Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

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 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1458709/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

Background

Aims of this study were to assess construct validity (dimensionality, measurement invariance, convergent and discriminant validity) and reliability of the previously developed Psychological and Social modules of the Portuguese Physical Literacy Assessment – Questionnaire (PPLA-Q).

Methods

Mokken Scale Analysis was used in 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).

Results

The 8 scales forming the 2 modules can be interpreted as moderate to strong Mokken scales with H coefficient ranging from .47 to .66; 4 of these scales had an interpretable Invariant Item Ordering (H^T>.30). Results suggest that all scales function similarly in male and female adolescents, except for the *Physical Regulation* scale which has shown evidence of a sex bias. All scales had good total score reliability ρ > .80, ranging from .93 to .94; regarding test-retest reliability: 3 scales had moderate to excellent reliability ($ICC_{95\%CI}$ lower bound ranging from .63 to .85. and upper bound from .84 to .95), with the remaining 5 scales presenting a large spread in estimated reliability likely due to change during COVID-19 lockdown. Scales score correlated as theoretically expected, with low to moderate across domain correlations providing support of convergent and discriminant validity.

Conclusions

Evidence supports the construct validity and score reliability of the psychological and social modules of the PPLA-Q to assess the psychological and social domains of Physical Literacy in the Portuguese PE context for grade 10–12 (15–18 years) adolescents. Further scrutiny of test-retest reliability is warranted in future research.

Background

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 (Physical Literacy for Life, 2021), and reap the widely documented physical, cognitive, affective and social benefits of PA (Australian Government Department of Health, 2019; World Health Organization, 2020). This would help counter the scenario where 27.5% of adults worldwide still fail to meet PA guidelines (Guthold et al., 2018). The scenario for adolescents (11–17 years) is much worse, with 81% failing to meet these guidelines (Guthold et al., 2020). Further studies in Portugal detail that among adolescents, high-schoolers (grade 10–12) seem to have lower PA levels and increased sedentary behavior than their younger peers (Baptista et al., 2012; Matos & Equipa Aventura Social, 2018).

Quality physical education (PE), as a mandatory, free and qualified environment, is exhorted as a central piece of the solution to address this issue through the development of PL (Guthold et al., 2020; UNESCO, 2015). If this goal is to

be achieved, assessment is an essential part of this endeavor to track and understand progress (Corbin, 2016).

The Portuguese Physical Literacy Assessment (PPLA) is a tool composed by two parts (a questionnaire, PPLA-Q (Mota et al., 2021), and an observational instrument, PPLA-O) developed to be used in PE to provide a feasible and holistic assessment of the PL of grade 10 to 12 students. This tool was inspired by the Australian Physical Literacy Framework (APLF; Sport Australia, 2019) and by the outcomes and didactic philosophy of the Portuguese PE syllabus (Ministry of Education [Ministério da Educação], 2001a, 2001b). PPLA-Q features three modules: psychological, social and cognitive; assessing a selection of elements from the APLF (Mota et al., 2021).

Among the psychological and social modules are elements that are posited as determinants of PA participation in adolescents (Cortis et al., 2017), and associated with multiple beneficial outcomes inside and outside of PE (Li et al., 2008; Pozo et al., 2018). For each of its eight elements, two Likert-type subscales were developed with differing difficulties: one to measure foundational skills and attitudes (*Foundation*), and another targeting a higher degree of development (*Mastery*). Both are posited to stand along a learning continuum, according to an integration of the learning taxonomies of Structure of Observed Learning Outcomes (Biggs & Collis, 1982), and Bloom's Affective taxonomy (Krathwohl et al., 1964). The choice to separate a continuum into two subscales was based on an initial Classical Test Theory-based development framework, whose dimensionality assessment methods (i.e., linear factor analysis) are prone to grouping together items (i.e., creating a method factor), based on difficulty (Sijtsma & Ark, 2021; van Schuur, 2003).

However, Item Response Theory models provide a solution to this issue, explicitly modelling difficulty as a parameter. Within this large class of models, nonparametric models (NIRT) – like those included in Mokken Scale Analysis (MSA; (Sijtsma & Ark, 2021)) have been pointed out as particularly useful for affective variables (e.g, Reise & Waller, 2009), since their underlying response processes might not conform to more rigidly defined response patterns implied by parametric models (van Schuur, 2003; Wind, 2017).

Previous research has highlighted differences in adolescents across sexes in both PA participation (Guthold et al., 2020) and other elements included in the PPLA-Q scales (Vaquero-Diego et al., 2020; Vasconcellos et al., 2019). However, before any meaningful comparisons can be drawn, differential item, and test, functioning (DIF and DTF) analysis are warranted, to provide evidence for the measurement invariance across sexes at item-level and scale-level (Gamerman et al., 2019; Moorer et al., 2001; Teresi et al., 2008). Another important psychometric quality, is test-retest reliability; which assesses temporal consistency of scores derived from an instrument at different time points (Price, 2017), distinguishing random short-term scores differences from true change (Polit, 2014).

This study was part of a larger project to validate all measures of PPLA and aimed to a) investigate dimensionality, convergent and discriminant validity, measurement invariance (DIF and DTF), and b) reliability (total-score and test-retest) of the psychological and social modules of the PPLA-Q in Portuguese grade 10 to 12 (15–18 years) adolescents through MSA.

Methods

Participants

Main study

A convenience sample was used consisting of 521 grade 10–12 students from 25 classes in 6 public schools in the metropolitan Lisbon area, out of 611 available students (15% attrition rate, down from 30% in previous pilot study). Due to COVID-19 restrictions, only schools with a PE preservice protocol with the Faculty of Human Kinetics were selected. To increase representativeness, recruitment was stratified by grade, and course major. We drew target percentage quotas according to student numbers reported for the school year of 2017/2018 (Ministério da Educação [Ministry of Education], 2019): 37%, 32% and 31% from grades 10, 11 and 12 respectively; regarding course major, initial target percentages were: Science, Technology, Engineering and Maths (STEM; 52%), Humanistic and Linguistics studies (29%), Economical studies (13%), and Visual Arts (6%). We also chose schools from as diverse as possible socioeconomic backgrounds – based on information in each schools' educational project. 13 students missed class on data collection's day and were removed from this study. Table 8 sums up the main characteristics of the final sample used in the analysis which adhered to target quotas within marginal variation (< 5%), conforming to acceptable sample sizes for the used methodology: Mokken Scale Analysis (MSA) (Mokkink et al., 2018; Straat et al., 2014).

Retest

A subsample of 72 students was used for retest application (Table 8). Minimum sample size (N = 64) estimation was based on a power analysis for an expected Intraclass Correlation Coefficient (ICC) of .80, with .10 precision in its 95% confidence interval (Bonett, 2002), accounting for 20% of subject attrition – using an online calculator (Arifin, 2020). Given time and COVID-19 constraints, no stratification was possible.

Table 8
Sample characteristics

	Baseline	Retest
Characteristic	**N = 508**[1]	**N = 72**[1]
Sex		
Female	299 (59%)	40 (56%)
Male	209 (41%)	32 (44%)
Age	16 (1)	16 (0)
Grade		
10	204 (40%)	-
11	137 (27%)	72 (100%)
12	167 (33%)	-
Major		
Economics	75 (15%)	-
Humanities	165 (32%)	
STEM	268 (53%)	72 (100%)
School		
School 1	39 (7.7%)	-
School 2	61 (12%)	-
School 3	21 (4.1%)	-
School 4	69 (14%)	-
School 5	207 (41%)	72 (100%)
School 6	111 (22%)	-
STEM – Sciences, Technology, Engineering and Math		
[1] *Statistic presented*: n (%); Mean(SD)		

Measures

PPLA-Q is a questionnaire developed to assess the psychological, social, and part of the cognitive domains of Physical Literacy in Portuguese adolescents. Evidence supporting its content validity has been previously established (Mota et al., 2021). The psychological and social modules, in their current development version (v0.6) are comprised of 46 and 43 Likert-type items, respectively, divided in eight elements: (1) *Motivation*, (2) *Confidence*, (3) *Emotional Regulation*, and (4) *Physical Regulation* in the psychological module; and (5) *Culture & Society*, (6) *Ethics*, (7) *Collaboration*, and (8) *Relationships* in the social module (Table 9 and Table 10). All items used a consistent 5-points unipolar response scale. Response points were fully labelled, using both numeric and verbal labels, (0 = *Not at all*; 1 =

Slightly; 2 = *Moderately*; 3 = *Quite a lot*; 4 = *Totally*), measuring student's identification with each of the statements (*How much do the following statements describe you?)*.

Procedures

Main study

The PPLA-Q was self-administered during PE classes to increase response rate, supervised by the lead author from January to March 2021. The short form of the *International Physical Activity Questionnaire* (IPAQ-SF; Craig et al., 2003) was also applied for further validation studies. Data collection initially started in paper format, however, due to COVID-19 lockdown only 3 out of 25 classes sampled used this format (n = 60). Data collection resumed in online format using LimeSurvey (LimeSurvey GmbH, 2021) for the remaining classes. Participants were informed of the questionnaire's goals, anonymity and encouraged to provided honest answers through a standardized initial instruction. Average completion time (n = 452) was 5.5 (2.2) and 4.6 (1.8) minutes for the social modules, respectively.

Retest

Second application of the PPLA-Q occurred in online format, 15 days apart from first application to reduce carryover effects (Nunnaly & Bernstein, 1994). IPAQ-SF was not applied to this recurrent sample. Remaining procedures were equal. Average completion time (n = 73) was 3.8 (1.2) minutes and 3.3 (1.1) minutes, for the psychological and social modules, respectively.

Analysis

All analysis were performed in RStudio (RStudio Team, 2020) with R 4.1.0 (R Core Team, 2020). Negatively stated items (S15, *Ethics* scale) and items P2 – P6 (*Motivation* scale) were reversed so that an increase in score would correspond to an increase in each assessed element. Resulting from the application in paper format, nine items had one missing response (0.2%). For these items, values were imputed using two-way imputation (Bernaards & Sijtsma, 2000).

Dimensionality

Given IRT and specifically, MSA's models cumulative nature (i.e., recognizing that different items might have different difficulty levels which might influence their endorsement; van Schuur, 2003), we chose to analyze each element in a single scale, coherent with the logic of a continuum that led to their development, instead of separating them into two different subscales based on difficulty.

Prior to MSA, Guttman errors – a non-parametric IRT (NIRT) person fit statistic – were calculated by scale and values that surpassed Tukey's upper fence were deemed as outliers (Zijlstra et al., 2011). These ranged from 23 to 35 students depending on the scale (5–7% of sample size). Sensibility analysis revealed that these outliers greatly affected the scalability coefficients for each scale, and so were removed from further analysis (Sijtsma & van der Ark, 2017).

The freeware RStudio (RStudio Team, 2020) with *R* version 4.1.0 (R Core Team, 2021) was used for all statistical analysis. MSA results and total-score reliability coefficients were calculated within the *mokken* package (Ark, 2012); while ICC were obtained with the *irr* package (Gamer et al., 2019).

MSA was used in a confirmatory manner to test the dimensionality and total-score reliability of each scale, through fitting of the polytomous Monotone Homogeneity Model (MHM) and polytomous Double Monotonicity Model (DMM).

Unidimensionality assumption was assessed using the 95% confidence intervals for scalability coefficients at item (H_i) and scale level (H). For H_i, a .30 cutoff was used (Ark, 2012): non-conforming items were eliminated one by one, after evaluating the impact on content representativeness and their scalability with other items in the scale. H for final scales were evaluated using the criteria of: $H \geq .50$, $.40 \leq H < .50$, and $.30 \leq H < .40$, for strong, medium and weak scales respectively (Sijtsma & Molenaar, 2002).

Local independence was assessed through the conditional association procedure (Sijtsma et al., 2015; Straat et al., 2016). Pairs of items flagged by the *mokken* package for positive local dependence (PLD; W_1 and W_2 statistic) or negative local dependence (NLD; W_3 statistic) were examined regarding their content, and the least representative item was deleted in each pair before the analysis was rerun.

Monotonicity and Invariant Item Ordering (IIO) were assessed through the *crit* statistic – which indicates the seriousness of violations for each of the assumptions – for each item, using a cutoff of crit < 40 (Stochl et al., 2012). Analysis of IIO was supplemented by graphical analysis of pairwise Item Response Functions (IRF) to assess non-intersection (Sijtsma et al., 2011; Wind, 2017). After IIO was established, *Htrans* (H^T) coefficient was calculated using Manifest Item Invariant Ordering to assess the accuracy and usefulness of said IIO; evaluation used the criteria of $H^T \geq .50$, $.40 \leq H^T < .50$, and $.30 \leq H^T < .40$ for high, medium and low accuracy, respectively (Ligtvoet et al., 2010).

For scales in which clusters of unscalable items and/or borderline scalability coefficients ($H_{i95\%CI} \approx .30$) were identified, further exploratory analysis was performed using both the Automatic Item Selection Procedure (AISP) and Genetic Algorithm (GA) features available in the *mokken* package. These were run from lower-bound c = .30 to .60 in incremental steps of .05 to detect changes in clustering patterns of items at different scalability thresholds (Hemker et al., 1995; Sijtsma & van der Ark, 2017). Clusters discovered with these features were then submitted to a confirmatory analysis, using the procedures previously presented.

Measurement invariance

We assessed whether DIF and DTF according to sex was present in each scale by calculating scalability for each item (H_i) and scale (H) for the female and male subgroup (Sijtsma & van der Ark, 2017; Wind, 2017). We then analyzed its difference, and its statistical significance (at p = .05): non-intersecting 95%CI for both coefficients were considered as evidence of statistically significant differences between sexes.

Reliability

Molenaar and Sijtsma ρ(1988) was calculated as an unbiased measure of test-score reliability for each of the final scales. Its interpretation follow the same cutoffs as those of Cronbach's α (Cronbach, 1951): with $\rho > .70$ considered as acceptable (Nunnaly & Bernstein, 1994) and $\rho > .80$ being recommended (Price, 2017). For comparison purposes with previous studies and readers accustomed to CTT, we also computed α coefficient.

To establish total score test-retest reliability we computed Intraclass Correlation Coefficient (ICC) and its 95%CI according to a single rater, absolute agreement, two-way mixed effect model (formula 2.1 in Koo & Li, 2016), using sum scores of the final scales in both time points (Liljequist et al., 2019). Evaluation followed the criteria of: ICC > .90, excellent; $75 \leq ICC < .95$, good; $.50 \leq ICC < .75$, moderate; and ICC < .50, poor test-retest reliability.

Discriminant and convergent validity

Bivariate Spearman correlations (and its 95% CI) were calculated among total summed scored using the *RVAideMemoire* (Hervé, 2021) package with 1000 bootstrap replications. These correlations were then disattenuated for measurement error using obtained ρ coefficients as $r_{xy}/\sqrt{\rho_x * \rho_y}$ (Murphy & Davidshofer, 2005), and used to evaluate discriminant validity (threshold of $r = .85$ to discern whether variables were statistically different) and convergent validity based on magnitude reported in similar studies. Interpretation of magnitudes followed (Hinkle et al., 2003) guidelines: r > .90, >.70, > .50, >.30, as very high, high, moderate, and low correlations, respectively.

Results

Item response frequencies and difficulty

Table 9 displays the response frequencies in each response category, as well as mean for each item in the psychological and social modules of the PPLA-Q. No response option had higher than 55% frequency, suggesting a balanced distribution of responses across options; 9 items (10%) had no responses in their lowest response option (0 – "*Not at all*"). As expected, items developed to represent a higher development in each element (i.e., *Mastery*) had overall lower mean values (i.e., higher difficulty) than their less complex (i.e., *Foundation*) counterparts.

Table 9
Percent response frequencies for the Psychological Module of the PPLA

Scale	Level	Label	Content[1]	Mean (SD)	Frequency per response option (%)				
					0	1	2	3	4
Motivation	Global	P1	I am motivated to practice PA	2.6 (1.0)	2	11	37	30	20
	Foundation	P2 [R]	I practice PA because others tell me I should	3.0 (1.1)	1	10	17	31	41
		P3 [R]	I feel guilty when I do not practice PA	2.0 (1.2)	11	30	25	20	14
		P4 [R]	I feel bad about myself when I do not practice PA	1.9 (1.2)	14	28	26	21	11
		P5 [R]	I feel pressured by others to practice PA	3.3 (1.0)	2	4	11	28	55
		P6 [R]	I practice PA because I feel others would be unhappy if I did not	3.6 (0.8)	0	2	7	18	72
	Mastery	P7	I practice PA because it is fun	2.5 (1.1)	4	11	31	34	20
		P8	I feel good when I practice PA	3.2 (0.8)	1	2	16	36	45
		P9	I consider PA a part of me	2.4 (1.3)	8	21	25	22	25
		P10	I value the benefits of PA	3.3 (0.8)	1	3	10	38	48
		P11	I see PA as a fundamental part of who I am	2.2 (1.2)	10	22	27	22	19
		P43	I feel more motivated to reach my goals because I practice PA	2.4 (1.1)	5	15	27	35	17
Confidence	Global	P13	I feel confident to practice PA	2.7 (1.1)	5	11	24	34	26
	Foundation	P14	* I am confident in my abilities	2.4 (1.0)	5	11	37	33	14
		P15	* I can participate with success	2.6 (0.9)	1	8	33	42	16
		P16	* I consider myself competent	2.5 (1.0)	4	11	36	34	15
		P17	* I have trust in my skills	2.5 (1.1)	3	14	33	29	20
		P18	* I feel good about the way I can participate	2.5 (1.0)	3	12	31	34	19
	Mastery	P19	* I can participate in PA that I consider challenging	2.5 (1.0)	2	16	31	36	15

Scale	Level	Label	Content[1]	Mean (SD)	Frequency per response option (%)				
					0	1	2	3	4
		P20	* I know how to become more confident in myself	2.2 (1.1)	8	20	33	28	12
		P21	* I feel competent even when I am criticized	2.3 (1.2)	7	18	28	29	18
		P22	* I believe in myself even when I lose	2.3 (1.1)	6	18	32	28	16
		P44	** I feel more confident in my skills because I practice PA	2.5 (1.1)	5	14	29	33	19
Emotional Regulation	Global	P23	* I can manage my emotions	2.4 (1.1)	5	17	31	32	15
	Foundation	P24	* I can recognize other's emotions	2.8 (0.9)	2	3	27	47	21
		P25	* I can recognize my emotions	2.8 (0.9)	2	6	24	43	26
		P26	* I am sensitive to the feelings of others	2.7 (0.9)	2	7	29	43	19
		P27	* I understand what others feel	2.6 (0.9)	2	6	35	42	16
		P28	* I can identify what I feel	2.7 (0.9)	2	8	26	44	19
	Mastery	P29	* I can anticipate what I will feel	2.2 (1.0)	5	19	41	28	8
		P30	* I can deal with difficulties rationally	2.6 (0.9)	1	10	36	38	15
		P31	* I can manage my emotions when necessary	2.5 (1.0)	3	10	35	37	15
		P32	* I have a good control of my emotions	2.3 (1.0)	4	15	36	33	12
		P45	** I am better at controlling my emotions because I practice PA	1.9 (1.2)	13	25	33	18	10
Physical Regulation	Global	P33	* I can manage my effort	2.6 (0.9)	1	8	33	44	15
	Foundation	P34	* I know when I am tired	3.3 (0.8)	0	1	10	45	44
		P35	* I can recognize changes in my breathing	3.3 (0.8)	1	1	10	44	44
		P36	* I can recognize changes in my heart rate	3.2 (0.8)	1	3	9	39	48
		P37	* I recognize my physical limits	2.8 (0.9)	1	4	13	42	40

Scale	Level	Label	Content[1]	Mean (SD)	Frequency per response option (%)				
					0	1	2	3	4
		P38	* I can recognize the effect that different intensities have in me	3.0 (0.8)	1	7	25	42	25
	Mastery	P39	* I use strategies to manage my effort	2.3 (1.0)	1	3	19	48	28
		P40	* I can anticipate when I will be fatigued	2.4 (1.0)	4	20	34	29	13
		P41	* I can control my fatigue	2.0 (1.0)	4	17	33	33	13
		P42	* I take action to improve my physical skills	2.9 (1.0)	6	27	40	22	6
		P46	** I am better at controlling my fatigue because I practice PA	2.5 (1.1)	2	9	24	33	32

[1] General item stem: "*How much do the following statements describe you?";* [R]Reverse-coded item

[*] Specific item stem: "In Physical Activity Contexts:"; [**] Specific item stem: "In the different contexts of my life:"

Table 10
Percent response frequencies for the Social Module of the PPLA

Scale	Level	Label	Content[1]	Mean(SD)	Frequency per response option (%)				
					0	1	2	3	4
Culture	Global	S1	I believe that the cultural aspects of PA are important (e.g., its rituals, terminology, clothing, values)	2.5 (1.1)	5	15	28	35	18
	Foundation	S2	I participate in PA rituals (e.g., greetings, hymns/chants, cheers, applauses)	1.6 (1.3)	29	23	22	15	12
		S3	I use specific PA terminology (e.g., names of technics and tactics, names of equipment, idioms)	2.2 (1.1)	7	24	30	26	14
		S4	I use specific clothing of the PA I am practicing	3.0 (1.1)	3	8	16	31	43
		S5	I watch PA events (e.g., competitions, spectacles, shows)	2.4 (1.2)	8	18	25	26	23
	Mastery	S6	I like to keep up with PA events (e.g., competitions, spectacles, shows)]	2.5 (1.3)	8	17	22	26	27
		S7	I am interested in the cultural aspects of PA (e.g., its rituals, terminology, clothing, values)]	1.9 (1.2)	11	26	31	22	11
		S8	I encourage others to watch PA events (e.g., competitions, spectacles, shows)	1.6 (1.3)	25	29	20	16	10
		S9	I encourage others to participate in each PA's culture (e.g., rituals, terminology, clothing)]	1.4 (1.2)	28	29	24	13	7
		S40	** I am more involved in other cultural activities (e.g., theater, music) because I practice PA	1.3 (1.2)	32	26	24	13	6
Ethics	Global	S12	* I try to behave correctly and justly	2.4 (0.7)	1	8	42	50	0
	Foundation	S13	* I respect my adversaries	3.4 (0.8)	1	1	9	37	52
		S14	* I follow the rules	3.4 (0.7)	0	1	7	44	48
		S15	* I cheat if it brings me benefits	3.2 (1.0)	2	4	12	32	50

Scale	Level	Label	Content[1]	Mean(SD)	Frequency per response option (%)				
					0	1	2	3	4
		S16	* I respect the decisions of authorities (e.g., referee, umpire, coach/teacher)]	3.2 (0.9)	1	3	16	39	41
		S17	* I behave according to fair-play / sport ethics 'principles	3.3 (0.8)	1	1	12	38	48
	Mastery	S18	* I understand the importance of fair play/ sport ethics' principles	3.5 (0.8)	1	1	9	27	62
		S19	* I take action to make others behave according to fair play/sport ethic	2.8 (1.0)	3	7	23	38	29
		S20	* I follow the rules, even if unsupervised	3.1 (0.8)	1	2	17	45	35
		S21	* I behave according to fair play/sport ethics' principles on my initiative	3.2 (0.9)	1	3	16	36	44
		S22	* I take action for others to follow the rules	2.7 (1.0)	3	8	27	39	22
		S41	** I am more honest and just because I practice PA	1.9 (1.2)	15	24	28	26	7
Collaboration	Global	S23	* I collaborate with others	3.3 (0.7)	0	1	15	47	36
	Foundation	S24	* I am sympathetic with others	3.2 (0.7)	0	1	15	51	32
		S25	* I control my behavior towards others	3.1 (0.7)	0	0	4	37	59
		S26	* I respect others	3.5 (0.6)	0	0	7	48	45
		S27	* I cooperate with others	3.4 (0.6)	1	4	17	39	39
	Mastery	S28	* I encourage others	3.1 (0.9)	3	4	23	40	30
		S29	* I care about others' success	2.9 (1.0)	1	3	21	44	31
		S30	* I help others achieve success	3.0 (0.8)	1	2	17	51	31
		S31	* I am helpful to others	3.1 (0.8)	11	19	33	28	8
		S42	** I collaborate more with others because I practice PA	2.0 (1.1)	8	20	35	29	8
Relationships	Global	S32	* I have a positive relationship with others	3.2 (0.7)	0	1	12	51	35
	Foundation	S33	* I interact with others	3.1 (0.8)	0	3	17	45	34

Scale	Level	Label	Content[1]	Mean(SD)	Frequency per response option (%)				
					0	1	2	3	4
		S34	* I share a common goal with others	2.8 (0.9)	2	6	27	39	26
		S35	* I feel close to others	2.7 (1.0)	2	8	29	41	21
		S36	* I feel a sense of camaraderie with others	2.8 (0.9)	2	7	24	44	22
	Mastery	S37	* I take action to improve my relationship with others	2.9 (0.9)	2	6	22	42	28
		S38	* I know how to improve my relationship with others	2.6 (0.9)	1	10	35	37	16
		S39	* I care about my relationship with others	2.9 (1.0)	4	6	20	41	30
		S43	** I have better relationships with others because I practice PA	2.0 (1.1)	12	20	32	28	8

[1] General item stem: "*How much do the following statements describe you?";* [R]Reverse-coded item

* Specific item stem: "In Physical Activity Contexts:"; ** Specific item stem: "In the different contexts of my life:"

Dimensionality

Scalability

In the psychological module, 9 items were deemed unscalable since the confidence interval for their H_i included the cut-off value of .30 (or a lower value) (Table 11–12): 4 of these items were in the *Motivation* scale, with items P3 and P4 – both pertaining to introjected regulation – displaying high scalability between each other (H_{ij} = .74); 3 were in the *Emotional Regulation* scale, where items P24, P26 and P27 had high scalability between each other (Hij = .64 to .78) suggesting the existence of an item cluster pertaining to evaluation of other's emotions (e.g., P27 – "I understand what others feel"); and the remaining item in the *Physical Regulation* scale.

In the social module, 6 items were in same condition: 2 in the *Culture* scale; 2 in the *Ethics* scale, one of which was the single reverse-scored item of the scale; 1 in the *Collaboration* scale; and 1 in the *Relationships* scale. None of the unscalable items displayed a clustering pattern (i.e., high scalability between otherwise unscalable items), however, 4 of these 6 unscalable items were developed to assess the highest level of development in each corresponding scale – the capability to transfer the social skills developed in a PA context to other contexts – which suggests that these items might need a conceptual overhaul to accurately target this phenomenon. All 15 items were removed in a stepwise manner, ensuring that the remaining items in each scale conformed to the .30 cutoff.

Local Independence

Using the Conditional Association procedure, 3 psychological module items were flagged for likely being in a PLD pair with other(s) item(s) in the same scale (Table 11–13 column 3; 1 in the *Motivation* scale, and 2 in the *Physical Regulation*). For the social module, this number increased to 8 items (Table 15–17, column 3; 1 in the *Culture* scale, 3

in the *Ethics* scale, 2 in the *Collaboration* scale and 2 in the *Relationships* scale). Most identified pairs were within the same lower-level structure (i.e., foundation or mastery), and within the same specific trait (e.g., P9 and P11 with the same motivational regulation) with similar wording. Within each pair, an item was chosen to be removed according to its content relevancy to the scale, resulting in the removal of 11 items total.

Monotonicity

Graphical analysis of each Item Response Function (IRF), supplemented by the *crit* statistic in the *mokken* package revealed no significant violations of the monotonicity assumption (all crit = 0). As such, all scales conformed with the Monotone Homogeneity Model, suggesting that the relative ordering of students according to each construct (scale) is consistent across its items.

Invariant Item Ordering (IIO) and total scalability

During IIO analysis of both the IRFs and the corresponding crit statistic, 2 items in the *Confidence* scale (P15 and P17), and 1 item in both the *Ethics* and the *Collaboration* (S16 and S25, respectively) scales revealed statistically significant intersections with other IRF within the same scale (crit > 40) and were removed so that scales conformed with the additional requirement of the Double Monotonicity Model. Table 19 displays the resulting scales' total scalability coefficients (H) and IIO coefficients (H^T). Based on the 95% confidence interval around their H point-estimate, 2 scales formed medium to strong (*Motivation*, and *Physical Regulation*), while the remaining 6 formed strong Mokken hierarchical scales (H lower bound > .50, point-estimates ranging from .50 to .66). Despite displaying formal IIO (through non-intersection of IRFs), 4 of the scales (*Confidence*, *Emotional Regulation*, *Collaboration*, and *Relationships*) had an estimated H^T lower that .30 (.08, .19, .26, and .22, respectively), suggesting that such ordering might be too inaccurate for practical purposes (Ligtvoet et al., 2010) – students might perceive different items as having equivalent difficulty. The remaining 4 scales displayed better prospects for such ordering, with their IIO accuracy as weak (*Motivation*, and *Culture*), medium (*Physical Regulation*) and strong (*Ethics*).

Additional dimensionality analysis – Exploratory Mokken Scalling

Motivation

We noticed a pattern of borderline $CI_{95\%}$ lower bound values for H_i in items P2 and P5 in the *Motivation scale*; additionally, as previously mentioned, items P3 and P4 showed a high degree of scalability between each other. At c = .30 both the AISP and GA algorithms clustered P3 and P4 into a separate scale, and at c = .35 the items formed 3 clusters, coherent with different motivational regulations in SDT (Ryan & Deci, 2017), with the more autonomous regulations clustered together (Cluster 1 – External regulation; Cluster 2 – Introjected Regulations, Cluster 3 – integrated and internal regulations); this pattern persisted at higher c values, with P10 (the single item pertaining to identified regulation) becoming unscalable past c = .45. Further confirmatory analysis of these clusters (Table 11, columns 5–7) revealed that they formed two strong Mokken scales (since Cluster 2 was composed of only two items, we did not consider it for this effect) conforming with the DMM: Cluster 1 (H = .61, H^T = .50) and Cluster 3 (H = .60, H^T = .56) – after removal of items flagged in local dependence pairs.

Table 11
Mokken Scaling Analysis (MSA) abbreviated results for the Motivation scale of the PPLA-Q; n = 481

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 284 / n_{male}= 197	Exploratory MSA (AISP + GA) – c = .45			
		Removed items	Final H_i [95%CI]	ΔH_i	Removed items	Cluster 1	Cluster 2	Cluster 3
P1	2.6 (0.9)		.56 [.52, .61]	− .01				.62 [.57, .67]
P2	3.0 (1.0)		.38 [.31, .45]	− .01		.58 [.49, .66]		
P3	1.9 (1.2)	us: − .01 [-.08, .06] (2)					.74 [.68, .80]	
P4	1.8 (1.2)	us: .02 [.04, .09] (1)					.74 [.68, .80]	
P5	3.3 (0.9)		.39 [.32, .46]	− .12		.64 [.56, .72]		
P6	3.6 (0.7)	us: .36 [.28, .43] (4)				.60 [.51, .69]		
P7	2.6 (1.0)		.46 [.40, .51]	.04				.57 [.52, .63]
P8	3.3 (0.8)		.50 [.45, .55]	.00				.60 [.54, .65]
P9	2.4 (1.2)	PLD_{P11} ($W_{1 =}$ 8.02) and PLD_{P2} ($W_{1 =}$ 7.33) (5)			PLD_{P8} ($W_{1 =}$ 2.86) and PLD_{P11} ($W_{1 =}$ 4.05) (2)			
P10	3.3 (0.7)	us: .33 [.27, .40] (3)			us			
P11	2.2 (1.2)		.54 [.49, .58]	− .05				.61 [.57, .66]
P43	2.5 (1.1)		.46 [.41, .51]	− .04	NLD_{P7}($W_{3 =}$ 8.36) (1)			

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 284 / n_{male}= 197	Exploratory MSA (AISP + GA) – c = .45			
		Removed items	Final H_i [95%CI]	ΔH_i	Removed items	Cluster 1	Cluster 2	Cluster 3
		H [95%CI]	.47 [.43, .51]	– .03		.61 [.53, .68]	.74 [.68, .80]	.60 [.56, .65]

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (4) Item removal order

DIF – Differential item functioning; us – unscalable item; PLD_k – positive local dependence (subscripted item pair); NLD – negative local dependence (subscripted item pair)

Table 12
Mokken Scaling Analysis (MSA) abbreviated results for the Confidence scale of the PPLA-Q; n = 474

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 279 / n_{male}= 195
		Removed items	Final H_i [95%CI]	ΔH_i
P13	2.7 (1.1)		.71 [.68, .75]	− .07
P14	2.4 (1.0)		.70 [.66, .74]	− .06
P15	2.6 (0.9)	IIO_{crit} = 48 (2)		
P16	2.5 (1.0)		.67 [.64, .71]	− .05
P17	2.5 (1.0)	IIO_{crit}= 94 (1)		
P18	2.5 (1.0)		.71 [.67, .74]	− .01
P19	2.5 (1.0)		.64 [.60, .68]	− .05
P20	2.1 (1.1)		.61 [.57, .66]	.00
P21	2.4 (1.1)		.65 [.60, .69]	− .07
P22	2.3 (1.1)		.64 [.60, .69]	− .06
P44	2.5 (1.1)		.58 [.53, .64]	− .09
		H [95%CI]	0.66 [0.62, 0.69]	− .05

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (2) Item removal order

DIF – Differential item functioning; us – unscalable item; IIO – Invariant Item Ordering

Table 13
Mokken Scaling Analysis (MSA) abbreviated results for the Emotional Regulation scale of the PPLA-Q; n = 482

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 285 / n_{male}= 197	Exploratory MSA (AISP + GA) – c = .45	
		Removed items	Final H_i [95%CI]	ΔH_i	Cluster 1	Cluster 2
P23	2.4 (1.0)		.56 [.50, .62]	.08	.56 [.50, .62]	
P24	2.9 (0.8)	us: .31 [.23, .38] (2)		-		.66 [.60, .72]
P25	2.9 (0.9)		.57 [.51, .63]	.02	.57 [.51, .63]	
P26	2.8 (0.8)	us: .26 [.18, .34] (3)		-		.71 [.66, .77]
P27	2.7 (0.8)	us: .28 [.20, .35] (4)		-		.69 [.63, .75]
P28	2.7 (0.9)		.58 [.52, .64]	.05	.58 [.52, .64]	
P29	2.2 (0.9)		.51 [.45, .57]	.01	.51 [.45, .57]	
P30	2.6 (0.9)		.57 [.52, .62]	.00	.57 [.52, .62]	
P31	2.5 (0.9)		.61 [.57, .66]	.08	.61 [.57, .66]	
P32	2.4 (1.0)		.64 [.60, .69]	.07	.64 [.60, .69]	
P45	1.9 (1.1)	us: .21 [.15, .28] (1)		-	us	
		H [95%CI]	.58 [.53, .62]	.05	.58 [.53, .62]	.69 [.63, .74]

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (4) Item removal order; * intersecting 95% confidence intervals

DIF – Differential item functioning; us – unscalable item

Table 14

Mokken Scaling Analysis (MSA) abbreviated results for the Physical Regulation scale of the PPLA-Q; n = 485

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 288 / n_{male}= 197	Exploratory MSA (AISP) – c = .45[1]		
		Removed items	Final H_i [95%CI]	ΔH_i	Removed items	Cluster 1	Cluster 2
P33	2.7 (0.8)		.46 [.40, .52]	− .13			.53 [.47, .59]
P34	3.3 (0.7)	us: .31 [.24, .39] (1)		-	PLD_{P35} (W_{1} = 2.96) (1)		
P35	3.3 (0.8)		.46 [.40, .53]	− .23*		.62 [.55, .69]	
P36	3.2 (0.8)		.45 [.38, .51]	− .16		.57 [.51, .64]	
P37	2.8 (0.9)		.41 [.35, .47]	− .12		.50 [.42, .58]	
P38	3.0 (0.8)		.49 [.43, .55]	− .16		.56 [.50, .63]	
P39	2.3 (1.0)		.52 [.47, .57]	− .20*			.57 [.52, .63]
P40	2.4 (1.0)	PLD_{P39} (W_{1} = 7.81) (2)		-	us		
P41	2 (0.9)		.46 [.40, .51]	− .20*			.58 [.53, .64]
P42	2.9 (1.0)		.42 [.36, .48]	− .14			.54 [.48, .59]
P46	2.5 (1.0)	PLD_{P39} (W_{1} = 7.10) (3)		-			.56 [.50, .61]
		H [95%CI]	.46 [.41, .50]	− .17*		.56 [.50, .62]	.56 [.51, .60]

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (3) Item removal order; *intersecting 95% confidence intervals

DIF – Differential item functioning; us – unscalable item; PLD_k – positive local dependence (subscripted item pair)

Table 15
Mokken Scaling Analysis (MSA) abbreviated results for the Culture scale of the PPLA-Q; n = 490

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 288 / n_{male}= 202
		Removed items	Final H_i [95%CI]	ΔH_i
S1	2.5 (1.0)	$PLD_{S7}(W_1 = 14.49)$ (3)		-
S2	1.6 (1.3)		.55 [.50, .60]	− .02
S3	2.2 (1.1)		.56 [.50, .61]	− .03
S4	3.0 (1.1)	us: .35 [.28,.41] (2)		
S5	2.4 (1.2)		.66 [.63, .70]	00
S6	2.5 (1.3)		.67 [.64, .71]	− .02
S7	2.0 (1.1)		.64 [.60, .69]	− .03
S8	1.6 (1.3)		.65 [.60, .69]	.01
S9	1.4 (1.2)		.63 [.59, .68]	− .04
S40	1.3 (1.2)	us: .23 [.17, .30] (1)		
		H [95%CI]	.62 [.59, .66]	− .02

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (3) Item removal order

DIF – Differential item functioning; us – unscalable item; PLD_k – positive local dependence (subscripted item pair)

Table 16
Mokken Scaling Analysis (MSA) abbreviated results for the Ethics scale of the PPLA-Q; n = 473

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 280 / n_{male}= 193
		Removed items	Final H_i [95%CI]	ΔH_i
S12	2.4 (0.7)		.49 [.42, .56]	.09
S13	3.4 (0.7)		.50 [.43, .57]	− .02
S14	3.4 (0.7)		.52 [.46, .59]	.00
S15	3.3 (0.9)	us: .27 [.19, .35] (2)		
S16	3.2 (0.8)	IIO_{crit} = 71 (6)		
S17	3.4 (0.7)		.59 [.53, .64]	.05
S18	3.5 (0.7)	PLD_{S21} ($W_{1\ =}$ 8.00) (3)		
S19	2.9 (1.0)		.53 [.46, .59]	− .01
S20	3.2 (0.7)	PLD_{S12} ($W_{1\ =}$ 4.74) (4)		
S21	3.2 (0.8)		.62 [.57, .67]	.02
S22	2.7 (0.9)	PLD_{S19} ($W_{1\ =}$ 7.91) (5)		
S41	1.9 (1.1)	us: .21 [.14, .27] (1)		
		H [95%CI]	.54 [.49, .59]	.02

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (5) Item removal order

DIF – Differential item functioning; us – unscalable item; PLD_k – positive local dependence (subscripted item pair)

Table 17
Mokken Scaling Analysis (MSA) abbreviated results for the Collaboration scale of the PPLA-Q; n = 490

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 290 / n_{male} = 200
		Removed items	**Final H_i [95%CI]**	**ΔH_i**
S23	3.3 (0.7)		0.66 [0.60, 0.71]	.06
S24	3.2 (0.7)		0.59 [0.52, 0.65]	.04
S25	3.1 (0.7)	IIO_{crit} =82 (4)		
S26	3.5 (0.6)		0.69 [0.63, 0.75]	.15
S27	3.4 (0.6)		0.71 [0.67, 0.76]	.10
S28	3.1 (0.8)		0.62 [0.56, 0.67]	.08
S29	3.0 (0.9)	PLD_{S29} ($W_{1 =}$ 2.47) (2)		
S30	3.0 (0.8)		0.62 [0.57, 0.68]	.11
S31	3.1 (0.7)	PLD_{S28} ($W_{1 =}$ 3.01) and PLD_{S27} ($W_{1 =}$ 3.46) (3)		
S42	2.0 (1.1)	us: .18 [.10, .25] (1)		
		H [95%CI]	.64 [.60, .69]	.09

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (4) Item removal order

DIF – Differential item functioning; us – unscalable item; PLD_k – positive local dependence (subscripted item pair); IIO – invariant item ordering

Table 18
Mokken Scaling Analysis (MSA) abbreviated results for the Relationships scale of the PPLA-Q; n = 482

Label	Mean (SD)	Confirmatory MSA		DIF n_{female} = 283 / n_{male}= 199
		Removed items	**Final H_i [95%CI]**	**ΔH_i**
S32	3.2 (0.7)		.64 [.59, .70]	− .02
S33	3.1 (0.8)		.66 [.61, .71]	.04
S34	2.8 (0.9)		.55 [.48, .61]	.08
S35	2.7 (0.9)	PLD_{S34} ($W_{1 =}$ 5.23) and PLD_{S38} ($W_{1 =}$ 4.52) (3)		
S36	2.8 (0.9)		.61 [.55, .67]	.03
S37	2.9 (0.9)		.64 [.59, .69]	.03
S38	2.6 (0.9)		.58 [.52, .64]	− .01
S39	2.9 (1.0)	PLD_{S37} ($W_{1 =}$ 8.31) (2)		
S43	2.0 (1.1)	us: .31 [.23, .38] (1)		
		H [95%CI]	.61 [.57, .66]	.02

Note: all items showed no violations of monotonicity assumption (crit = 0)

(1) – (3) Item removal order

DIF – Differential item functioning; us – unscalable item; PLD_k – positive local dependence (subscripted item pair)

Emotional Regulation

Similarly, the clustering pattern observed in the Emotional Regulation scale in items P24, P26 and P27 led us to perform the same procedure. At c = .30, the results differed in both algorithms: while the AISP algorithm pointed to clustering of P24, P26 and P27, the GA algorithm kept the whole scale intact – in both cases P45 was unscalable. At c = .40, both algorithms return the same results, clustering these 3 items. There two clusters both form strong Mokken scales – Cluster 1 (H = .58, H^T = .19), Cluster 2 (H = .69, H^T = .08) – conforming with the DMM (Table 13). While Cluster 1 is equivalent to the final scale for this element, the Cluster 2 is formed by items assessing whether the respondent can recognize and identify emotions in others, during the practice of PA.

Physical Regulation

During confirmatory analysis of the *Physical Regulation* scale, 6 out of 8 items revealed a pattern of lowly scalable item (with H_i around .40); as such, we chose to further study its dimensionality. Until c = .40, both algorithms suggested a single cluster of items, beyond that point, clustering patterns differed among algorithms, with the AISP c = .45 solution approaching the a priori (*Foundation* difficulty) pattern of items - P40 as unscalable. For interpretability,

we chose this solution to perform the confirmatory procedure; both clusters formed strong Mokken scales, conforming to the DMM – Cluster 1 (H = .56, H^T = .20), Cluster 2 (H = .56, H^T = .30) – after removal of P34 for being flagged for PLD with P35 (Table 14).

Measurement invariance

To assess whether items presented differential item functioning (DIF) according to sex, we calculated the scalability coefficients – both item and total – for each final scale. Items P5 ("*I feel pressured by others to practice PA*"), S26 (" *I respect others*"), S27 ("*I cooperate with others*"), S30 ("*I help others achieve success*") presented a difference in item scalability (i.e., DIF) according to sex (H_i difference > .10; Table 11–17, column 4) – P5 with higher scalability for males, and S26,S27 and S30 with higher scalability for females – however all these were not statistically significant (p > .05; i.e., their 95% confidence interval overlap) and produced no appreciable effect on total scale scalability (H difference < .10; non-DTF).

The *Physical Regulation* scale showed slight to moderate differences in item scalability (DIF) in all its items (ranging from .12 to .23) with statistically significant differences in items P35 ("*I can recognize changes in my breathing*"), P39 ("*I use strategies to manage my effort*") and P41 ("*I can control my fatigue*"); resulting in a statistically significant difference in total scalability (H difference = .17; DTF) and borderline total scalability for females (H = 0.38 [0.32, 0.43]). To further investigate these differences, we calculated reliability coefficients for both subsamples (not shown in tables): female (ρ = .81, α = .79) and male (ρ = .89, α = .88). These results suggest that females and male might have interpreted differently concepts related with physical signs and fatigue during PA.

Reliability

Test-score reliability

Table 19, columns 4 and 5, sums up the reliability coefficients for the final scales. *ρ* estimates ranged from .83 to .94, above the recommended cut-off of .80. Similarly, *α* coefficients were very close to *rho*, and all above .80 as well (ranging from .83 to .91).

Test-retest reliability

ICC (Table 19, column, 6) revealed two different scenarios regarding the test-retest reliability of sum-scores obtained according to the final scales. The *Confidence*, *Emotional Regulation* and *Culture* scales showed moderate to excellent reliability across both time points ($ICC_{95\%CI}$ lower bound ranging from .63 to .85. and upper bound from .84 to .95) (Koo & Li, 2016); while the remaining scales showed a large spread in their confidence interval, ranging from inexistent (although ICC can be negative, they should be regarded as zero ; Matheson, 2019) to moderate/good reliability, indicative of great variability in scores between time points. Mean scores decreased from the first application to the second application in all but the *Culture* scale, while maintaining the same spread (Table 19, columns 7 and 8).

Table 19. Scalability, Invariant Ordering, and Reliability indexes for the Psychological and Social modules of PPLA

Subscale – number of items	Dimensionality		Reliability				
	Scalability H [95% CI]	Invariant Item Ordering H^T (item ordering)[1]	Test-score		Test-Retest 15-day interval N = 72		
			Molennar-Sijtsma ρ	Cronbach's α	$ICC_{2.1}$ [95% CI][2]	Mean Scores Baseline (SD)	Mean Scores Retest (SD)
Psychological							
Motivation – 7 items	.47 [.43, .51]	.33 (P5, P8, P2, P1, P7, P43, P11)	.83	.83	.70 [.10, .88]	19.8 (4.8)	17.1 (4.3)
Confidence – 9 items	.66 [.62, .69]	.08 (P13, P17, P44, P16, P19, P14, P21, P22, P20)	.94	.93	.91 [.85, .95]	22.2 (7.1)	21.2 (7.1)
Emotional Regulation - 7 items	.58 [.53, .62]	.19 (P25, P28, P30, P31, P23, P32, P29)	.90	.88	.76 [.63, .84]	17.5 (4.9)	16.6 (4.8)
Physical Regulation – 8 items	.46 [.41, .50]	.41 (P35, P36, P38, P42, P37, P33, P39, P41)	.84	.84	.43 [-.08, .72]	22.3 (4.7)	17.6 (4.4)
Social							
Culture – 7 items	.62 [.59, .66]	.32 (S6, S5, S3, S7, S2, S8, S9)	.91	.91	.88 [.82, .92]	14.0 (6.5)	14.3 (7.0)
Ethics – 6 items	.54 [.49, .59]	.52 (S13, S14, S17, S21, S22, S12)	.86	.85	.22 [-.05, .57]	18.7 (3.2)	12.1 (3.1)

Collaboration – 6 items	.64 [.60, .69]	.26 (S26, S27, S23, S24, S28, S30)	.88	.87	.23 [-.06, .58]	19.4 (3.3)	12.8 (3.2)
Relationships – 6 items	.61 [.57, .66]	.22 (S32, S33, S37, S36, S34, S38)	.88	.88	.48 [-.06, .75]	17.0 (3.9)	13.4 (3.9)

ICC – Intraclass Correlation Coefficient;

[1]Invariant Item Ordering method used was Manifest Item Invariant Ordering (Ligtvoet et al., 2011)

[2]Intraclass Correlation formula 2.1 – two-way mixed effects model accounting for single measurement (Koo & Li, 2016)

Discriminant and convergent validity

Point-estimate disatenuated correlations among scales-scores within the Psychological domain ranged from .31 to .83 (Table 20). Of these, *Emotional Regulation* was the lowest common correlate. *Motivation* and *Confidence* correlated above the .85 threshold (upper CI bound), showing higher correlation than warranted for two theoretically distinct scales. Point-estimate disatenuated correlations within the Social module ranged from .21 to .69 (Table 20). Of these, *Culture* was the lowest common correlate; with other scales correlating moderately to strongly. Correlations across domains were generally low, except for Culture and Relationships, which showed moderate disatenuated correlations with scale-scores in the Psychological domain.

Table 20
Bivariate Correlation (Spearman) Matrix

Scale	Psychological domain				Social domain			
	1.	2.	3.	4.	5.	6.	7.	8.
1. Motivation		.83 [.77, .87]	.31 [.22, .42]	.61 [.53, .70]	.54 [.46, .62]	.27 [.17, .37]	.26 [.15, .36]	.42 [.32, .51]
2. Confidence	.73 [.69, 77]		.47 [.37, .54]	.65 [.54, .69]	.50 [.43, .60]	.22 [.13, .31]	.22 [.12, .31]	.45 [.36, .53]
3. Emotional Regulation	.27 [.19, .36]	.43 [.35, .51]		.49 [.38, .54]	.18 [.08, .27]	.26 [.15, .37]	.19 [.09, .28]	.22 [.12, .33]
4. Physical Regulation	.51 [.44, .58]	.57 [.50, .64]	.43 [.35, .49]		.44 [.35, .52]	.42 [.31, .51]	.37 [.27, .48]	.46 [.37, .54]
5. Culture	.47 [.39 .54]	.46 [.38, .53]	.16 [.07, .25]	.39 [.31, 46]		.21 [.10, .29]	.23 [.14, .32]	.36 [.26, .45]
6. Ethics	.23 [14, .31]	.20 [.10, .29]	.23 [.14, .31]	.36 [.28, .44]	.18 [.10, .27]		.74 [.64, .77]	.52 [.42, .59]
7. Collaboration	.22 [.14, .31]	.20 [.11, .29]	.17 [.08, .26]	.32 [.23, .39]	.20 [.11, .29]	.64 [.58, .70]		.69 [.60, .73]
8. Relationships	.36 [.28, .44]	.41 [.32, .48]	.20 [.11, .28]	.40 [.32, .47]	.32 [.23, .40]	.45 [.37, .52]	.61 [.54, .67]	
Note: Raw bivariate correlation below diagonal, disattenuated correlations above diagonal								

Discussion

This study sought to establish evidence for construct validity and reliability of the psychological and social modules of the PPLA-Q in grade 10 to 12 (15–18 years) adolescents through investigation of their dimensionality, measurement invariance and reliability (total-score and test-retest).

Dimensionality

We used Mokken Scale Analysis (MSA) to gather evidence on the dimensionality of each of the eight scales composing the psychological and social modules of the PPLA-Q. Most local dependencies occurred within items initially designed for the same difficulty (i.e., foundation or mastery), and within the same specific trait (e.g., P9 and P11 with the same motivational regulation) with similar wording. This was expected since scale development ensured a desirable degree of redundancy (DeVellis, 2017).

All eight scales, after removal of offending items, adhered to the assumptions of the MHM (scalability, local independence, and monotonicity), with total scale scalability coefficients estimates (H) ranging from .46 to .62 – thus evaluated as moderate to strong scales. This values support the convergent validity (at item-level) of each scale (Sijtsma et al., 2011). Sum scores of items in these scales can, as such, be considered a sufficient indicator of the position in latent trait of each individual (Wind, 2017).

For all eight scales, the additional invariant item ordering (IIO) assumption held – assessed through the method of Manifest IIO (Ligtvoet et al., 2010) – as such, they adhered to the DMM. This evidence supports the interpretation that

an invariant order of items' difficulty can be established across different ranges of development, for all students, in the respective constructs (Wind, 2017), as warranted in the initial development of these scales. However, four of these scales had a H^T coefficient lower than .30 (*Confidence*, *Emotional Regulation*, *Collaboration* and *Relationships*), meaning that their IRF are too close together and that respondents might find difficult to distinguish between neighbor item, in difficulty terms (Sijtsma et al., 2011). Albeit still presenting an overall valid assessment of the position of a student (and items) on a continuum of difficulty, no specific use of this ordering (e.g., application of scales from an estimated difficulty point onward) is recommended for these four scales.

For the *Motivation* scale, items generally formed a difficulty continuum from controlled to more autonomous forms of motivation (Table 19) with weak accuracy (H^T =.33). Despite this, the continuum found does not entirely adhere to the posited order of the *Organism Integration* mini-theory of *Self-Determination Theory* regulations *(*Ryan & Deci, 2017): P8 ("*I feel good when I practice PA*"), developed to assess intrinsically regulated motivation was deemed easier (i.e., higher mean score) than P2–targeting externally regulated motivation at the diametrical side of the theoretical continuum. We argue that this might be due to the wording of P8 targeting a general well-being perception, which makes it easier to endorse that the more targeted expressions of intrinsic motivation like pleasure or satisfaction. As such, we recommend rewriting this item so that it more closely adheres to expected difficulty range. Similarly, P7 (developed to assess intrinsically regulated motivation, mean = 2.6) and P11 (integrated regulation, mean = 2.2) switched places, as the first is usually expected to be the most autonomous form of motivational regulation. This result agrees with previous results of bifactor modelling suggesting (Howard et al., 2016) that these two regulations might be closely placed in the continuum. To the intended application of the scale, however, this switch might have little consequence, as we generally discuss in the next paragraphs.

For the *Physical Regulation* scale, items formed a moderate accurate (H^T= .41) continuum from identifying physiological signs of effort and awareness of physical limits to using strategies to manage effort during PA, adhering to the *a priori* expectations. P42 ("*I take action to improve my physical skills*", mean = 2.9) wording might need to be adjusted in the future, as it appears to be interpreted as identical difficulty-wise as P37 ("*I recognize my physical limits*", mean = 2.8) – as evidenced by near-touching IRF – as both were to have different difficulties by design (i.e., P42 developmentally more complex than P37).

For the *Culture* scale, items formed a weakly accurate (H^T= .32) continuum from participation in the movement culture through use of specific PA terminology, to endorsing and encouraging others to so as well. Albeit designed to be among the easier items in this scale, S2 ("*I participate in PA rituals (e.g., greetings, hymns/chants, cheers, applauses*") figured, difficulty-wise, among the harder items in this scale; this might result from a misunderstanding regarding the concept of what "rituals" in a movement context truly mean, despite examples being provided in the item, as such, this item might merit further scrutiny in the future. Also, S6 ("*I like to keep up with PA events (e.g., competitions, spectacles, shows)*") wording might also be refined, to differentiate itself from S5 ("*I watch PA events (e.g., competitions, spectacles, shows)")* in terms of difficulty.

For the *Ethics* scale, items formed a strongly accurate (H^T= .51) continuum from immature forms (i.e., pragmatic) to mature forms (i.e., value-based) of moral development, adhering to the *a priori* development expectations based on Gibbs (2014)'s model.

Items developed to figure as global items (P1, P13, P23, P33, S1, S12, S23, S32) - to act as convergent validity indicators in future analysis (Cheah et al., 2018) – showed adequate scalability in all scales, strengthening the evidence for their convergent validity, as these were developed to generally represent each latent construct. Only in one of the scales (*Culture*) was one of these items (S1) flagged for local dependence – likely due to similar wording –

and removed. Difficulty-wise, in scales with interpretable IIO ($H^T > 30$), they figured in the middle to more difficult part of the difficulty continuum (i.e., lower mean score); this, again, was to be expected, as these were based on the operational definition of each element which state the development of each skill/construct in its final stages. Nonetheless, the usefulness of these items should be further examined (i.e., whether they are invaluable for scale scalability and validity), since their removal might result in a slight increase in feasibility in subsequent applications of this questionnaire, with no content representation trade-off.

Item developed to assess *Relational Thinking*, the highest development stage in the *Structure of Observed Learning Outcomes* taxonomy (Biggs & Collis, 1982) – items P43-P46, S40-S44 – did not fit the tested models, either for being unscalable or for being in local dependence pair; exception to this observation were the items in the *Motivation* and *Confidence* scales. These dealt with the degree to which skills developed in PA contexts are applied in context of the student's life. We argue that this might be due to: 1) endorsement of these items being highly dependent on the capacity of the respondent to draw a connection between his actual psychological and social skills in PA to their application in other contexts (e.g., being able to apply emotional regulation strategies developed or recurrently applied in PA contexts, to daily stressing occurrences), which by itself might be a different skill altogether – as is inferred from the most recent version of the Australian PL framework (Sport Australia, 2019); 2) the wording might not be clear enough to capture this phenomena among adolescents. As such, further efforts should be done to refine these items, and subsequently analyze their dimensionality – either as part of each of the scales, or as a separate latent trait by itself.

Additional exploration of the dimensionality – using the Automated Item Selection Procedures (AISP) and Genetic Algorithms (GA) at lower-bound c = .45) – of the *Motivation*, *Emotional Regulation* and *Physical Regulation* scales revealed an alternative cluster structure for these scales. Generally, at lower c values, these algorithms captured the higher-level constructs (i.e., unidimensional elements), while increasingly higher c values retrieved the lower-level constructs (i.e., foundation and mastery levels) and even specific subtraits within these (Straat et al., 2013). The clustering pattern of the *Motivation* scale throughout different lower-bound c values is coherent with previous research that posit that different motivational regulations differ not only in degree, but also in kind (Howard et al., 2020), with a general underlying continuum structure (Howard et al., 2016). Here, introjected regulation items were the exception, as they tended to cluster away from the remaining items at lower c values. These results suggest that this specific regulation might stand-out from all others, and along with the clustering of autonomous motivations is coherent with previous results on adolescents (Navarro et al., 2021; Vasconcellos et al., 2019).

For the *Emotional Regulation* scale, clustering patterns suggests that the identification of emotions in themselves and in others might be two different skills, although we initially equated them as part of the same continuum. Finally, the *Physical Regulation* results conform with the interpretation of a continuum underlying the development of all its skills, with two strong lower-level clusters of dimensionality within, coherent with the *a priori* construction of *Foundation* and *Mastery* levels. We argue that, although these alternative dimensionality structure could be supported for these proposed unidimensional scales, given that the aim of these scales is to be integrated within an overarching assessment framework for PL, instead of specific theory development on their construct, they possess enough dimensionality to be used to locate an individual in each of these latent traits, as evidenced by their total scalability coefficients.

Additionally, refinements to item's difficulty in scales with below standard, or borderline IIO accuracy ($H^T \approx .30$) are warranted, to better target different development stages across each construct. Parametric IRT models might support this effort, although its restrictive assumptions regarding Item Response Functions might not fit the functions we observed in this study.

For ease of interpretation and comparability between scales, we recommend that scores on this scale be transformed into a 0-100 metric using the maximum possible number of summed points as upper bound. Given that scales have mostly a balanced number of items designed to measure Foundational, and Mastery skills, a middle point score (50%) can be used as a heuristic cut-score to identify students transitioning into a deeper phase of learning.

Measurement Invariance

DIF and DTF analysis results suggest that all scales function similarly in male and female adolescents, except for the *Physical Regulation* scale which has shown evidence of a sex bias, despite obtaining borderline total scalability and acceptable reliability for females. This sex bias might stem from a different interpretation of these items by females. As such, we advise caution on the interpretation and comparison of between-sexes differences in resulting sum-scores in further studies. Since previous literature in this construct is sparse, further investigation and refinement of this construct and items is recommended through complementary statistic (e.g., Logistic Regression/ parametric Item Response Theory; Choi et al., 2011) and qualitative methodologies.

Reliability

All scales have shown evidence of adequate test-score reliability, further supporting the use of a sum-score. These estimates were, as expected, an improvement upon those obtained during the pilot phase (Mota et al., 2021), where 37% of scales failed to reach adequate reliability.

Intraclass Correlation Coefficient (ICC) results were mixed regarding adequacy of test-retest reliability, likely affected by true change in constructs measured. Given the timeframe of both applications, we argue that lockdown and school closure in Portugal might be one of the culprits, resulting in decreased scores in constructs that were otherwise expected to be stable in a 15-day interval. This decrease was especially evident in constructs related with social skills (*Ethics*, *Collaboration*, and *Relationships*), which were severely hampered during lockdown. Any use of these scales for the purpose of longitudinal studies should be further researched in a normal context to establish evidence for its test-retest reliability and enable valid detection of change across time (Marconcin et al., 2021), preferably through usage of IRT methods like growth models.

Discriminant and convergent validity

Disattenuated bivariate correlation suggested that the *Motivation* and *Confidence* scales might not show adequate discriminant validity (upper bound bordering on the usual .85 guideline; Brown, 2015), and thus might be measuring a very similar construct. Previous research has identified a moderate to strong correlation between very similar constructs (r = .64; Sweet et al., 2012), not differing much from the raw correlation we obtained. As such, these findings should tentatively bear on further studies, since disatenuated correlation are known to provide over inflated estimates (Murphy & Davidshofer, 2005). As further studies will integrate these scales as part of a larger structure, using both these scales as indicators of a common higher-order domain, along with other scales in the same domain - as generally posited by our model (Mota et al., 2021) – will allow for more robust interpretations. Similarly, refinement of items as previously suggested and further replications might evaluate this correlation. If this finding is replicated, we then suggest collapsing these scales to improve feasibility of the questionnaire.

Correlations among the Relationships scale-score and both Confidence and Motivation are coherent in magnitude with those observed in previous studies (Sweet et al., 2012), proving support for convergent validity of these scales, which measure constructs akin to Perceived Relatedness and Perceived Competence - core psychological needs of SDT (Ryan & Deci, 2017). Other correlations supported by similar results in literature include also that of Collaboration – Motivation (Li et al., 2008). These results, along with low to moderate correlations among constructs

in different domains provide support for convergent and discriminant validity of these scales. This assertion could be further supported by higher-order modelling in next phases of validation of PPLA.

Strengths and limitations

This study builds on the preliminary reliability evidence collected during pilot testing of the PPLA-Q (Mota et al., 2021) to refine the quality of the scales of its psychological and social modules. To do this, we used MSA, a non-parametric scaling technique which models these scales using a cumulative model, recognizing that items differ in their difficulty along a latent trait – providing an improvement over CTT models used in linear factor analysis (van Schuur, 2003). This conception closely aligns with the a priori specification of an underlying learning continuum with multiple stages. The resulting scales from this study can be feasibly applied in a PE context, given that their score can be derived via summing (i.e., sum-score), to provide an assessment of the students' position on each of these skills.

Despite the pandemic context imposed by COVID-19, we managed to recruit a diverse sample, closely 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, some caution should be used when generalizing findings using these scales, without further evidence of its adequacy in other contexts. Also, further test-retest reliability with a more diverse sample, under stabler circumstances and, preferably using IRT-based procedures should preclude any interpretation of longitudinal data based on these scales.

Also, despite being a useful and powerful method with increasing traction in instrument development, we also would like to acknowledge reports of the limited value of MSA for assessing dimensionality (Smits et al., 2012); as such, complementary methods for assessing dimensionality could be further employed in the future.

Conclusions

We have shown evidence in support of the dimensionality, convergent and discriminant validity, and test-score reliability of the eight scales of the psychological and social modules of the PPLA-Q, resultant of refinement through Mokken Scale Analysis; as such, sum of all final items in each scale (Additional File 1) can be used as an indicator of each latent construct. Further refinement to wording of items is warranted to increase the accuracy of the difficulty ordering within each scale, and discriminant validity of the *Motivation* and *Confidence* scales. We identified differential item and test functioning across sexes in one of the scales (*Physical Regulation*), which should be further scrutinized before any between-sexes comparisons are made on this construct, while all other scales have obtained evidence in support of their measurement invariance. These scales will be integrated into the PPLA framework to provide a feasible and holistic assessment of the individual journey of each grade 10–12 (15–18 years) student in Portuguese PE. Further test-retest reliability evidence should be collected before application in studies where the focus is in changes in construct scores 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 invaluable contribution of Filomena Araújo, António Rosado, António Rodrigues, Sofia Santos, Dean Dudley, and of all the tireless PE teachers and students which participated in these studies and kept the movement flame going during lockdown. We express a debt of gratitude to the entire R community for their selflessness and professionalism. The lead author would also like to thank his co-authors for their ever-present guidance and support during his PhD project.

References

1. Arifin, W. N. (2020). *Sample size calculator*. http://wnarifin.github.io
2. Ark, L. A. van der. (2012). New Developments in Mokken Scale Analysis in R. Journal of Statistical Software, *48*(1), 1–27. https://doi.org/10.18637/jss.v048.i05
3. Australian Government Department of Health. (2019). *Australian 24-Hour Movement Guidelines for Children (5–12 years) and Young People (13–17 years): An Integration of Physical Activity, Sedentary Behaviour, and Sleep*. Australian Government Department of Health.
4. Baptista, F., Santos, D. A., Silva, A. M., Mota, J., Santos, R., Vale, S., Ferreira, J. P., Raimundo, A. M., Moreira, H., & Sardinha, L. B. (2012). Prevalence of the Portuguese population attaining sufficient physical activity. Medicine

and Science in Sports and Exercise, *44*(3), 466–473. https://doi.org/10.1249/MSS.0b013e318230e441

5. Bernaards, C. A., & Sijtsma, K. (2000). Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable. Multivariate Behavioral Research, *35*(3), 321–364. https://doi.org/10.1207/S15327906MBR3503_03
6. Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of Observed Learning Outcomes)*. Academic Press.
7. 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
8. Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.
9. Cheah, J.-H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM: On using single-item versus multi-item measures in redundancy analyses. International Journal of Contemporary Hospitality Management, *30*(11), 3192–3210. https://doi.org/10.1108/IJCHM-10-2017-0649
10. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. Journal of Statistical Software, *39*(8), 1–30. https://doi.org/10.18637/jss.v039.i08
11. Corbin, C. B. (2016). Implications of Physical Literacy for Research and Practice: A Commentary. Research Quarterly for Exercise and Sport, *87*(1), 14–27. https://doi.org/10.1080/02701367.2016.1124722
12. Cortis, C., Puggina, A., Pesce, C., Aleksovska, K., Buck, C., Burns, C., Cardon, G., Carlin, A., Simon, C., Ciarapica, D., Condello, G., Coppinger, T., D'Haese, S., De Craemer, M., Di Blasio, A., Hansen, S., Iacoviello, L., Issartel, J., Izzicupo, P., … Boccia, S. (2017). Psychological determinants of physical activity across the life course: A "DEterminants of DIet and Physical ACtivity" (DEDIPAC) umbrella systematic literature review. PLOS ONE, *12*(8), e0182709. https://doi.org/10.1371/journal.pone.0182709
13. Craig, C. L., Marshall, A. L., Sjöström, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., Pratt, M., Ekelund, U., Yngve, A., Sallis, J. F., & Oja, P. (2003). International physical activity questionnaire: 12-country reliability and validity. Medicine and Science in Sports and Exercise, *35*(8), 1381–1395. https://doi.org/10.1249/01.MSS.0000078924.61453.FB
14. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, *16*(3), 297–334. https://doi.org/10.1007/BF02310555
15. DeVellis, R. (2017). *Scale Development: Theory and Applications* (4th ed.). SAGE Publications Ltd. https://us.sagepub.com/en-us/nam/scale-development/book246123
16. 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
17. Gamerman, D., Gonçalves, F. B., & Soares, T. M. (2019). Differential Item Functioning. In *Handbook of Item Response Theory Volume 3*. Chapman and Hall/CRC.
18. Gibbs, J. C. (2014). *Moral development and reality: Beyond the theories of Kohlberg, Hoffman, and Haidt* (Third edition). Oxford University Press.
19. Guthold, R., Stevens, G. A., Riley, L. M., & Bull, F. C. (2018). Worldwide trends in insufficient physical activity from 2001 to 2016: A pooled analysis of 358 population-based surveys with 1·9 million participants. The Lancet Global Health, *6*(10), e1077–e1086. https://doi.org/10.1016/S2214-109X(18)30357-7
20. Guthold, R., Stevens, G. A., Riley, L. M., & Bull, F. C. (2020). Global trends in insufficient physical activity among adolescents: A pooled analysis of 298 population-based surveys with 1·6 million participants. The Lancet Child

& Adolescent Health, *4*(1), 23–35. https://doi.org/10.1016/S2352-4642(19)30323-2

21. Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of Unidimensional Scales From a Multidimensional Item Bank in the Polytomous Mokken I RT Model. Applied Psychological Measurement, *19*(4), 337–352. https://doi.org/10.1177/014662169501900404
22. Hervé, M. (2021). *RVAideMemoire: Testing and Plotting Procedures for Biostatistics* (R package version 0.9–80) [Computer software].
23. Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (Vol. 663). Houghton Mifflin College Division.
24. Howard, J. L., Gagné, M., & Morin, A. J. S. (2020). Putting the pieces together: Reviewing the structural conceptualization of motivation within SDT. Motivation and Emotion, *44*(6), 846–861. https://doi.org/10.1007/s11031-020-09838-2
25. Howard, J. L., Gagné, M., Morin, A. J. S., & Forest, J. (2016). Using Bifactor Exploratory Structural Equation Modeling to Test for a Continuum Structure of Motivation. Journal of Management, *44*(7), 2638–2664. https://doi.org/10.1177/0149206316645653
26. 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
27. Krathwohl, Bloom, & Masia (Eds.). (1964). *Taxonomy of education objectives: The classification of education goals: Handbook 2—Affective domain*. David McKay.
28. Li, W., Wright, P. M., Rukavina, P. B., & Pickering, M. (2008). Measuring Students' Perceptions of Personal and Social Responsibility and the Relationship to Intrinsic Motivation in Urban Physical Education. Journal of Teaching in Physical Education, *27*(2), 167–178. https://doi.org/10.1123/jtpe.27.2.167
29. Ligtvoet, R., van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an Invariant Item Ordering for Polytomously Scored Items. Educational and Psychological Measurement, *70*(4), 578–595. https://doi.org/10.1177/0013164409355697
30. Ligtvoet, R., van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous Latent Scales for the Investigation of the Ordering of Items. Psychometrika, *76*(2), 200–216. https://doi.org/10.1007/s11336-010-9199-8
31. Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation – A discussion and demonstration of basic features. PLOS ONE, *14*(7), e0219854. https://doi.org/10.1371/journal.pone.0219854
32. LimeSurvey GmbH. (2021). *LimeSurvey: An Open Source survey tool*. LimeSurvey GmbH. http://www.limesurvey.org
33. Marconcin, P., Werneck, A., Peralta, M., Ihle, A., Gouveia, E., Ferrari, G., Sarmento, H., & Marques, A. (2021). *The Effects of Physical Activity on Mental Health During the COVID-19 Pandemic: A Systematic Review*. https://doi.org/10.21203/rs.3.rs-1026835/v1
34. Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. PeerJ, *7*, e6918. https://doi.org/10.7717/peerj.6918
35. Matos, M. G., & Equipa Aventura Social. (2018). *A Saúde dos Adolescentes Portugueses após a Recessão—Dados nacionais 2018*. Faculdade de Motricidade Humana. http://aventurasocial.com/arquivo/1437158618_RELATORIO%20HBSC%202014e.pdf
36. Ministério da Educação. (2001a). *Programa Nacional Educação Física: Ensino Secundário*. DES.

37. Ministério da Educação. (2001b). *Programa Nacional Educação Física (Reajustamento): Ensino Básico 3^{oCiclo}*. DEB.
38. Ministério da Educação [Ministry of Education]. (2019). *Infoescolas—Estatísticas do Ensino Básico e Secundário*. http://infoescolas.mec.pt/
39. Mokkink, L. B., de Vet, H. C. W., Prinsen, C. a. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, *27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4
40. Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. Kwantitatieve Methoden: Nieuwsbrief Voor Toegepaste Statistiek En Operationele Research, *9*(28), 115–126.
41. Moorer, P., Suurmeijer, Th. P. B. M., Foets, M., & Molenaar, I. W. (2001). Psychometric properties of the RAND-36 among three chronic disease (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. Quality of Life Research, *10*(7), 637–645. https://doi.org/10.1023/A:1013131617125
42. 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
43. Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed). Pearson/Prentice Hall.
44. Navarro, J., Escobar, P., Miragall, M., Cebolla, A., & Baños, R. M. (2021). Adolescent Motivation Toward Physical Exercise: The Role of Sex, Age, Enjoyment, and Anxiety. Psychological Reports, *124*(3), 1049–1069. https://doi.org/10.1177/0033294120922490
45. Nunnaly, J., & Bernstein, I. (1994). *Psychometric Theory*. McGraw-Hill.
46. Physical Literacy for Life. (2021). *What is Physical Literacy*. https://physical-literacy.isca.org/update/36/what-is-physical-literacy-infographic
47. 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
48. Pozo, P., Grao-Cruces, A., & Pérez-Ordás, R. (2018). Teaching personal and social responsibility model-based programmes in physical education: A systematic review. European Physical Education Review, *24*(1), 56–75. https://doi.org/10.1177/1356336X16664749
49. Price, L. R. (2017). *Psychometric Methods Theory into Practice*. The Guilford Press.
50. R Core Team. (2020). *R: A language and environment for statistical compution*. R Foundation for Statistical Computing. http://www.R-project.org/
51. Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. Annual Review of Clinical Psychology, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553
52. RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC. http://www.rstudio.com/
53. Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.
54. Sijtsma, K., & Ark, L. A. van der. (2021). *Measurement models for psychological attributes*. CRC Press.
55. Sijtsma, K., Meijer, R. R., & Andries van der Ark, L. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. Personality and Individual Differences, *50*(1), 31–37. https://doi.org/10.1016/j.paid.2010.08.016

56. Sijtsma, K., & Molenaar, I. (2002). *Introduction to Nonparametric Item Response Theory*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984676
57. Sijtsma, K., Straat, J. H., & van der Ark, L. A. (2015). Goodness-of-Fit Methods for Nonparametric IRT Models. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative Psychology Research* (Vol. 140, pp. 109–120). Springer International Publishing. https://doi.org/10.1007/978-3-319-19977-1_9
58. Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. British Journal of Mathematical and Statistical Psychology, *70*(1), 137–158. https://doi.org/10.1111/bmsp.12078
59. Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken Scale Analysis as a Dimensionality Assessment Tool: Why Scalability Does Not Imply Unidimensionality. Applied Psychological Measurement, *36*(6), 516–539. https://doi.org/10.1177/0146621612451050
60. Sport Australia. (2019). *Australian Physical Literacy Framework.* https://nla.gov.au/nla.obj-2341259417
61. Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. BMC Medical Research Methodology, *12*(1), 74. https://doi.org/10.1186/1471-2288-12-74
62. Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Methodological artifacts in dimensionality assessment of the hospital anxiety and depression scale (HADS). Journal of Psychosomatic Research, *74*(2), 116–121. https://doi.org/10.1016/j.jpsychores.2012.11.012
63. Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum Sample Size Requirements for Mokken Scale Analysis. Educational and Psychological Measurement, *74*(5), 809–822. https://doi.org/10.1177/0013164414529793
64. Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2016). Using Conditional Association to Identify Locally Independent Item Sets. Methodology, *12*(4), 117–123. https://doi.org/10.1027/1614-2241/a000115
65. Sweet, S. N., Fortier, M. S., Strachan, S. M., & Blanchard, C. M. (2012). Testing and integrating self-determination theory and self-efficacy theory in a physical activity context. Canadian Psychology/Psychologie Canadienne, *53*(4), 319–327. https://doi.org/10.1037/a0030280
66. Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychology Science Quarterly, *50*(4), 538.
67. UNESCO. (2015). *Quality Physical Education (QPE): Guidelines for policy makers*. UNESCO Publishing.
68. van Schuur, W. H. (2003). Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. Political Analysis, *11*(2), 139–163. https://doi.org/10.1093/pan/mpg002
69. Vaquero-Diego, M., Torrijos-Fincias, P., & Rodriguez-Conde, M. J. (2020). Relation between perceived emotional intelligence and social factors in the educational context of Brazilian adolescents. Psicologia: Reflexão e Crítica, *33*(1), 1. https://doi.org/10.1186/s41155-019-0139-y
70. Vasconcellos, D., Parker, P., Hilland, T., Cinelli, R., Owen, K., Kapsal, N., Lee, J., Antczak, D., Ntoumanis, N., Ryan, R., & Lonsdale, C. (2019). Self-determination theory applied to physical education: A systematic review and meta-analysis. Journal of Educational Psychology, *112*(7), 1444–1469. https://doi.org/10.1037/edu0000420
71. Wind, S. A. (2017). An Instructional Module on Mokken Scale Analysis. Educational Measurement: Issues and Practice, *36*(2), 50–66. https://doi.org/10.1111/emip.12153
72. World Health Organization. (2020). *WHO guidelines on physical activity and sedentary behaviour*. World Health Organization. https://apps.who.int/iris/handle/10665/336656

73. Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in Questionnaire Data: Can They Be Detected and Should They Be Removed? Journal of Educational and Behavioral Statistics, *36*(2), 186–212. https://doi.org/10.3102/1076998610366263

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.pdf