified as a single indicator, and its error variance was constrained to $(1 - \text{reliability}) \times \text{variance} \times (\text{indicator})$ (50, 51). Estimation of the last three models (F4-F6) used bounded parameters to stabilize the solution (52). Initial estimation of model F5 resulted in a negative variance (Heywood case); however, changing the marker indicator resolved this issue. For the sixth model, a five-point global indicator of PL ("I can lead a healthy and active life") was inserted into the model. A sensitivity analysis compared MLR estimation with weighted least square mean and variance adjusted (WLSMV) estimation, with no substantial differences in fit indices or parameters [WLSMV Comparative Fit Index (CFI) = 0.96, Root Mean Square Error of Approximation (RMSEA) = 0.05 (0.04–0.06), Standardized Root

TABLE 2 Indices and statistics used, along with thresholds and descriptors.

Index/Statistic	Purpose	Guideline	Descriptor	Reference
Factor-based methods				
Robust χ^2	Absolute Fit	$p > 0.05$		(53, 54)
SRMR		< 0.08		
Robust RMSEA	Approximate Fit	≤ 0.06		
Robust RMSEA 90% CI		0.10 not included in interval		
CFI		≥ 0.95		
Scaled χ^2 difference tests	Model comparison	$p < 0.05$		(58)
AIC		Lower values indicate better fit		(59)
BIC				
Modification indices	Local fit	> 3.84		
Standardized covariance residuals		$< 1.96 $		
Standardized factor loading	Convergent validity	> 0.71	Excellent	(60)
		> 0.63	Very Good	
		> 0.55	Good	
		> 0.45	Fair	
		> 0.32	Poor	
Inter-factor correlations	Discriminant validity	< 0.85		(59)
Omega coefficient (ω and ω_s)	Total score and subscale-score reliability	> 0.80	Good	(61)
		> 0.70	Acceptable	(62)
Omega hierarchical coefficient (ω_H and ω_{HS})	Tenability of interpretation of a sole total score	> 0.80		(63, 64)
ECV _{ss}	Whether the use of subscales adds unique information			
ECV	Essential unidimensionality	PUC > 0.80 , or ECV > 0.60 and $\omega_H > 0.70$		(65)
PUC				
I-ECV	Common variance attributable to general factor in each indicator	$> 0.80-0.85$		(66)
Composite-based methods				
Bootstrapped-based d_L and d_G (1,000 replications)	Global fit of the model (vs. saturated model)	$< 95\%$ quantile of distribution		(21, 67)
Bootstrapped-based SRMR (1,000 replications)			< 0.08	
RMS $_{\theta}$			< 0.12	
Indicator weight magnitude	Local fit			
Indicator weight statistical significance		$p < 0.05$		
VIF	Multicollinearity and Suppressor Effects	< 3.3		(68)
				(69)
Indicator correlations		> 0.90	Very high	(70)
Composite correlations		0.70–0.90	High	
		0.50–0.70	Moderate	
		0.30–0.50	Low	
		0.00–0.30	Negligible	
R^2 adj.	Explained variance	> 0.25	Large	(71)
		$> .09$	Medium	
		$> .01$	Small	
f^2	Predictive validity	> 0.35	Large	(71)
		> 0.15	Medium	
		> 0.02	Small	

CFI, Comparative Fit Index; RMSEA, Root Mean Square Error of Approximation; SRMR, Standardized Root Mean Square Residual; AIC, Akaike's Information Criteria; BIC, Bayesian Information Criteria; ECV, Explained Common Variance; PUC, Percentage of Uncontaminated Correlations; I-ECV, item ECV; RMS, Root Mean Square Error Covariance; d_L , squared Euclidean distance; d_G , geodesic distance; VIF, Variance Inflation Factor; CI, confidence interval.

Mean Square Residual (SRMR) = 0.033], as such, we present the results for the MLR estimation for comparability with other models.

Model fit and selection

All indices used to assess model fit are summarized in Table 2. Guided by the literature (53, 54), thresholds of SRMR, CFI, and RMSEA were used as guidelines for quantifying global fit, rather than as strict rules (50, 55, 56). Only solutions that achieved acceptable or borderline

global fit were summarized. Specific bifactor indices were calculated for the final model (i.e., asymmetrical bifactor) using the *BifactorIndicesCalculator* package (57) in *RStudio*.

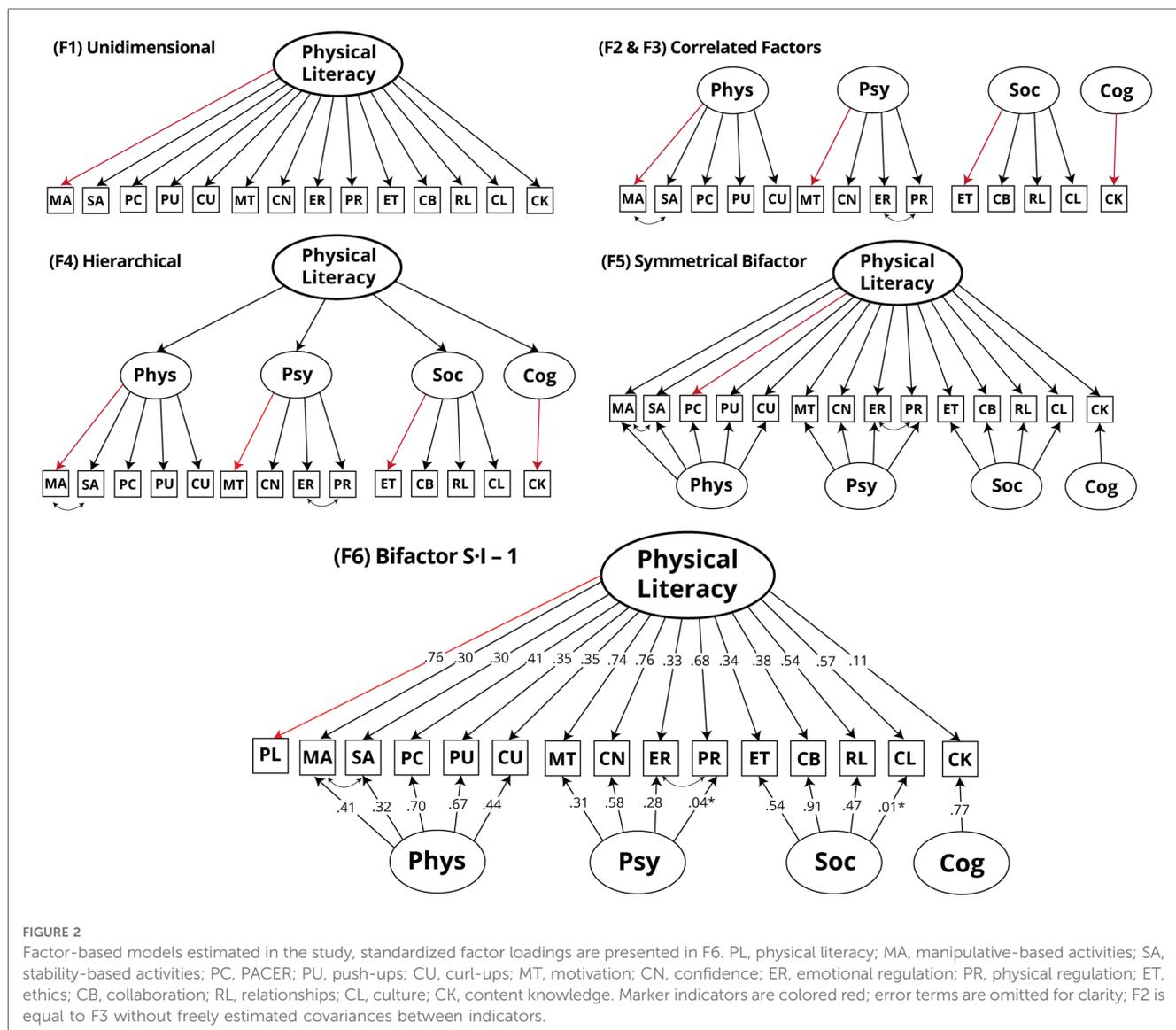
Confirmatory composite analysis

Our initial postulated model conceptualized PL and its domains as composites. Therefore, we used *Confirmatory*

TABLE 3 Pearson and polyserial bivariate correlation matrix.

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	1.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	
1. Self-reported PA-vigorous	—																					
2. Self-reported PA-moderate	0.38	—																				
3. Self-reported PA-walking	0.25	0.36	—																			
4. PACER	0.31	0.01	-0.04	—																		
5. Push-ups	0.29	0.03	-0.07	0.61	—																	
6. Curl-ups	0.24	0.05	-0.03	0.44	0.41	—																
7. Shoulder stretch, right	-0.13	-0.02	-0.08	-0.03	0.08	0.24	—															
8. Shoulder stretch, left	-0.10	-0.05	-0.06	-0.02	-0.04	0.00	0.75	—														
9. Sit-and-reach, right	-0.09	0.05	-0.03	-0.14	-0.07	-0.05	0.31	0.32	—													
10. Sit-and-reach, left	-0.07	0.08	0.01	-0.14	-0.05	-0.01	0.32	0.30	0.93	—												
11. Manipulative-based activities	0.23	0.05	0.01	0.37	0.43	0.34	-0.23	-0.25	-0.23	-0.22	—											
12. Stability-based activities	0.18	0.06	-0.01	0.31	0.35	0.27	-0.06	-0.11	-0.03	-0.01	0.79	—										
13. Content knowledge	-0.01	-0.08	-0.07	0.17	0.10	0.10	0.19	-0.03	0.15	0.14	-0.06	-0.05	—									
14. Motivation	0.45	0.13	0.03	0.40	0.35	0.30	-0.05	-0.04	0.05	0.09	0.33	0.31	0.09	—								
15. Confidence	0.40	0.10	0.04	0.49	0.45	0.37	-0.08	0.02	-0.01	0.03	0.37	0.33	0.02	0.73	—							
16. Emotional regulation	0.06	0.06	0.05	0.19	0.21	0.11	-0.13	-0.04	-0.03	-0.02	0.16	0.14	0.03	0.28	0.42	—						
17. Physical regulation	0.24	0.10	0.14	0.28	0.19	0.25	-0.13	-0.09	0.05	0.06	0.20	0.17	0.006	0.50	0.56	0.44	—					
18. Culture	0.31	0.16	0.06	0.26	0.22	0.23	-0.10	0.07	0.10	0.10	0.21	0.20	0.10	0.47	0.46	0.16	0.36	—				
19. Ethics	0.03	0.03	0.13	-0.08	0.01	0.04	-0.03	0.03	0.11	0.13	0.05	0.04	0.11	0.21	0.16	0.23	0.32	0.16	—			
20. Collaboration	0.08	0.07	0.06	-0.13	0.01	0.02	-0.15	-0.09	0.05	0.07	0.04	0.01	0.04	0.23	0.19	0.17	0.32	0.23	0.63	—		
21. Relationships	0.21	0.10	0.13	0.07	0.07	0.07	-0.17	-0.05	-0.03	-0.01	0.13	0.14	0.03	0.36	0.39	0.20	0.39	0.34	0.42	0.63	—	

PA, physical activity; PACER, progressive aerobic cardiovascular endurance run. All correlations with Shoulder Stretch variables are polyserial correlations.



Composite Analysis (CCA) (72) to mimic the analysis done through CFA and compare both measurement models.

All composite models were estimated in *cSEM* 0.4.0.9000 (73) using the PLS estimator with 1,000 bootstrap replications. All cases with missing data on any of the study variables were deleted (final $N=443$) since no other options are available in *cSEM* at the time of writing. In parallel with the CFA analysis, three models were estimated in mode B (Figure 3). They included a single composite, correlated composite, and a second-order model of PL (using the “two-stage approach”) (74, 75). No bifactor model was estimated, since no literature exists to substantiate it in composite fashion.

To assess the impact of unit weights, models constrained to equal weights for each indicator in the respective composite were estimated. To identify the models, three single indicator factors (one for each intensity of self-reported PA) were inserted into the model as outcomes of the modeled composites. All indices used to assess model fit are listed in Table 2.

Results

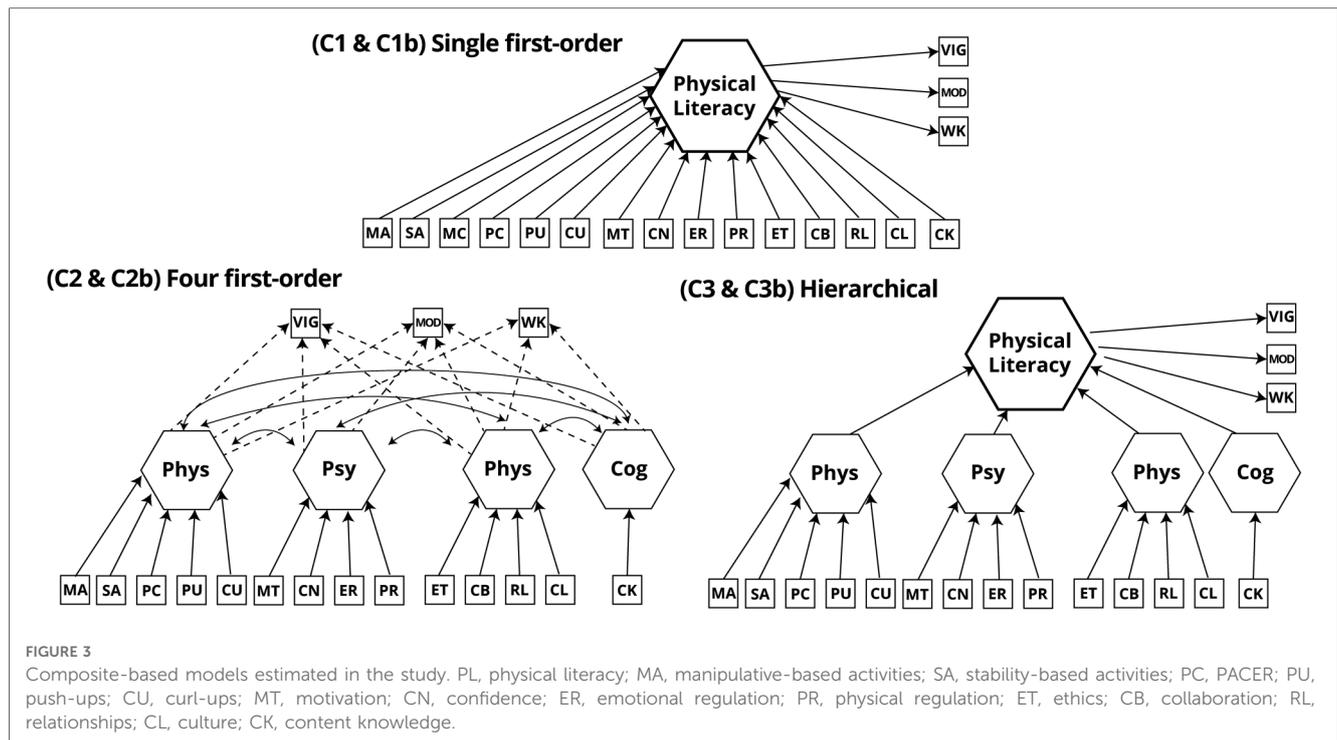
Preliminary analysis

Bivariate correlations between all indicators displayed results compatible with the *a priori* factorial structure, i.e., indicators in the same domain correlated higher with indicators in the same domain than with those of other domains (Table 3). An exception to this was the correlations of the flexibility indicators (*Shoulder Flexibility* and *Sit-and-Reach*), which displayed either no correlation or low negative correlations with other indicators postulated to be in the Physical domain (*PACER*, *Push-ups*, *Curl-ups*, and *Movement Competence* factors). We removed these indicators from the following models.

Confirmatory factor analysis

Model fit

The bifactor S-I-1 model (F6) showed the best absolute fit (SRMR) and relative fit (CFI and RMSEA) to the data



attaining an acceptable fit (see **Table 4**). The symmetrical bifactor model also achieved acceptable values in all indices. Despite not achieving acceptable fit by conservative standards, models F3 and F4 had fit indices close to more lenient standards of 0.90 for CFI, and borderline to RMSEA and SRMR of 0.08 (50).

There was an improvement in model fit for the baseline correlated factors (F2) over the unidimensional F1. Analysis of modification indices (MI) for the F2 model revealed several large values (largest MI = 372.25); however, only two theoretically plausible modifications emerged. The first was to free a residual covariance between both indicators of *Movement Competence* (*Manipulative-based activities* and *Stability-based activities*) as this might be due to a teacher's observation method factor. The second was between the *Emotional Regulation* and *Physical Regulation* indicators, which shared similarities in the wording of the items. We specified a *post-hoc* error covariance between these indicators, resulting in the F3 model. This model was kept for testing in further models. F3 was an improvement over F2 according to all indices.

MI analysis of F3 suggested the model could be further improved by allowing a cross-loading of the *Culture* indicator on the Psychological domain. While this might be theoretically defensible as some items in this scale deal with similar self-related concepts to those of the latter factor, we kept this parameter constrained since the following bifactorial specification would assess whether a general trait could best account for this implied correlation. Fitting the symmetrical bifactor solution (F5) resulted in an improvement over F3; however, MI analysis revealed that the largest MI (55.51) regarded a correlation between two group factors (Physical and Psychological, the two highest correlating first-order factors in the F3 solution).

Therefore, the asymmetrical model was fit (F6), which resulted in better overall indices, lower MI, and most residuals below the 1.96 threshold. No direct comparison was possible due to the insertion of a global indicator of PL to estimate F6.

Convergent and discriminant validity

Model F3 mean factor loadings were very good, with correlations between factors ranging from 0.01 to 0.66 supporting discriminant validity. The correlation pattern was uneven, though, with the Psychological factor showing moderate correlations with only the Physical and Social factors.

Mean factor loadings in the group factors of the F5 and F6 solutions were lower than in F3. This was expected since, in these models, the group factors represent residual variance not explained by the general factor. Between these models, there was a marginal increase ($\Delta = 0.02$) in mean factor loadings in the general factor, and a concomitant reduction in two of the group factors (i.e., Psychological and Social).

In the asymmetrical bifactor model (F6), three indicators had excellent loadings on the general factor (Physical Literacy), one had very good, one had good, and the remaining seven indicators had borderline (~ 0.32) to poor loadings. *Content Knowledge* had no statistically significant loading (see **Table 5**). Loadings on the group factors (residual variance not explained by the general factor) ranged from 0.32 to 0.70, 0.04 to 0.58, 0.01 to 0.91, and 0.77, on the Physical, Psychological, Social, and Cognitive domains, respectively. Except for the indicators in the Psychological domain, all indicators had average higher loadings on their group factors than on the general factor (see **Table 4**). After accounting for the general factor (PL), half of the inter-factor correlations became negative, with the remaining three showing a decrease in their positive correlations.

TABLE 4 Model fit, mean factor loadings, and inter-factor correlations for factor-based models.

Fit measure	First-order models			Second-order models		
	Unidimensional (F1)	Correlated factors (F2)	Correlated factors (F3) ^a	Hierarchical (F4)	Bifactor (F5)	Bifactor S-I-1 (F6)
MLR χ^2	1,217.07 (77), $p < .001$	675.20 (72), $p < .001$	329.15 (70), $p < .001$	378.38 (72), $p < .001$	182.02 (62), $p < .001$	160.62 (69), $p < .001$
Robust CFI	0.55	0.78	0.90	0.89	0.95	0.97
Robust RMSEA (90% CI)	0.18 (0.17–0.19)	0.13 (0.12–0.14)	0.09 (0.08–0.10)	0.09 (0.08–0.10)	0.07 (0.06–0.08)	0.05 (0.04–0.06)
SRMR	0.12	0.11	0.09	0.09	0.06	0.04
AIC	58,075.108	57,452.753	57,110.133	57,150.092	56,991.07	
BIC	58,253.850	57,652.773	57,318.665	57,350.112	57,233.65	
χ^2 robust different test		F1 vs. F2: $\chi^2 = 319.51$, $\Delta df = 5$, $p = < .001$	F2 vs. F3: $\chi^2 = 7,338.7$, $\Delta df = 2$, $p = < .001$	F4 vs. F3: $\chi^2 = 151.35$, $\Delta df = 2$, $p = < .001$	F3 vs. F5: $\chi^2 = 712.34$, $\Delta df = 8$, $p = < .001$	
Mean factor loadings						
General					0.44 (0.20)	0.46 (0.20)
Physical			0.61 (0.15)		0.44 (0.16)	0.51 (0.17)
Psychological			0.69 (0.21)		0.32 (0.14)	0.30 (0.22)
Social			0.65 (0.22)		0.55 (0.36)	0.48 (0.37)
Cognitive			0.77		0.77	0.77
Factor correlations (SE)						
Physical–Psychological			0.67 (0.03)			0.48 (0.11)
Physical–Social			0.02 (0.06)			–0.33 (0.07)
Physical–Cognitive			0.19 (0.06)			0.14 (0.08)
Psychological–Social			0.42 (0.07)			–0.20 (0.12)
Psychological–Cognitive			0.06 (0.06)			–0.14 (0.09)
Social–Cognitive			0.09 (0.07)			0.00 (.06)

CFI, comparative fit index; RMSEA, root mean square error of approximation; MLR, maximum likelihood robust; SRMR, standardized root mean square residual; AIC, Akaike’s information criteria; BIC, Bayesian information criteria; SE, standard error.

^aCorrelated residuals between emotional regulation and physical regulation, and manipulative-based activities and stability-based activities.

Variance and reliability

Regarding model-based reliabilities in the final model (F6), the total PL score (i.e., summing all indicators, after normalization) attained good reliability, being estimated that 89% of its variance was due to both the general and group factors ($\omega = 0.89$). An estimated 71% of total score variance was due to individual differences in the general factor. Reliabilities for the subscale scores were all acceptable or good, except for the Cognitive score (ω_s from 0.60 to 0.82). Based on the relationship between ω_s and ω_{Hs} , three subscales attained the recommended thresholds for adding statistical value over and beyond that of the total score (recommended $\omega_{Hs} = 0.212$, 0.192, 0.192, and 0.244, respectively).

Similar results were estimated by the Explained Common Variance (ECV): 49% of the total common variance (inherent to both general and group factors) is explained by the general factor (see Table 5). Of all indicators, only three (*Motivation*, *Physical Regulation*, and *Culture*), achieved the tentative 0.80–0.85 threshold for item ECV (I-ECV) (66) and can be regarded as essentially being influenced by the general trait alone. Except in the Psychological domain, the most reliable variance in indicators was explained by their respective group factor, resulting in marked dimensional uniqueness ($ECV_{ss} = 0.56, 0.22, 0.71, \text{ and } 0.99$; Physical, Psychological, Social, and Cognitive group factors, respectively). Based on ECV_{ss} and ω_s , all group

factors attain the recommended value for dimensional uniqueness (i.e., warrant interpretation) [recommended $\omega_s = 0.479, 0.815, 0.479, \text{ and } 0.479$; (63)]. Finally, 79% of all correlations were saturated by the general factor (Percentage of Uncontaminated Correlations, $PUC = 0.79$), bordering on the 80% recommendation (65) for consideration of essential unidimensionality in future SEM measurement models.

Confirmatory composite analysis

Model fit

The single composite models (C1 and C1b) showed the best absolute fit to the data, with all estimated values below or bordering their critical value suggesting excellent fit to the data (see Table 6). Both C2 and C3 models provide an acceptable fit to the data, with estimated borderline SRMR (both cases) and RMS_θ below their thresholds, despite having estimated global fit indices bordering above the critical value. In terms of unit-weighted models, summing every indicator with equal weights to produce a total score (C1b) reproduced the observed relationships in the model better than the alternative sum (also with equal weights per indicator) into domain scores (assuming each domain as an emergent variable; C2b).

TABLE 5 Item parameters and model-based indices for the bifactor S*1-1 model (F6).

Subtest	General		Physical		Psychological		Social		Cognitive		I-ECV
	Standardized loading (SE)	S ²									
PL global	0.76 (0.03)	0.58									1.00
Manipulative-based activities	0.30 (0.05)	0.09	0.41 (0.06)	0.17							0.35
Stability-based activities	0.30 (0.05)	0.09	0.32 (0.06)	0.10							0.46
PACER	0.41 (0.05)	0.16	0.70 (0.04)	0.49							0.25
Push-ups	0.35 (0.05)	0.12	0.67 (0.05)	0.45							0.21
Curl-ups	0.35 (0.05)	0.12	0.44 (0.05)	0.19							0.38
Motivation	0.74 (0.04)	0.55			0.31 (0.11)	0.10					0.85
Confidence	0.76 (0.04)	0.58			0.58 (0.08)	0.34					0.63
Emotional regulation	0.33 (0.06)	0.11			0.28 (0.08)	0.08					0.58
Physical regulation	0.68 (0.04)	0.47			0.04 (0.09)	0.00					0.996
Ethics	0.34 (0.06)	0.12					0.54 (0.04)	0.30			0.28
Collaboration	0.38 (0.06)	0.14					0.91 (0.06)	0.83			0.15
Relationships	0.54 (0.04)	0.29					0.47 (0.04)	0.22			0.57
Culture	0.57 (0.04)	0.33					0.01 (0.05)	0.00			0.999
Content knowledge	0.11 (0.05)	0.01							0.77 (0.02)	0.59	0.02
ECV/ECV _{SS}		0.49		0.71		0.23		0.61		0.98	
ω_5		0.89		0.76		0.82		0.80		0.60	
ω_H/ω_{HS}		0.71		0.52		0.15		0.42		0.59	
PUC		0.79									

PL, physical literacy; S², squared standardized loading; SE, standard error; PACER, progressive aerobic cardiovascular endurance run; ECV, Explained Common Variance; PUC, Percentage of Uncontaminated Correlations; I-ECV, item ECV; ω_5 , Omega coefficient subscale; ω_H/ω_{HS} , Omega hierarchical coefficient/Omega hierarchical subscale.

TABLE 6 Model fit and inter-factor correlations for composite-based models (C1–C3b); n = 443.

Fit measure	First-order				Second-order	
	Single first-order composite (C1)	Single first-order composite-unit-weighted (C1b)	Four first-order composites (C2)	Four first-order composites-unit-weighted (C2b)	Hierarchical (C3)	Hierarchical-unit-weighted (C3b) ^a
RMS _θ	<0.001	<0.001	0.04	0.03		
SRMR (critical value 95%)	0.050 (0.054)	0.059 (0.059)	0.078 (0.068)	0.087 (0.061)	0.071 (0.060)	0.090 (0.062)
d _L (critical value 95%)	0.391 (0.443)	0.534 (0.533)	0.928 (0.707)	1.147 (0.563)	0.767 (0.547)	1.228 (0.586)
d _G (critical value 95%)	0.085 (0.106)	0.118 (0.113)	0.179 (0.146)	0.215 (0.138)	0.169 (0.134)	0.224 (0.140)
Construct correlations						
Physical–Psychological			0.49 (0.04)	0.47 (0.03)		
Physical–Social			0.27 (0.06)	0.12 (0.04)		
Physical–Cognitive			0.11 (0.05)	0.06 (0.04)		
Psychological–Social			0.53 (0.05)	0.49 (0.04)		
Psychological–Cognitive			0.07 (0.05)	0.05 (0.05)		
Social–Cognitive			0.05 (0.05)	0.09 (0.05)		

RMS, root mean square error covariance; SRMR, standardized root mean square residual; d_L, squared Euclidean distance; d_G, geodesic distance. Statistically significant weights (p < 0.05) are bolded.

^aOnly the first-order composite is produced by summing the indicators with equal weights; total score for hierarchical score is obtained by optimally weighting its first-order scores.

TABLE 7 Item parameters and total effects of single first-order composite models (C1 and C1b); n = 443.

Composite ← indicators	β (SE)	VIF	Indicator correlation
PL←			−0.14 to 0.77
Manipulative-based activities	0.20 (0.14)	2.88	
Stability-based activities	−0.25 (0.13)	2.55	
PACER	0.15 (0.16)	2.03	
Push-ups	0.12 (0.15)	1.89	
Curl-ups	0.09 (0.10)	1.39	
Motivation	0.61 (0.13)	2.39	
Confidence	0.13 (0.10)	3.14	
Emotional regulation	−0.17 (0.10)	1.41	
Physical regulation	0.03 (0.13)	1.84	
Ethics	−0.07 (0.12)	1.76	
Collaboration	−0.05 (0.14)	2.36	
Relationships	0.15 (0.12)	1.89	
Culture	0.18 (0.11)	1.38	
Content knowledge	−0.14 (0.10)	1.10	
Optimal weights (C1)	Total effects	f²	R² adj.
	β (SE)		
IPAQ ← PL			
Vigorous	0.52 (0.03)	0.36	0.27
Moderate	0.16 (0.06)	0.03	0.02
Walking	0.07 (0.07)	0.00	0.00
Unit weights (C1b)			
IPAQ ← PL			
Vigorous	0.39 (0.04)	0.18	0.15
Moderate	0.12 (0.05)	0.02	0.01
Walking	0.07 (0.05)	0.01	0.00

β, standardized weights; SE, standard error; VIF, variance inflation factor; R² adj., adjusted R²; PACER, progressive aerobic cardiovascular endurance run; PL, physical literacy; IPAQ, international physical activity questionnaire. Statistically significant (p < 0.05) weights are bolded.

Variance

Standardized weights in the single composite model (C1; Table 7) ranged from −0.25 to 0.61, with *Motivation* being the only indicator with a statistically significant result. All other indicators do not contribute beyond this indicator. High correlations (<0.70) existed between some indicators, with corresponding Variance Inflation Factor (VIF) ranging from 2.26 to 3.14, suggesting the existence of suppressor effects.

In the correlated composites model (C2; Table 8), standardized weights increased for most variables. Six indicators had non-statistically significant weights and two were borderline (p ≈ 0.05). VIF values at indicator level were lower, with all composites showing high correlations among some of its indicators, with unexpected, inverted signs in three. Correlations among composites ranged from negligible (0.05) to moderate (0.53) (70) (see Table 6), with correlations with the Cognitive composite being all negligible and non-statistically significant (0.05–0.11).

Weights, VIF, and indicator correlation mostly kept their magnitude in the second-order composite (see Table 9). The exception was an increase in weights for *Push-ups*, *Curl-ups*, and *Physical Regulation* indicators, and a decrease in *Emotional Regulation* and *Manipulative-based Activities*. First-order weights on the second-order composite attributed a higher relative contribution to the Psychological composite (β = 0.69), in explaining variance in the second-order composite of PL. Analysis of the first-order loadings (bivariate correlations) suggested that the Physical and Social composites (loadings = 0.68, not shown) were still important in composing this second-order composite, despite explaining substantially fewer amounts of variance.

TABLE 8 Item parameters and total effects of four first-order composite models (C2 and C2b); *n* = 443.

Composite ← indicators	β (SE)	VIF	Indicator correlation
Physical←		1.33	0.29–0.77
Manipulative-based activities	0.41 (0.20), <i>p</i> = 0.050	2.82	
Stability-based activities	−0.23 (0.19)	2.50	
PACER	0.53 (0.17)	1.72	
Push-ups	0.28 (0.19)	1.79	
Curl-ups	0.22 (0.15)	1.34	
Psychological←		1.70	0.29–0.73
Motivation	0.70 (0.13)	2.22	
Confidence	0.40 (0.15)	2.57	
Emotional regulation	−0.22 (0.11), <i>p</i> = 0.065	1.35	
Physical regulation	0.06 (0.15)	1.64	
Social←		1.40	0.13–0.63
Ethics	−0.02 (0.22)	1.64	
Collaboration	−0.24 (0.24)	2.22	
Relationships	0.59 (0.17)	1.69	
Culture	0.76 (0.14)	1.12	
Cognitive←		1.01	—
Content knowledge	1.0		
Optimal weights (C2)	Total effects β (SE)	<i>f</i>²	<i>R</i>² adj^a
IPAQ ← Physical			
Vigorous	0.19 (0.05)	0.04	0.25
Moderate	−0.04 (0.06)	0.00	0.04
Walking	−0.08 (0.07)	0.01	0.02
IPAQ ← Psychological			
Vigorous	0.36 (0.05)	0.10	
Moderate	0.08 (0.06)	0.00	
Walking	0.04 (0.07)	0.00	
IPAQ ← Social			
Vigorous	0.06 (0.05)	0.00	
Moderate	0.15 (0.06)	0.02	
Walking	0.13 (0.08)	0.01	
IPAQ ← Cognitive			
Vigorous	−0.04 (0.04)	0.00	
Moderate	−0.08 (0.05)	0.01	
Walking	−0.08 (0.04)	0.01	
Unit weights (C2b)			
IPAQ ← Physical			0.17
Vigorous	0.22 (0.05)	0.04	0.02
Moderate	−0.00 (0.06)	0.00	0.03
Walking	−0.09 (0.06)	0.00	
IPAQ ← Psychological			
Vigorous	0.25 (0.05)	0.05	
Moderate	0.11 (0.06)	0.01	
Walking	0.08 (0.07)	0.00	
IPAQ ← Social			
Vigorous	0.03 (0.06)	0.00	
Moderate	0.08 (0.07)	0.01	
Walking	0.12 (0.06)	0.01	
IPAQ ← Cognitive			
Vigorous	−0.02 (0.05)	0.00	
Moderate	−0.08 (0.04)	0.01	
Walking	−0.09 (0.05)	0.01	

β , standardized weights; SE, standard error; VIF, variance inflation factor; *R*² adj., adjusted *R*²; PACER, progressive aerobic cardiovascular endurance run; PL, physical literacy; IPAQ, international physical activity questionnaire. Statistically significant (*p* < 0.05) and borderline weights are bolded. ^aResults displayed refer to the combined effect on vigorous, moderate, and walking intensities, respectively.

TABLE 9 Item parameters and total effects of the hierarchical composite models (C3 and C3b); *n* = 443.

Composite ← indicators	β (SE)	VIF	Indicator correlation
First-order			
Physical ←			0.29–0.77
Manipulative-based activities	0.19 (0.14)	2.82	
Stability-based activities	0.09 (0.14)	2.50	
PACER	0.54 (0.11)	1.72	
Push-ups	0.21 (0.12), <i>p</i> = 0.072	1.79	
Curl-ups	0.29 (0.10)	1.34	
Psychological←			
Motivation	0.44 (0.08)	2.22	0.29–0.73
Confidence	0.56 (0.10)	2.57	
Emotional regulation	−0.10 (0.07)	1.35	
Physical regulation	0.16 (0.08), <i>p</i> = 0.06	1.64	
Social←			0.13–0.63
Ethics	0.08 (0.12)	1.64	
Collaboration	−0.14 (0.14)	2.22	
Relationships	0.55 (0.09)	1.69	
Culture	0.73 (0.07)	1.12	
Cognitive←			—
Content Knowledge	1.0	—	
Second-order			
PL←			
Physical	0.27 (0.13)	1.42	
Psychological	0.69 (0.13)	1.93	
Social	0.22 (0.15)	1.48	
Cognitive	−0.14 (0.10)	1.02	
Optimal weights (C3)	Total effects β (SE)	<i>f</i>²	<i>R</i>² adj.
IPAQ–PL			
Vigorous	0.48 (0.04)	0.30	0.23
Moderate	0.16 (0.05)	0.03	0.02
Walking	0.09 (0.06)	0.01	0.01
Unit weights (C3b)			
IPAQ–PL			
Vigorous	0.41 (0.04)	0.20	0.17
Moderate	0.14 (0.05)	0.02	0.02
Walking	0.09 (0.06)	0.01	0.01

β , standardized weights; SE, standard error; VIF, variance inflation factor; *R*² adj., adjusted *R*²; PACER, progressive aerobic cardiovascular endurance run; PL, physical literacy; IPAQ, international physical activity questionnaire. Statistically significant (*p* < 0.05) weights and borderline weights are bolded.

All optimally weighted approaches (C1, C2, and C3) explained similar amounts of variance in self-reported vigorous PA, with the single composite model having marginally higher values (*R*² adj. = 0.27; **Table 7**). The variance explained on moderate and walking self-reported PA was negligible. Using the correlated composite (C2) approach revealed different contributions by composite; Psychological and Physical domains had a higher effect size (*f*²) on vigorous PA, while the Social domain had a low effect size on moderate PA (71). Unit-weighting produced reductions in all effect sizes compared to optimally weighted composites, with the greatest reduction in the single composite (C1b). Again, the correlated composites model (C2b) revealed a sharper decrease in the contributions of the Psychological composite, with others maintaining their relative magnitudes.

Discussion

This paper aimed to assess the construct validity of the PPLA and the adequacy of using a PL total score and respective subscores. It also investigated the practical and ontological implications of the different methods for PL. In summary, results from factor-based methods suggested that an asymmetrical bifactor model with correlated group factors (F6) provided the best fit to the data. This suggests the existence of a common trait underlying reliable individual variation of response, simultaneous with significant uniqueness in each domain. In composite-based methods, results suggested that the single first-order composite models (C1 and C1b) provided the best fit, while the hierarchical model (C3) provided the most interpretable solution for comparison purposes. Overall, evidence supported the construct validity and reliability of the PPLA for its intended use as an integrated tool to measure PL as a multidimensional construct in 15- to 18-year-old Portuguese students in a physical education setting, with comparable results across both factor-based and composite-based methods.

Factor-based methods

Our results from factor-based methods suggest that the best-fitting representation of a measurement model for the PPLA is an asymmetrical bifactor model (F6) with correlated group factors. These findings differed from those found in other PL measuring batteries. In the most recent construct validation effort of the *Canadian PL Assessment* (CAPL), a second-order factor was modeled to account for correlations between domains of PL (10). However, the authors did not report fitting a bifactorial model. Similarly, in a validation of a PL measuring model for children and youth (24), a second-order factor model was chosen as the best representation of the data (with a bifactorial model providing an inadmissible solution). In our study, estimation of a second-order factor model provided a worse fitting (compared to both a correlated factors model and bifactor models) and an inadmissible solution to the data.

We gather that this might stem from an artifact produced by an uneven pattern of correlations among factors, which does not suggest a direct underlying common cause (i.e., factor). While the Physical factor correlated highly with the Psychological factor and moderately with the Cognitive factor, it did not correlate sufficiently with the Social factor. Despite using different measures and operational definitions of the constructs, the CAPL's correlations among factors followed a similar pattern, which then resulted in one of the posited first-order factors (*Knowledge and Understanding*) having a poor loading (0.21) on the second-order Physical Literacy factor (10). Similar results emerged in our study, providing evidence against a second-order model interpretation, with the first-order factor mediating the effect of PL on each indicator. A bifactorial model represents direct effects of the general factor (PL) on indicators, with the asymmetrical version allowing for correlations among group

factors, resulting in a better fit than that of its symmetrical counterpart. This suggests that the orthogonality constraint was overly restrictive and that the PL general factor fails to account for all shared variance among domains.

Our results from the bifactor model (F6) analysis suggest the existence of a common trait underlying reliable individual variation of responses (i.e., Physical Literacy), albeit not with the strength required for a meaningful statistical interpretation of a total-PPLA score in isolation. Instead, the complementary use of unit-weighted subscale scores has added value over and beyond the single total score since they present enough dimensional uniqueness. A noteworthy exception is that of the interpretation of the Psychological subscale. This occurs since indicators in this domain seem to be saturated by the general factor. Any interpretation of differences on this subscale would be biased by shared variance across domains. A tentative interpretation of this fact can be given by the prominent role of psychological variables in predicting PA in both adolescents (76, 77) and adults (78); similarly, these variables might play a mediating role between other domains.

Similarly, despite achieving borderline values to be considered essentially measuring the single trait of PL, high values of ECV_{SS} and moderate values of ECV suggest that further research in SEM contexts should use a bifactor measurement model for the PPLA. This would also allow testing of different effects of the general factor and group factors.

Comparison with composite-based methods

The single composite models (both optimally and unit-weighted) attained the best fit, with optimally weighted first-order and hierarchical factor (C2 and C3) providing borderline adequate approximate fit to the data. Of these, the latter provided the most interpretable solution in terms of individual contribution of indicators, since it reduces the possibility of multicollinearity. Despite attaining excellent fit by all metrics, the single optimal-weighted model (C1) had non-statistically significant weights for all but the *Motivation* indicator. This could be explained by the existence of high correlations among indicators and the number of indicators estimated in the same composite.

Although no assumptions regarding covariation of the indicators are made in a composite model, high correlation patterns among indicators will generally result in multicollinearity and cause suppression effects, co-occurrence of positive and negative weights (i.e., “flipped signs”), and preclude a meaningful interpretation of these weights in general (69). Since a multiple regression is used to estimate the weights of indicators, these are competing for explained variance, increasing the chance of non-statistically significant weights to be estimated. This phenomenon was minimized in the correlated composites, and mainly in the hierarchical model (C3), where most indicators had statistically significant (or borderline) weights, with the expected direction.

Comparing across methodologies, both the hierarchical composite model and asymmetrical bifactor model attained

similar results. In the former, both *Manipulative-based* and *Stability-based Activities*, along with *Emotional Regulation*, *Ethics*, and *Collaboration* indicators, did not contribute to explaining variance in their respective composites over and beyond other indicators. While the magnitudes of standardized loadings in the latter obtained by these indicators in the general factor of PL, along with I-ECV, were poor. A similar case occurred with the *Content Knowledge* indicator. Its poor performance was carried into the first-order weight. Analyzing the first-order weights and the inverse of the ECV_{SS} suggests a similar pattern in that the Psychological indicators contribute more to the general factor/higher-order composite ($\beta = 0.69$, $ECV_{GS} = 0.78$), whereas the Physical ($\beta = 0.27$, $ECV_{GS} = 0.29$) and Social ($\beta = 0.22$, $ECV_{GS} = 0.39$) indicators contribute similarly, and the Cognitive indicator contributes marginally ($\beta = -0.14$, $ECV_{GS} = 0.02$). This is parallel to our earlier discussion on the absorption of the most of the Psychological indicators' variance into the general PL factor (asymmetrical bifactor model).

Thus, these models could be further improved by dropping indicators with statistically nonsignificant weights (69), indicators with low I-ECV, or with poor loadings on both the general and group factors (64). This, however, might compromise content validity of the PPLA and meaningful interpretation of these indicators within their group factors. We recommend that before any removal of indicators is undertaken, this analysis should be replicated in a large independent sample and outside of COVID-19 restrictions, which might change the effects on how different elements of PL correlate with each other and concomitantly on the measurement models. Future development of instruments to measure the remaining elements of the Cognitive domain (i.e., *Tactics* and *Rules*) might draw a different global picture for the construct.

Further parallels can be drawn between results in the different methodologies. In both correlated factors models, correlations among the different domains maintained a similar relative pattern. The Psychological domain was moderately correlated with the Physical and Social domains, with the remaining correlations being lower. A noteworthy difference is an increase in correlation among the Physical and Social domains in the composite model, which could be attributed to a difference in indicator weighting between models.

In conclusion, evidence in favor of a measurement model with a higher-order PL construct (either represented by an asymmetrical bifactor or composite hierarchical model) was mostly robust across methods, with comparable results. Regarding the use of a total summed score, the methods present slightly different results. In the composite-based methods, total score was an adequate representation of an emergent PL variable, while in the factor-based methods, this total score does not quite reach the uniqueness (ω_H) needed to represent a singular latent variable. Based on this, we advise calculation of a total summed PL score, along with domain scores, which should be interpreted conjointly in applied settings.

Conceptual implications for physical literacy

From a reflective perspective, a bifactor view seems the most empirically and conceptually plausible one, since it is tenable that

transversal broadband meta-learning or disposition influences all different elements in a movement context, while domain-specific processes inherent to the different physical, affective, social, and cognitive skills originate clusters of highly interdependent variance. This seems compatible with the APLF's conceptualization of a higher learning state where learning in one element is transferable between elements and domains [*Transfer and Empowerment*, akin to the *Relational Abstract* level of the Structure of Observed Learning Outcomes (79)].

Also, if a higher-order common factor perspective is tenable, then different domains and elements are theoretically interchangeable since they are merely a sample of the infinite indicators and facets that could be chosen to measure PL. While it seems plausible that a different set of indicators could be selected according to the research questions and applications at hand, it might diminish the integrated perspective that researchers have been seeking all along if proper care is not taken to ensure representation across all domains. The asymmetrical bifactor model, however, offers a compromise solution as it becomes possible to acknowledge that while indicators are interchangeable within a domain, domains themselves are not interchangeable (80, 81) and are essential to defining the PL construct. Other plausible interpretations include that of PL as a network of interconnected latent variables that may or may not correlate [i.e., similar to our F3 model, and initial efforts of the CAPL (23)]. This could, however, compromise the place of PL in educational policy discussion, since it would present no added value as a whole variable and could easily be dissected based on convenience.

From a diametrical perspective, viewing PL as an emergent variable through a pragmatic lens (i.e., as a composite, assumed without measurement error or disturbance terms) could also be plausible. As such, rather than being an existing phenomenon to be measured, PL would instead be an umbrella term to designate and index a nomological network of variables that form a sum higher than its individual parts to predict movement-related outcomes throughout the life course, without a singular common cause (82, 83). This would also recognize that selected indicator variables might share a distal common cause mediated through a complex chain of mediators and moderators. This is a vision more compatible with the epistemic phenomenology position of the *Whiteheadian* school of thought, wherein everyone might have a different pattern of correlations (including no correlation) among domains and elements depending on their personal understanding and development of PL. A risk, however, to this interpretation is the possibility of interpretational confounding with data-derived weights (optimal weights), which could compromise the theoretical standing of the concept and similarly hinder meaningful progress in educational practices if care is not taken in interpretation and dissemination. A solution for this might be the use of unit-weighted composites (as shown), or *a priori-defined* weights based on theory or intended usage.

Alternatively, a causal-formative framework could also be used, given PL's composite-based nature (22, 67). The scope of PL, in this case, would be directly defined by its composing domains and would require that all domains of PL be included when estimating the model, which would reinforce its holistic nature. This would

view PL as an aggregated latent variable composed of multiple non-exchangeable domains whose variation could be explained by variation in only a specific set of elements that did not mandate concomitant variation in all elements (22, 83). Further research should seek to reconcile and/or discuss these paradigms.

Since our results are compatible with both the common factor and composite interpretations, we take a practical stance and tentatively recommend the common factor lens of analysis. This is the implicit foundation of both Classical Test Theory (CTT) and IRT which afford access to a robust analysis toolkit (e.g., FIML estimation) to explore dimensionality, score adequacy, and response patterns (17–19). It may also afford the possibility of disentangling the impact of different group factors on intended outcomes of PL by controlling for the general PL factor. Nonetheless, further comparison of practical impacts on derived scores, with different datasets and under different conditions, might be warranted to determine the adequacy of the conceptual interpretation described.

Strengths and limitations

A major strength of this study was the comparison between two different methodologies to draw inferences about the construct validity and reliability of the PPLA. Second, to our knowledge, this is the first study to demonstrate the application of bifactor models to a PL assessment tool that assessed the adequacy of interpretation or use of scores and subscores. Third, our study builds upon measures that have gathered evidence of construct validity and reliability at item-level using Item Response Theory methodologies that provide accurate estimates, and the capability to study item quality and psychometric behavior.

Some limitations of our study include *post-hoc* modifications to initially hypothesized models (i.e., correlated residuals) and the need to use bounded estimation for the factor-based higher-order models. We did not account for multilevel grouping within data (i.e., schools and classes), which could also hold some bias over the results.

Despite mimicking the relative composition of grade 10–12 student population in Portugal according to both grade and course major, our sample was a convenience one. All these points warrant caution before generalizing any of our findings outside of this sample, without further cross-validation with a larger independent sample and multilevel estimation. This is also a requirement if scores derived from PPLA are used as antecedent or precedent variable(s) in extended studies.

Similarly, to assess whether studied relationships among constructs hold across different population groups, measurement invariance should be assessed for the full model, along with its predictive validity on meaningful outcomes (e.g., objectively measured PA, wellbeing), which was not a focus in this study. We also recognize that PL could cogently be modeled using equivalent or alternative models and encourage further research.

Another limitation created using IRT-calibrated measures at indicator level was the incapability to account for measurement error at the lower abstraction level (which is one of SEM's

strengths). This could have attenuated correlations among first-order factors and biased our overall interpretations. Future methods to account for this should be used.

A particular conceptual limitation was the elimination of flexibility indicators in this version of the PPLA. We argue these indicators are relevant to the whole-picture PL and their inclusion should be considered and scrutinized in future efforts.

Conclusion

Using both confirmatory factor analysis and confirmatory composite analysis, we gathered evidence supporting the construct validity and reliability of the PPLA as an integrated tool to measure PL as a multidimensional construct in 15- to 18-year-old Portuguese students. Out of all the estimated models, the bifactor model enabled richer conclusions on the tenability and interpretation of total and subscales (per PL domain) scores. Nonetheless, a composite model description also seems preliminary tenable and useful for predicting self-reported PA. Present results provide evidence that a general trait of PL accounts for a considerable amount of variance in all indicators, albeit with insufficient strength to be interpreted in isolation, along with clear domain-specific variance.

Based on this, we suggest calculation of a total summed PL score, along with domain scores, which should be interpreted conjointly in applied settings. While the former provides a heuristic summary for a quick comparison of different classes and schools in low-stakes settings, the latter allows for a more meaningful interpretation of students' PL profiles and needs. We also recommend the use of total scores per element/indicator in contexts that would benefit from the detail they provide. Most research settings would benefit from using a bifactor measurement model, which enables disentanglement of the variance attributed to the general PL trait and its domains.

We believe all these options will offer flexible solutions for both practitioners—generating specific feedback for students, families, teachers, and schools—and researchers—supporting, e.g., efforts to monitor quality PE practices and the longitudinal impact of educational policies and/or specific interventions on PL. This will contribute to a better understanding of the development of PL and, ultimately, more meaningful PL journeys.

Data availability statement

The datasets presented in this article are not readily available because participants of this study did not explicitly agree for their data to be shared publicly. Requests to access the datasets should be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Council of Faculty of Human Kinetics. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

JMo led the writing of the draft manuscript and prepared figures and tables as part of his Ph.D. thesis. JMa and MO actively supported the definition of the project and participated in instrument development and revision along all phases (as Ph.D. supervisors of JMo). All authors contributed significantly to the final manuscript in terms of content, layout, and analysis. All authors contributed to the article and approved the submitted version.

Funding

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

References

- Physical Literacy for Life. What is Physical Literacy (2021). Available at: <https://physical-literacy.isca.org/update/36/what-is-physical-literacy-infographic> (Accessed December 15, 2021)
- Organisation for Economic Co-operation and Development. The Future of Education and Skills: Education 2030. Organisation for Economic Co-operation and Development (2018). Available at: [https://www.oecd.org/education/2030/E2030%20Position%20Paper%20\(05.04.2018\).pdf](https://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf) (Accessed December 7, 2021).
- UNESCO. *Quality Physical Education (QPE): guidelines for policy makers*. Paris: UNESCO Publishing (2015).
- World Health Organization. *WHO guidelines on physical activity and sedentary behaviour*. Geneva: World Health Organization (2020). Available at: <https://apps.who.int/iris/handle/10665/336656> (Accessed August 30, 2021).
- Ydo Y. Physical literacy on the global agenda. *Prospects*. (2021) 50(1):1–3. doi: 10.1007/s11125-020-09524-8
- Barnett LM, Mazzoli E, Hawkins M, Lander N, Lubans DR, Caldwell S, et al. Development of a self-report scale to assess children's perceived physical literacy. *Phys Educ Sport Pedagogy*. (2022) 27(1):91–116.
- Cairney J, Veldhuizen S, Graham JD, Rodriguez C, Bedard C, Bremer E, et al. A construct validation study of PLAYfun. *Med Sci Sports Exerc*. (2018) 50(4):855–62. doi: 10.1249/MSS.0000000000001494
- Chen ST, Tang Y, Chen PJ, Liu Y. The development of Chinese assessment and evaluation of physical literacy (CAEPL): a study using Delphi method. *Int J Environ Res Public Health*. (2020) 17(8):2720. doi: 10.3390/ijerph17082720
- Gandrieau J, Schnitzler C, Derigny T, Potdevin F. *Évaluation de la littératie physique: création d'un outil de mesure pour les jeunes adultes*. In: *SEPAPS 2020*; (2021).
- Gunnell KE, Longmuir PE, Barnes JD, Belanger K, Tremblay MS. Refining the Canadian assessment of physical literacy based on theory and factor analyses. *BMC Public Health*. (2018) 18(Suppl 2):131–45. doi: 10.1186/s12889-018-5899-2
- Mohammadzadeh M, Sheikh M, Houminian Sharif Abadi D, Bagherzadeh F, Kazemnejad A. Design and psychometrics evaluation of adolescent physical literacy questionnaire (APLQ). *Sport Sci Health*. (2022) 18(2):397–405. doi: 10.1007/s11332-021-00818-8
- Physical Literacy for Life Consortium. Physical literacy for life self-assessment tools (2021). Available at: <https://physical-literacy.isca.org/tools/> (Accessed December 5, 2021).
- Sum RKW, Cheng CF, Wallhead T, Kuo CC, Wang FJ, Choi SM. Perceived physical literacy instrument for adolescents: a further validation of PPLI. *J Exerc Sci Fit*. (2018) 16(1):26–31. doi: 10.1016/j.jesf.2018.03.002
- Sport Australia. Australian Physical Literacy Framework (2019). Available at: <https://nla.gov.au/nla.obj-2341259417> (Accessed March 4, 2020).
- Ministério da Educação. *Programa nacional educação física: Ensino secundário*. Lisboa: DES (2001).
- Ministério da Educação. *Programa nacional educação física (reajustamento): ensino básico 3º Ciclo*. Lisboa: DEB (2001).
- Mota J, Martins J, Onofre M. Portuguese Physical literacy assessment questionnaire (PPLA-Q) for adolescents: validity and reliability of the psychological and social modules using mokken scale analysis. *Percept Mot Skills*. (2023) 130(3):958–83. doi: 10.1177/00315125231159688
- Mota J, Martins J, Onofre M. Portuguese Physical literacy assessment–observation (PPLA-O) for adolescents (15–18 years) from grades 10–12: development and initial validation through item response theory. *Front Sports Act Living*. (2022) 4:1033648. doi: 10.3389/fspor.2022.1033648
- Mota J, Martins J, Onofre M. 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. *Res Sq*. (2022). doi: 10.21203/rs.3.rs-1458688/v2
- Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. *Psychol Methods*. (2000) 5(2):155–74. doi: 10.1037/1082-989X.5.2.155

Acknowledgments

We would like to acknowledge the invaluable contribution of all the tireless PE teachers and students who participated in these studies. Similarly, we thank António Rosado for his helpful insights during this project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fspor.2023.1192025/full#supplementary-material>

21. Henseler J. *Composite-based structural equation modeling: analyzing latent and emergent variables*. New York, NY: The Guilford Press (2021) (Methodology in the social sciences).
22. Sarstedt M, Hair JF, Ringle CM, Thiele KO, Gudergan SP. Estimation issues with PLS and CBSEM: where the bias lies! *J Bus Res.* (2016) 69(10):3998–4010. doi: 10.1016/j.jbusres.2016.06.007
23. Longmuir PE, Boyer C, Lloyd M, Yang Y, Boiarskaia E, Zhu W, et al. The Canadian assessment of physical literacy: methods for children in grades 4 to 6 (8 to 12 years). *BMC Public Health.* (2015) 15(1):767. doi: 10.1186/s12889-015-2106-6
24. Cairney J, Clark H, Dudley D, Kriellaars D. Physical literacy in children and youth—a construct validation study. *J Teach Phys Educ.* (2019) 38(2):84–90. doi: 10.1123/jtpe.2018-0270
25. Mota J, Martins J, Onofre M. 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.* (2021) 21(1):2183. doi: 10.1186/s12889-021-12230-5
26. Kock N. *WarpPLS*. Laredo, TX: ScriptWarp Systems (2020).
27. Kock N, Hadaya P. Minimum sample size estimation in PLS-SEM: the inverse square root and gamma-exponential methods: sample size in PLS-based SEM. *Inf Syst J.* (2018) 28(1):227–61. doi: 10.1111/isj.12131
28. Direção-Geral da Educação. Faculdade de Motricidade Humana. FITescola (2015). Available at: <https://fitescola.dge.mec.pt/home.aspx> (Accessed December 22, 2021)
29. Sijtsma K, van der Ark LA. A tutorial on how to do a mokken scale analysis on your test and questionnaire data. *Br J Math Stat Psychol.* (2017) 70(1):137–58. doi: 10.1111/bmsp.12078
30. Molenaar IW, Sijtsma K. Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschr Voor Onderwijsres.* (1984) 9(5):257–68.
31. Molenaar IW. A weighted Loevinger H-coefficient extending Mokken scaling to multcategory items. *Kwant Methoden.* (1990) 12(37):97–117.
32. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc.* (2003) 35(8):1381–95. doi: 10.1249/01.MSS.0000078924.61453.FB
33. Kim Y, Park I, Kang M. Convergent validity of the International Physical Activity Questionnaire (IPAQ): meta-analysis. *Public Health Nutr.* (2013) 16(3):440–52. doi: 10.1017/S1368980012002996
34. Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. *Int J Behav Nutr Phys Act.* (2011) 8:115. doi: 10.1186/1479-5868-8-115
35. RStudio Team. *RStudio: integrated development for R*. Boston, MA: RStudio, PBC (2020). Available at: <http://www.rstudio.com/>.
36. R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing (2020). Available at: <http://www.R-project.org/>.
37. Tierney N, Cook D, McBain M, Fay C. naniar: data structures, summaries, and visualisations for missing data (2021). Available at: <https://CRAN.R-project.org/package=naniar>.
38. Revelle W. psych: procedures for psychological, psychometric, and personality research (2021). Available at: <https://CRAN.R-project.org/package=psych> (Accessed October 5, 2021).
39. Little RJA. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc.* (1988) 83(404):1198–202. doi: 10.1080/01621459.1988.10478722
40. Korkmaz S, Goksuluk D, Zararsiz G. MVN: an R package for assessing multivariate normality. *R J.* (2014) 6(2):151–62. doi: 10.32614/RJ-2014-031
41. Leys C, Delacre M, Mora YL, Lakens D, Ley C. How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *Int Rev Soc Psychol.* (2019) 32(1):5. doi: 10.5334/irsp.289
42. Klein O, Delacre M. *Routliers: Robust Outliers Detection*. (2019). Available at: <https://CRAN.R-project.org/package=Routliers>
43. Fox J. polycor: polychoric and polyserial correlations (2019). Available at: <https://CRAN.R-project.org/package=polycor>.
44. Sport Australia. The Australian Physical Literacy Framework (2019). Available at: <https://nla.gov.au/nla.obj-2341259417> (Accessed April 3, 2020).
45. Roseel Y. Lavaan: an R package for structural equation modeling. *J Stat Softw.* (2012) 48(2):1–36. doi: 10.18637/jss.v048.i02
46. Huber PJ. *The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (1967). Vol. 5.1. p. 221–34.
47. Yuan KH, Bentler PM. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociol Methodol.* (2000) 30(1):165–200. doi: 10.1111/0081-1750.00078
48. Arbuckle J. Full information estimation in the presence of incomplete data. In: Marcoulides GA, Schumacker RE, editors. *Advanced structural equation modeling*. Mahwah, NJ: Psychology Press (1996):243–78.
49. Dong Y, Peng CYJ. *Principled missing data methods for researchers*. SpringerPlus. (2013) 2(222). doi: 10.1186/2193-1801-2-222
50. Gana K, Broc G. *Structural equation modeling with lavaan*. London: ISTE Ltd (2019). 273 p (Mathematics and statistics).
51. Kline RB. *Principles and practice of structural equation modeling*. 4th ed. New York: The Guilford Press (2016). 534 p. (Methodology in the social sciences).
52. Jonckere JD, Roseel Y. Using bounded estimation to avoid nonconvergence in small sample structural equation modeling (2021). Available at: <https://osf.io/l7z6j/> (Accessed November 26, 2021).
53. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model Multidiscip J.* (1999) 6(1):1–55. doi: 10.1080/10705519909540118
54. Schreiber JB, Nora A, Stage FK, Barlow EA, King J. Reporting structural equation modeling and confirmatory factor analysis results: a review. *J Educ Res.* (2006) 99(6):323–38. doi: 10.3200/JOER.99.6.323-338
55. Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociol Methods Res.* (2008) 36(4):462–94. doi: 10.1177/0049124108314720
56. Marsh HW, Hau KT, Wen Z. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct Equ Model Multidiscip J.* (2004) 11(3):320–41. doi: 10.1207/s15328007sem1103_2
57. Dueber D. BifactorIndicesCalculator (2021). Available at: <https://CRAN.R-project.org/package=BifactorIndicesCalculator>.
58. Satorra A, Bentler PM. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika.* (2001) 66(4):507–14. doi: 10.1007/BF02296192
59. Brown TA. *Confirmatory factor analysis for applied research*. 2nd ed. New York: The Guilford Press (2015). 462 p. (Methodology in the social sciences).
60. Comrey AL, Lee HB. *A first course in factor analysis*. New York: Psychology Press (1992).
61. Price LR. *Psychometric methods theory into practice*. New York: The Guilford Press (2017).
62. Nunnally J, Bernstein I. *Psychometric theory*. New York: McGraw-Hill (1994).
63. Dueber D. A bifactor approach to dimensionality assessment. University of Kentucky Libraries (2020). Available at: https://uknowledge.uky.edu/edsc_etds/59/ (Accessed November 29, 2021).
64. Rodriguez A, Reise SP, Haviland MG. Evaluating bifactor models: calculating and interpreting statistical indices. *Psychol Methods.* (2016) 21(2):137–50. doi: 10.1037/met0000045
65. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling: a bifactor perspective. *Educ Psychol Meas.* (2013) 73(1):5–26. doi: 10.1177/0013164412449831
66. Stucky BD, Edelen MO. Using hierarchical IRT models to create unidimensional measures from multidimensional data. In: Reise SP, Revicki DA, editors. *Handbook of item response theory modelling: applications to typical performance assessment*. New York: Routledge (2015):183–206.
67. Benitez J, Henseler J, Castillo A, Schuberth F. How to perform and report an impactful analysis using partial least squares: guidelines for confirmatory and explanatory IS research. *Inf Manage.* (2020) 57(2):103168. doi: 10.1016/j.im.2019.05.003
68. Diamantopoulos A, Sigauw JA. Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. *Br J Manag.* (2006) 17(4):263–82. doi: 10.1111/j.1467-8551.2006.00500.x
69. Cenfetelli B. Interpretation of formative measurement in information systems research. *MIS Q.* (2009) 33(4):689. doi: 10.2307/20650323
70. Hinkle DE, Wiersma W, Jurs SG. *Applied statistics for the behavioral sciences*. Boston, MA: Houghton Mifflin College Division (2003), Vol. 663.
71. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: L. Erlbaum Associates (1988). 567 p.
72. Schuberth F. Confirmatory composite analysis using partial least squares: setting the record straight. *Rev Manag Sci.* (2020) 15(1):1311–45. doi: 10.1007/s11846-020-00405-0
73. Rademaker M, Schuberth F. cSEM: composite-based structural equation modeling (2020). Available at: <https://m-e-rademaker.github.io/cSEM/>.
74. Schuberth F, Rademaker ME, Henseler J. Estimating and assessing second-order constructs using PLS-PM: the case of composites of composites. *Ind Manag Data Syst.* (2020) 120(12):2211–41. doi: 10.1108/IMDS-12-2019-0642
75. van Riel ACR, Henseler J, Kemény I, Sasovova Z. Estimating hierarchical constructs using consistent partial least squares: the case of second-order

composites of common factors. *Ind Manag Data Syst.* (2017) 117(3):459–77. doi: 10.1108/IMDS-07-2016-0286

76. Babic MJ, Morgan PJ, Plotnikoff RC, Lonsdale C, White RL, Lubans DR. Physical activity and physical self-concept in youth: systematic review and meta-analysis. *Sports Med.* (2014) 44(11):1589–601. doi: 10.1007/s40279-014-0229-z

77. Park H, Kim N. Predicting factors of physical activity in adolescents: a systematic review. *Asian Nurs Res.* (2008) 2(2):113–28. doi: 10.1016/S1976-1317(08)60035-3

78. Amireault S, Godin G, Vézina-Im LA. Determinants of physical activity maintenance: a systematic review and meta-analyses. *Health Psychol Rev.* (2013) 7(1):55–91. doi: 10.1080/17437199.2012.701060

79. Biggs J, Collis K. *Evaluating the quality of learning: the SOLO taxonomy (structure of observed learning outcomes)*. New York: Academic Press (1982).

80. Eid M. Multi-faceted constructs in abnormal psychology: implications of the Bifactor S-1 model for individual clinical assessment. *J Abnorm Child Psychol.* (2020) 48(7):895–900. doi: 10.1007/s10802-020-00624-9

81. Eid M, Geiser C, Koch T, Heene M. Anomalous results in G-factor models: explanations and alternatives. *Psychol Methods.* (2017) 22(3):541–62. doi: 10.1037/met0000083

82. Bollen KA, Bauldry S. Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychol Methods.* (2011) 16(3):265–84. doi: 10.1037/a0024448

83. Bollen KA, Diamantopoulos A. In defense of causal-formative indicators: a minority report. *Psychol Methods.* (2017) 22(3):581–96. doi: 10.1037/met0000056