Universidade de Lisboa

Faculdade de Motricidade Humana

Portuguese Physical Literacy Assessment (PPLA): Development and Validation of an Instrument for Adolescents in Physical Education

João Mota Rodrigues

Supervisor: Doctor Marcos Teixeira de Abreu Soares Onofre
Co-supervisor: Doctor João Filipe da Silva Figueira Martins

Thesis written for the degree of Doctor in Education in the Specialty of Didactics of Physical Education and Sports

Thesis by compilation of articles, according to point a) of nº2 of article 31º from *Decree-Law* number 230/2009

2022

Universidade de Lisboa

Faculdade de Motricidade Humana

Portuguese Physical Literacy Assessment (PPLA): Development and Validation of an Instrument for Adolescents in Physical Education

João Mota Rodrigues

Supervisor: Doctor Marcos Teixeira de Abreu Soares Onofre
Co-supervisor: Doctor João Filipe da Silva Figueira Martins

Thesis written for the degree of Doctor in Education in the Specialty of Didactics of Physical Education and Sports

Jury:

President:

- Doutor Francisco José Bessone Ferreira Alves
  Presidente do Conselho Científico
  Faculdade de Motricidade Humana da Universidade de Lisboa

Members:

- Doutor António Fernando Boleto Rosado
  Professor Catedrático
  Faculdade de Motricidade Humana da Universidade de Lisboa
- Doctor Dean Alan Dudley
  Associate Professor
  Macquarie School of Education, Macquarie University (Sydney, Australia)
- Doutor Marcos Teixeira de Abreu Soares Onofre
  Professor Associado
  Faculdade de Motricidade Humana da Universidade de Lisboa
- Doutor Cláudio Filipe Guerreiro Farias
  Professor Auxiliar
  Faculdade de Desporto da Universidade do Porto
- Doutor António José Mendes Rodrigues
  Professor Auxiliar
  Faculdade de Motricidade Humana da Universidade de Lisboa

Funding by University of Lisbon
2022

To the giants on whose shoulders I stand.

Funding

This research work was funded by a PhD Scholarship from the University of Lisbon PhD Scholarship Program 2017.

Portuguese Physical Literacy Assessment (PPLA): Development and Validation of an Instrument for Adolescents in Physical Education

Abstract

The main purpose of this PhD thesis was to develop and validate a novel criterion-referenced Physical Literacy (PL) assessment system for application in Portuguese PE for grade 10-12 adolescents (15-18 years): the Portuguese Physical Literacy Assessment (PPLA). Inspired by the Australian Physical Literacy Framework (APLF), this tool is comprised of two instruments assessing the physical, cognitive, psychological, and social domains of PL: 1) PPLA-Questionnaire (PPLA-Q) and 2) the PPLA-Observation tool (PPLA-O).

The first is a self-administered questionnaire with three modules, each respectively designed to assess the psychological, social, and part of the cognitive domains of PL; while the latter is an instrument with two modules that uses teacher-reported data to assess the physical and the remainder of the cognitive domain of PL. PPLA development and validation process is presented through five scientific papers: the first and fourth present the development of both instruments, marrying quantitative and qualitative methods; while the second, third and fourth establish evidence for the content and construct validity (dimensionality, measurement invariance across sex, and convergent and discriminant validity), as well as reliability (score and test-retest) at element-level, within each of the four domains assessed. Finally, the fifth articles focus on the integration of the full PPLA measurement model with all domains and elements, establishing its construct validity and reliability.

Overall, the PPLA emerges as a highly feasible tool for the PE context that can be completed in around 20 minutes (students filling in the PPLA-Q) plus time spent by PE teachers in data insertion/copying into the PPLA-O spreadsheet. Its measurement model is best represented through an asymmetrical bifactor model, allowing for disentangling the variance associated with a general PL trait - referent to a transversal broadband meta-learning or disposition in movement settings – from variance of specific group factors (domains).

PPLA can be used to provide a detailed and feasible assessment of each student's PL journey, and to support pedagogical decisions (at local, regional, and national level) towards a more meaningful and targeted PE environment to promote PL learning. Further research is warranted in replicating these findings outside an imposition-laden COVID-19 setting, along with multiple fine-tuning to the PPLA. Similarly, adaptation of this tool to other age-ranges and its use as an aid in monitoring and advocating for PL inside a quality PE setting are open threads for future work.

Keywords: physical literacy, assessment, physical education, development, validation, high-school, adolescence.

Avaliação Portuguesa da Literacia Física (PPLA): Desenvolvimento e Validação de um Instrumento para Adolescentes em Educação Física

Resumo

O principal objetivo desta tese de doutoramento foi desenvolver e validar um novo sistema de avaliação de Literacia Física (LF), baseado num referencial criterial, para aplicação em adolescentes portugueses do 10º ao 12º ano de escolaridade durante as aulas de EF: *Portuguese Physical Literacy Assessment* (PPLA). Inspirando-se no modelo australiano de LF (APLF), este sistema é composto por dois instrumentos que avaliam os domínios físico, cognitivo, psicológico e social da LF: 1) PPLA-Questionário (PPLA-Q) e 2) PPLA-Observação (PPLA-O).

O primeiro é um questionário autoadministrado com três módulos, cada um concebido, respetivamente, para avaliar os domínios psicológico, social, e parte do domínio cognitivo da LF; já o último é um instrumento com dois módulos que utiliza dados reportados pelos professores de EF para avaliar o domínio físico e o resto do domínio cognitivo da LF. O seu processo de desenvolvimento e validação é apresentado através de cinco artigos científicos: o primeiro e quarto apresentam o desenvolvimento de ambos os instrumentos, através de métodos qualitativos e quantitativos; já o segundo, terceiro e quarto artigo estabelecem evidência para a validade de conteúdo e construto (dimensionalidade, invariância de medição em diferentes sexos, e validade convergente e discriminante) e fiabilidade (*score* e teste-reteste) do PPLA ao nível micro (elementos da LF). Finalmente, o quinto artigo foca-se na integração do modelo completo de medição do PPLA com todos os seus domínios e elementos, estabelecendo evidência da sua validade de construto e fiabilidade.

Globalmente, o PPLA emerge como uma ferramenta altamente viável para aplicação em contexto de EF que pode ser completada em cerca de 20 minutos (preenchimento do PPLA-Q pelos alunos) acrescidos do tempo de inserção de dados, por parte do professor, na folha de cálculo do PPLA-O. O seu modelo de medição é melhor representado através de um modelo bifatorial assimétrico, que permite separar a variância associada a uma competência geral de LF - referente a uma metaaprendizagem ou disposição transversal às diversas competências e contextos de movimento - da variância específica associada a cada domínio; permitindo também a análise independente do efeito ou relação de cada domínio com variáveis de interesse futuro. O PPLA pode ser utilizado para facultar uma avaliação detalhada e acessível do percurso efetuado por cada estudante na sua LF, e para apoiar decisões pedagógicas (a nível local, regional e nacional) com fim a tornar a EF um ambiente mais orientado para promover a aprendizagem deste conjunto de competências. De futuro, é necessária mais investigação no sentido de replicar os resultados obtidos nestes estudos, fora de um cenário fortemente limitado pela pandemia de COVID-19; e de efetuar múltiplas afinações ao PPLA. Do mesmo modo, são tópicos abertos para trabalho futuro a adaptação desta ferramenta a outras faixas etárias e a sua utilização como suporte na monitorização e disseminação da LF como um dos focos de desenvolvimento de EF de qualidade.

Palavras-chave: literacia física, avaliação, educação física, desenvolvimento, validação, ensino secundário, adolescência.

Table of Contents

Tables and Figures

Tables List

Figures

Abbreviations

PL – Physical Literacy

PA – Physical Activity

PE – Physical Education

APLF – Australian Physical Literacy Framework

PPES – Portuguese Physical Education Syllabus

PPLA – Portuguese Physical Literacy Assessment

FA – Factor Analysis

CTT – Classical Test Theory

IRT – Item Response Theory

Symbols

θ – Latent trait

ρ_{xx} - Reliability

κ – Agreement coefficient

ω / ω_H – Omega coefficient / omega hierarchical coefficient (congeneric reliability)

α – Alpha coefficient (tau-equivalent reliability)

β – standardized regression weight

λ – standardized factor loading

Δ – difference

χ^2 – chi-square statistic

Introduction

Physical Literacy (PL) is a concept based on lifelong holistic learning acquired and applied in movement and physical activity (PA) contexts (Sport Australia, 2019). Arguably, the most seminal contribution to the development of the concept in modern pedagogy have been the works of Margaret Whitehead (2001, 2007, 2010), which conceptualized PL as the motivation, confidence, physical competence, understanding and knowledge to maintain physical activity throughout the life course; a definition that was later adopted by the International Physical Literacy Association (2017) and by the Canada's PL Consensus Statement (Tremblay et al., 2018).

Notwithstanding its lifelong development, sowing the seeds of PL during school-age seems critical, as participation in early childhood might predict adherence to active lifestyles throughout life (Telama, 2009; Telama et al., 2014), counteracting the high levels of physical inactivity observed in adolescents and adults (Guthold et al., 2018, 2020). In this line, PL development is implicit in the World Health Organization updated guidelines for PA (2020), and explicitly argued as the main outcome of quality physical education (PE) (UNESCO, 2015). The latter is a privileged environment – mandatory, free and qualified – for learning the life skills and values needed for active and global citizenship (Onofre, 2017), with a relevant effect on PA participation of adolescents (Uddin et al., 2020). Thus, many authors have underlined the need to operationalize this concept in school curricula and educational policies (Corbin, 2016a; Dudley, 2015; Dudley, Cairney, et al., 2017)

Despite a general consensus on the ultimate goal of PL – sustained lifelong PA participation (Whitehead, 2013a, 2013b) –, its proposed conceptualization and constituent elements differ across sources (L. Edwards, Bryant, Keegan, Morgan, & Jones, 2017; Liu & Chen, 2021; Martins et al., 2020). These range from philosophically-driven conceptualizations, like Whitehead's PL original proposition (Whitehead, 2001) – rooted in the philosophical tenets of monism, phenomenology, and existentialism – to diametrical conceptualizations focusing solely on one of its aspects (e.g., fundamental movement skills) (D. B. Robinson et al., 2018). Although recognized as a rich theoretical concept, the former might lack pragmaticism to be implemented in PE practice (L. Edwards, Bryant, Keegan, Morgan, & Jones, 2017; Longmuir & Tremblay, 2016): while the later might deviate from the holistic nature of PL, compromising crucial elements like pleasure and enjoyment in taking part in PA (Pot et al., 2018).

This tension also translates into PL assessment, a crucial element in its implementation and practice (Corbin, 2016a). As conceived by Whitehead, PL favors an ipsative-referenced frame (Standal, 2016), based on an idealist/interpretative paradigm (L. Edwards, Bryant, Keegan, Morgan, Cooper, et al., 2017), which presents a diametrical view to normative references that underlie most standardized testing settings. In both arenas – conceptual and measurement - a middle-ground compromise in the form of adequate criterion-referenced assessment might offer a tenable and comparable solution across different contexts: using clear and measurable outcomes, while honoring most of the philosophical-driven premises that define the concept (D. B. Robinson et al., 2018; Young et al., 2019).

To this end, a team of Australia-based researchers developed the *Australian Physical Literacy Framework* (APLF) (Keegan et al., 2019; Sport Australia, 2019), an evidence -based, integrated model of PL in the *physical*, *cognitive*, *psychological* and *social* domains with 30 different elements. This model was designed with implementation in mind – be that by practitioners, policy-makers, and researchers – and is novel in that it explicitly acknowledges the contribute that PL might play in cultural and social participation. It also provides a clear focus on a learning continuum, inspired by the *Structure of Observed Learning Outcomes* taxonomy (Biggs & Collis, 1982), designed to include individuals in different states of their PL journey: from their first steps (*pre-foundational*) to higher stages of proficiency (*transfer & empowerment*) (Keegan et al., 2017, 2019).

A few assessment instruments have been developed, under diverse conceptual models (L. Edwards, Bryant, Keegan, Morgan, Cooper, et al., 2017; Liu & Chen, 2021; Shearer et al., 2021). Of these, the most prolific research-wise have been the Canadian Assessment of Physical Literacy (Francis et al., 2016; Gunnell et al., 2018), and the Physical Literacy Assessment for Youth (Cairney et al., 2018). Both tools integrate observational procedures and self-report, and feature overall good feasibility in PE (Shearer et al., 2021) but lack options for older adolescents (15-18 years), a critical age range in Portugal in which adolescents tend to present lower levels of PA (Baptista et al., 2012; Martins, Marques, et al., 2019; Matos & Equipa Aventura Social, 2018). This age group is therefore a priority target in the Portuguese PE setting for PL development, contributing to lifelong and meaningful engagement in physical activity.

Research goals and thesis structure

The main purpose of this PhD thesis was to develop and validate a novel criterion-referenced PL assessment system for application in Portuguese PE for grade 10-12 (15-18 years) adolescents: the Portuguese Physical Literacy Assessment (PPLA). This tool is comprised of two instruments (Figure 1): PPLA-Questionnaire (PPLA-Q) and the PPLA-Observation tool (PPLA-O). The first is a self-administered questionnaire with three modules, each respectively designed to assess the psychological, social, and part of the cognitive domain of PL; while the latter is an instrument with two modules that uses teacher-reported data to assess the physical and the remainder of the cognitive domain of PL. Both instruments make use of the APLF model's domains, elements and learning continuum conceptualization to provide a detailed and feasible assessment of each student's PL journey, that may support pedagogical decisions (at local, regional, and national level) towards a more meaningful and targeted PE environment to promote PL learning. This process is presented through five scientific papers[1]: one published, and four in the final stages of preparation for submission. The choice to compose this thesis of scientific papers written in English, affords a unique opportunity to contact early on with the peer-reviewing, and publishing process - essential to the current *ethos* of academia; and to contribute to the international dialogue about PL measurement in a timely and accessible manner.

[1] ***Other scientific work during the doctorate (not part of the thesis):***

Published Papers:

Martins, J., Onofre, M., **Mota, J.**, Murphy, C., Repond, R.-M., Vost, H., Cremosini, B., Svrdlim, A., Markovic, M., & Dudley, D. (2020). International approaches to the definition, philosophical tenets, and core elements of physical literacy: A scoping review. *PROSPECTS*. https://doi.org/10.1007/s11125-020-09466-1 [Impact Factor: 0.67]

Peer-reviewed oral communications:

Mota, J., Martins, J., & Onofre, M. (2021). Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q): Development, content validation and pilot testing. *Book of Abstracts for AIESEP 2021*, 311.

Martins, J., Onofre, M., **Mota, J.**, Murphy, C., Repond, R.-M., Vost, H., Cremonesi, B., Svrdlin, A., & Markovic, M. (2019). A review of different international approaches to the definition and core elements of physical literacy. *Book of Abstracts for AIESEP 2019*, 496.

Prior to presentation of the main body of research, a general overview of the research design choices, and relevant measurement theories and models will be given in Chapter 1. This overview presents a description of conceptual and methodological frameworks that are common across all following chapters, which would be unwarranted in published work due to its extension. Chapter 1 also provides a summary of the main methods used throughout the project. All other PL-related content is reviewed extensively inside Chapter 2 to 6 (paper chapters). After the presentation of the main research, a final synthesis of the results, as well as conclusions, limitations, and future perspectives will be addressed in Chapter 7. In the interest of parsimony, a single references list is provided at the end of this thesis, along with a single repository of Additional Files.

The main research work was developed through multiple phases of development and validation, mapped in Figure 1 to the Chapters of this thesis to frame the reader's understanding and expectations. Validation entailed collecting evidence from multiple sources on validity and reliability of the intended instruments (American Educational Research Association et al., 2014; Hubley & Zumbo, 2011; Mokkink et al., 2018; Prinsen et al., 2018) and overarching PL model. This was done in a bottom-up approach to ensure minimal propagation of error: element-level validation led to domain-level, and PL-level (Figure 1; right panel).

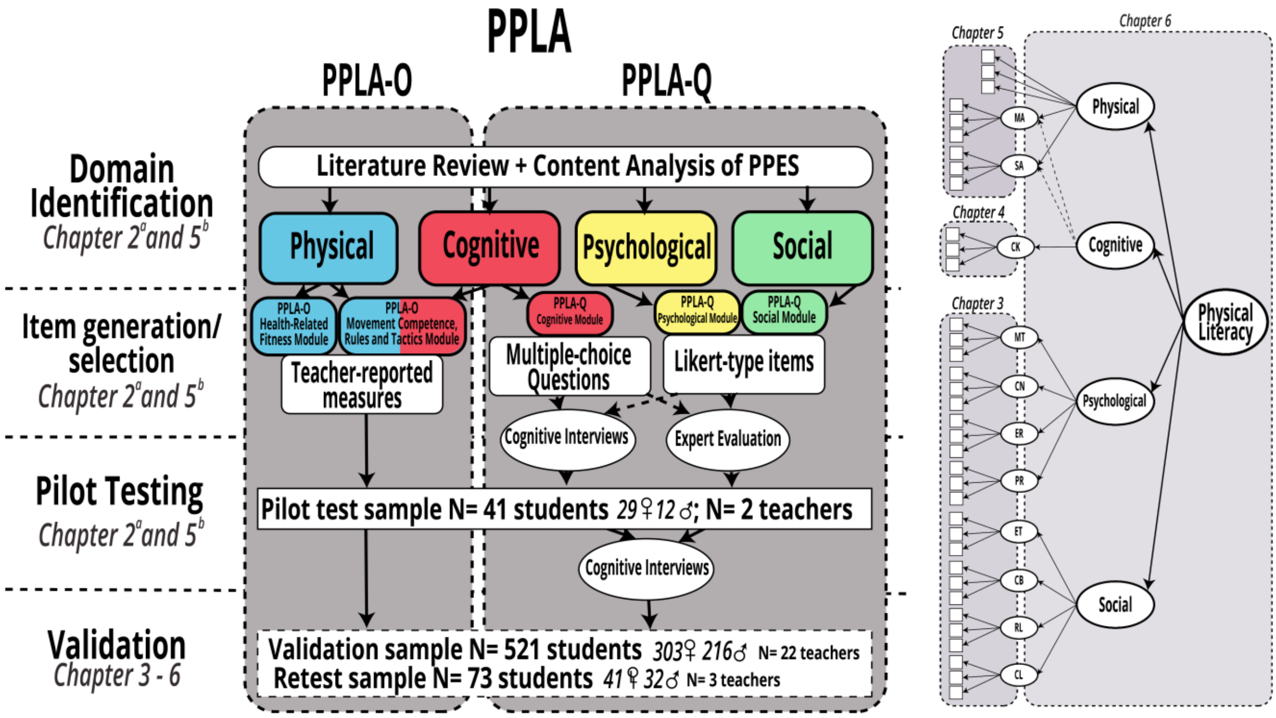

PPLA
PPLA-O
PPLA-Q
Domain Identification
Chapter 2a and 5b
Item generation/ selection
Chapter 2a and 5b
Pilot Testing
Chapter 2a and 5b
Validation
Chapter 3 - 6
Literature Review + Content Analysis of PPES
Physical
Cognitive
Psychological
Social
PPLA-O Health-Related Fitness Module
PPLA-O Movement Competence, Rules and Tactics Module
PPLA-Q Cognitive Module
PPLA-Q Psychological Module
PPLA-Q Social Module
Teacher-reported measures
Multiple-choice Questions
Likert-type items
Cognitive Interviews
Expert Evaluation
Pilot test sample N= 41 students 29♀12♂; N= 2 teachers
Cognitive Interviews
Validation sample N= 521 students 303♀ 216♂ N= 22 teachers
Retest sample N= 73 students 41♀ 32♂ N= 3 teachers
Chapter 5
Chapter 6
Chapter 4
Chapter 3
MA
SA
CK
MT
CN
ER
PR
ET
CB
RL
CL
Physical
Cognitive
Psychological
Social
Physical Literacy

Figure 1. Left panel: Thesis map of Portuguese Physical Literacy Assessment (PPLA) development and validation phases; a - PPLA-Questionnaire (PPLA-Q); b – PPLA-Observation (PPLA-O). Right panel: Thesis map of the Portuguese Physical Literacy Assessment (PPLA) validation[2]

Legend: MA – Manipulative-based Activities; SA – Stability-based Activities; CK – Content Knowledge; MT-Motivation; CN-Confidence; ER – Emotional Regulation; PR – Physical Regulation; ET - Ethics ; CB - Collaboration ; RL - Relationships ; CL – Culture.

Paper 1 (Chapter 2)

Full title: Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) for adolescents (15-18 years) from grades 10-12: development, content validation and pilot testing

Short title: PPLA-Q Development, Content Validity and Preliminary Construct Validity

Paper 1 will present the development, content validation and preliminary construct validity and reliability for the PPLA-Questionnaire (PPLA-Q), one of two instruments of the PPLA system. Here will also be detailed the rationale and design choices common to both tools. Much of what was presented in the preceding

[2] *Note: although the final PPLA model corresponds to a bifactor representation, a hierarchical second-order model is presented here for simplicity of presentation.*

introduction will be echoed through this, and other chapters, due to their publication as mostly self-contained pieces of literature. This paper was adapted from its published version in form only (i.e., citation style, common references, and additional files) to coherently integrate with the remaining chapters, with no alterations to its content.

Paper 2 (Chapter 3)

Full title: 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

Short title: PPLA-Q Psychological and Social Modules Construct Validity and Reliability

Paper 2 will detail the assessment of multiple aspects of construct validity and reliability of the Psychological and Social modules of the PPLA-Q through Non-Parametric Item Response Theory models (Mokken Scale Analysis), namely their dimensionality, measurement invariance (across sexes), convergent and discriminant validity, score reliability, and test-retest reliability. As will be common across the following papers, scoring implications and suggested revisions to these modules are discussed.

Paper 3 (Chapter 4)

Full title: 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

Short title: PPLA-Q Cognitive Module Construct Validity and Reliability

Paper 3 details the assessment of construct validity and reliability of the Cognitive module of the PPLA-Q: a) internal structure/dimensionality, measurement invariance (across sexes), reliability (score and test-retest). These were based on Parametric Item Response Theory (PIRT) models, which permitted a deeper analysis of item's behavior and scoring implications at different ranges of knowledge (latent continuum).

Paper 4 (Chapter 5)

Full title: Portuguese Physical Literacy Assessment Observation (PPLA-O) for adolescents (15-18 years) from grades 10-12: development and validation through Item Response Theory

Short title: PPLA-O Development, Construct Validity and Reliability

Paper 4 introduces the development of the PPLA-Observation (PPLA-O), the second instrument of the PPLA, using content analysis of the Portuguese PE syllabus and literature review. It also details the assessment of construct validity (dimensionality, measurement invariance, convergent and discriminant validity) and reliability (score) of one of its modules (Movement Competence, Rules and Tactics) through multidimensional PIRT models.

Paper 5 (Chapter 6)

Full title: Portuguese Physical Literacy Assessment for adolescents (15-18 years) from grades 10-12: validation using Confirmatory Factor Analysis and Confirmatory Composite Analysis

Short title: Full PPLA Construct Validity and Reliability

Paper 5 analyzes the construct validity and reliability of the full PPLA model, integrating variables from both instruments (PPLA-Q and PPLA-O) through two different measurement paradigms (reflective and composite-formative), culminating in a pragmatic choice for an asymmetrical bifactor measurement model. It also provides a literature review of conceptual issue germane to ontology of measurement models, and implications for PL assessment of both these paradigms are discussed, along with multiple scoring recommendations.

CHAPTER 1 - General Overview of Research Paradigm and Methods

Research paradigm

Underlying any research effort are specific paradigms composed of ontological and epistemic assumptions regarding the phenomenon under study that will influence the choice of study design and methodologies used (Creswell & Creswell, 2018).

As hinted during the introductory note, the Whiteheadian school of thought approaches the conceptualization of PL through an essentially constructivist /interpretative worldview, assuming epistemic phenomenology and existentialism (i.e., positing that each individual experiences and construes the world in an inherently unique perspective, influenced by previous experiences and characteristics; Pot et al., 2018; Whitehead, 2010), as well as an ontological monism (i.e., reality is a whole without independent parts; embodied experience must be understood as a whole). As such, through this frame of reference, PL would be better studied using qualitative approaches and ideographic methodologies (Burrell & Morgan, 2019), which focus on understanding the multiple realities than are spanned within the individual sphere in an holistic interaction between all attributes/domains of PL, refuting any attempt at assessing this construct using standardized procedures (L. Edwards, Bryant, Keegan, Morgan, Cooper, et al., 2017; L. Edwards, Bryant, Keegan, Morgan, & Jones, 2017).

Diametrically, a strictly postpositivist worldview usually underlies the consideration and assessment of PL attributes in a reductionist fashion (e.g., movement competence/fundamental movement skills, physical fitness), entailing the assumption of ontological realism and epistemic positivism (i.e., reality is tangible through usage of objective measures; Creswell & Creswell, 2018). Through this lens of analysis, PL would be better studied through quantitative, nomothetic methodologies (Burrell & Morgan, 2019), focusing on classical standards of validity and reliability – concerned solely with the group-level relationships, or criteria external to the individuals being assessed (Hubley & Zumbo, 2011; Markus & Borsboom, 2013).

Sport Australia's view on PL takes a middle ground positioning, assuming a pragmatic paradigm. Through this, a problem-centered approach is taken, recognizing that the ontological and epistemic considerations are important, but focus is on applying the concept in research, and educational practice (Keegan et al.,

2019) through an integrated operationalization and understanding of all dimensions that underlie PL; recognizing that assessment must consider both measurement validity standards and contextual variables (Barnett et al., 2019). As such, mixed methodologies that integrate a triangulation of data sources (i.e., quantitative, data-driven inferences; and qualitative, meaning-driven inferences) might be more adequate for this positioning (Creswell & Creswell, 2018). Similarly, mixed-methods designs align with the most updated vision on validity as centered on the premise of marrying quantitative with qualitative methods to analyze the elicited response process, and its meaning to the individuals being assessed (American Educational Research Association et al., 2014; Chan, 2014).

Thus, a pragmatic paradigm was used throughout this thesis, using a convergent mixed methods approach (Creswell & Creswell, 2018) in which both qualitative methods (e.g., cognitive interviews, content analysis, qualitative expert evaluation) and quantitative methods (e.g., quantitative expert evaluation, Structural Equation Modelling [SEM]) were used to produce a cabal understanding of the problem of the assessment of PL in Portuguese PE setting. A systemic approach supported the domain identification, measure design process and validation, recognizing that although PL is a single complex and integrated conceptual framework, development in its multiple attributes is best tackled through identification and learning on key critical areas that will, according to their specificity, require different learning strategies on the part of the PE teachers, and of the student. Thus, throughout the following chapters, the term *holistic* will take the meaning of a systemic, integrated view, rather than a truly canonical view of holism. Similarly, throughout the research described in Chapter 2-6, methods were chosen based on their contextual usefulness, in constant dialogue between deduction (previous theory) and induction (understanding of the data at hand, and its originating conditions), towards this thesis' main goal, as will be exemplified in the next section.

Measurement

Similarly to the positions expressed above, decisions regarding the ontology of PL and paradigm choice are central to the definition of the measurement model and methods used to validate it (J. R. Edwards & Bagozzi, 2000; Henseler, 2021; Sarstedt et al., 2016): While a vision closer to anti-realism espoused by the *Whiteheadian* base of phenomenology (Whitehead, 2010) would lend itself to assuming PL as a social and personal construction – an artifact/emergent variable without ascribed existence or common cause, with the different attributes or elements forming/defining PL, as composite indicators (e.g., Henseler, 2021); a vision centered on canonical measure development and realism would point itself to a latent variable interpretation, using attributes/element as interchangeable reflective indicators of PL (Figure 2). There are, however, other practical implications that are germane to this choice (e.g., scoring of results obtained by the instrument). As such, in Chapter 2 an initial hypothesized measurement model will be put forth, based on the idea that at macro-level PL could be seen as an emergent variable composed of four non-interchangeable correlating domains (also emergent variables), which are finally composed of elements. Most of these elements were initially hypothesized as being an emergent variable due to methodological limitations of the most used Classical Test Theory-based analysis (factor analysis [FA]). As the work progressed, it became clear that other Item Response Theory-based methodologies could be used, and these elements considered as latent variables (i.e., unidimensional constructs) – this will be addressed in Chapter 3, when the dimensionality of the Psychological and Social elements are studied. Finally, in chapter 6 the hypothesis that PL could be considered an emergent variable is tested, and both practical and conceptual implications discussed through a pragmatic lens – there will be also place to a literature review that deepens the relationships presented in Figure 2.

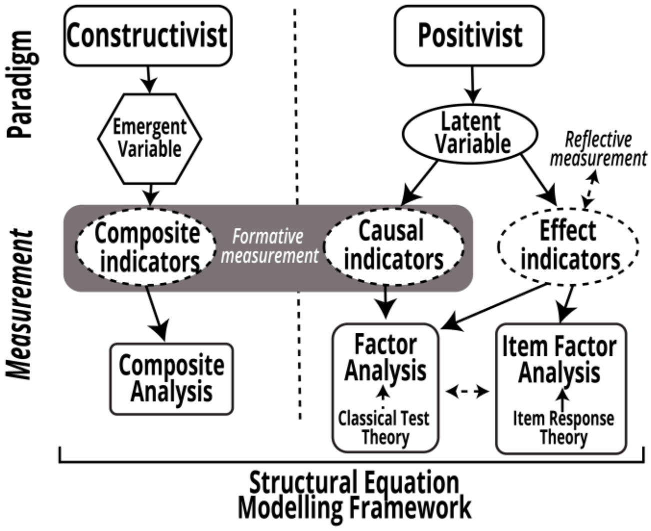


Figure 2. Heuristic relationship between research paradigms and measurement approaches and methods

In the following sections, an overview of relevant measurement methods in an unitary perspective of SEM (Henseler, 2021) will be presented to frame the understanding of chapters 2 to 6.

Measurement theory – reflective measurement

Validity evidence in a reflective measurement setting can be obtained according to two widely researched measurement theories: Classical Test Theory (CTT), and Item Response (IRT). Each presents contextual advantages (Raykov & Marcoulides, 2016), and as such, were used throughout work presented in later chapters in a complementary manner. Also, since the dominant methodologies in movement sciences are still CTT-based (Ntoumanis & Myers, 2016), a deeper focus will be given here to IRT methodologies.

Classical Test Theory and derived methodologies

Canonical CTT – as presented in the work of early authors - is a theory majorly concerned about error measurement, and thus reliability, making no claims about the existence of an underlying latent variable, nor implying any specific model (Markus & Borsboom, 2013; Sijtsma & Ark, 2021). Analysis is this lens included calculation of α coefficient (Cronbach, 1951), and item analysis (difficulty and discrimination index, and distractor analysis). These were conducted as preliminary means to assess item quality in Chapter 2 (*Content Knowledge* module), given their low sample size requirement.

Historically, the need to explicitly model latent variables brought the need of more robust statistical methodologies, and thus the onset of FA methodologies (e.g., Confirmatory Factor Analysis [CFA]), which are the base for most instrument development work in the psychological and social sciences (Immekus et al., 2019). These methods have the advantage of comporting estimation of large nomological networks with flexibility to constrain or free any possible parameter (Brown, 2015; R. B. Kline, 2016), and as such were used to study the macro dimensionality (domain-level and higher) of the PPLA measurement model in Chapter 6.

Despite its flexibility, regular FA is not equipped to deal with categorical variables (e.g., Likert-type items, correct/incorrect responses on a knowledge test), being designed to deal with continuous variables, with a linear response function (i.e., where higher level on the latent trait, should produce a higher score on the item; Reise et al., 2000). Some estimators in CFA (i.e., WLSMV) are able to partially circumvent this limitation, however require larger sample sizes to attain stable solutions (Brown, 2015), and do not provide the useful tools delineated below.

This framework also assumes that unidimensional scales are composed of items that present equal frequency distribution (same mean and standard deviation), thus not accounting for difficulty variations in items (van Schuur, 2003); which might lead to emergence of method factors pertaining to groupings of difficulty (Reise et al., 2000; Sijtsma & Ark, 2021).

Item Response Theories and derived methodologies

Item response theory presents a family of models that are inherently designed to use all information available in categorical items without resorting to WSLMV estimation in FA (Bock & Gibbons, 2021). In line with similarities with regular FA – assuming an underlying trait that influences responses to items - these models can be envisioned as a non-linear analogous versions of FA (Immekus et al., 2019), named collectively as Item Factor Analysis (IFA; (Wirth & Edwards, 2007). This non-linear character also aligns closely to the idea of a learning continuum posed in the APLF (Keegan et al., 2019). Due to this, IRT models present many widely documented advantages for item-level analysis (DeMars, 2010; Dima, 2018; Embretson & Reise, 2000; Hambleton et al., 1991; Singh, 2004), including:

a) explicitly modelling the interaction between the item characteristics and a person's ability/trait (denoted by the Greek theta letter θ) (Meijer & Tendeiro, 2018); allowing the estimation of essentially sample-independent parameters to evaluate model, item and person fit;

b) the possibility to analyze item and test information (and thus reliability) at different θ ranges, (Hambleton et al., 2010): i.e., the more information provided by a test at a particular level of θ, the smaller the measurement errors will be at this level. This differs from the usual CTT and FA's practice of using a general summary of reliability for all levels across θ (e.g., α, or McDonald's ω); and approaches the exhorted view of validity as contingent on intended application and interpretation of a test (American Educational Research Association et al., 2014);

c) allowing the use of mixed-format tests (e.g., tests composed of single-choice items and Likert scales), with no unbalanced impact upon tests scores (Embretson & Reise, 2000);

d) providing a robust visual inspection toolkit: including *item characteristic curves* (ICC; Figure 3) – which depicts the probability of correct response in function of θ, also named *option characteristic curve*s (OCC; Figure 4) in ordinal and nominal models; the *item information function* (IIF) – which depicts the amount of empirical information given by each item across θ (Toland, 2014); and *test information function* (TFT) which depicts the sum of all IIF included in the test, proving a complete picture of the test's ability to accurately measure different levels of θ, and allowing tailoring to target specific ability levels (Sijtsma & Ark, 2021).

IRT Models assumptions

All (unidimensional) IRT models bear three assumptions, that must be assessed before meaningful interpretation of parameters: 1) unidimensionality; 2) latent monotonicity; 3) local independence. (Finch & French, 2019). Unidimensionality implies that items homogenously represent a single latent trait – when this assumption is not tenable, multidimensional IRT (MIRT) models can be used (Reckase, 2009). Latent monotonicity implies that a student with a higher latent trait level will obtain a higher score on items. Local independence implies that, controlling for the latent trait, students' response to an item should not be influenced by their response on other items on the scale.

Non-parametric IRT

Two main classes of models can be conceived in IRT: non-parametric (NIRT), and parametric. NIRT models are so named due to imposing less restriction on the expected response pattern of items (Sijtsma & Ark, 2021). These are 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).

One of the available methodologies to explore NIRT application is Mokken Scale Analysis (MSA). MSA assesses the fit of two NIRT models initially proposed for dichotomous data (e.g., correct or wrong responses; Mokken, 1971), and later generalized for polytomous data (e.g., Likert-type scales; Molenaar, 1990, 1997): 1) monotone homogeneity model (MHM), and 2) double monotonicity model (DMM). If the MHM fits the data, individuals can be ranked on the latent trait based on their total score on the scale (sum score of all items). The DMM adds upon the above three canonical assumptions, that of Invariant Item Ordering (IIO), requiring non intersection of item response functions (analogous to item/option characteristic curve) (Sijtsma et al., 2011; Sijtsma & van der Ark, 2017). If the DMM fits the data, items can be ordered according to their difficulty (i.e., mean score), presenting an order that is equal across different-ability students, forming a hierarchical Mokken scale. This methodology was the base of the work conducted in Chapter 3.

Parametric IRT

Parametric IRT models for dichotomous data are named according to the number of item parameters that are freely estimated (Figure 3). These can be 1) discrimination/slope; 2) difficulty/threshold; and 3) guessing. The discrimination parameter (also a) can be interpreted as the strength of the relationship between the item and θ (De Ayala, 2009; DeMars, 2010); higher discriminating items are thus ideal; its FA analogous are factor loadings (Finch & French, 2019). The difficulty parameter (also b), as its name implies, can be interpreted as how hard an item is to correctly answer (i.e., the θ point at which the student is estimated to have 50% chance to answer correctly in a 1-parameter, or 2-parameter model; Nering, 2010).

Both the difficulty parameter and person θ estimate are on the same metric (interpretable as a z-score; Toland, 2014); and there is no ideal difficulty value, as it will depend upon which θ range the test intends to measure accurately (a departure from the idea of CTT that items should have a middle range difficulty) – its FA analogous are item intercepts. Finally, the guessing, or pseudo-chance parameter (also *c*) corresponds to the probability that a student will correctly answer an item without having achieved the estimated θ required (Finch & French, 2019); it has no analogous in FA.

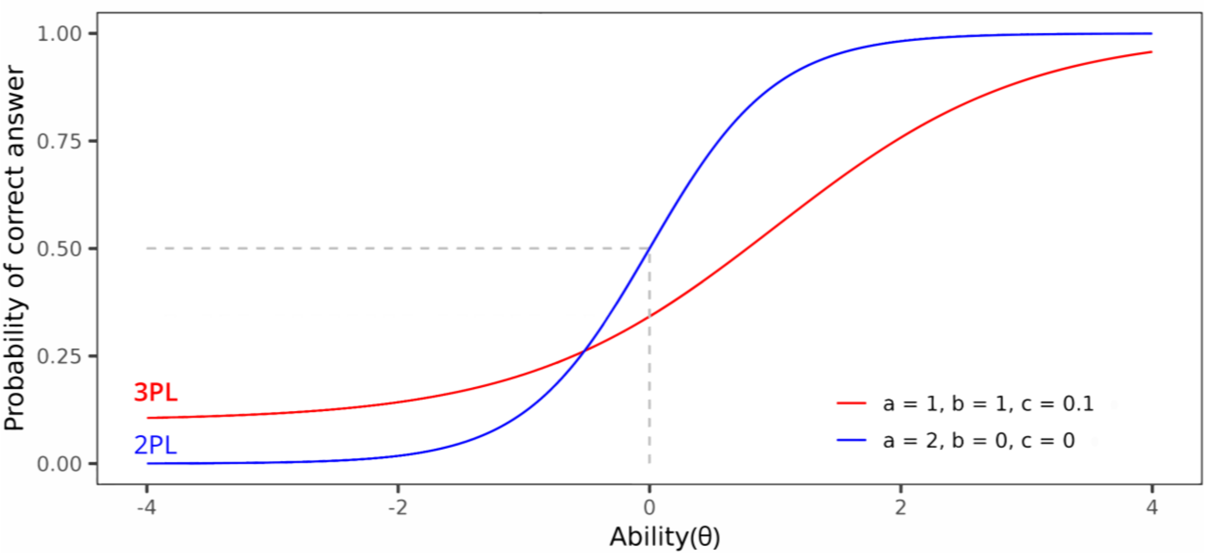


Figure 3. Item Characteristic Curve for a 2-parameter logistic (2PL) and a 3-parameter logistic (3PL) item - plot created using ShinyItemAnalysis (Martinková & Drabinová, 2019)

The 1-parameter logistic model (1PL; credited to Rasch, 1960) constrains items to have equal discrimination (i.e., freely estimates item's difficulty). This model implies a *tau-equivalent* model (McNeish & Wolf, 2020), and will result in a test score that is a simple transformation of the raw sum-score (Wu et al., 2016). The 2-parameters logistic model (2PL; Birnbaum, 1968) freely estimates discrimination parameters for all items, and uses them to weight responses. The 3-parameters logistic model (3PL; Birnbaum, 1968) freely estimates all three parameters mentioned above.

All dichotomous IRT models above discard information present in distractors (i.e., incorrect options), meaning that a response to distractors with different levels of correctness have no bearing on the estimation of θ for each student. In practice, each distractor will probably elicit different degrees of partial knowledge (Desjardins &

Bulut, 2018). Recognizing this, ordinal and nominal IRT models can be applied to extract further information from distractors. The 2PL equivalent for ordinal level data (in items which the level of correctness of each distractor is known beforehand), the graded response model (GRM; Samejima, 1969) can be applied to estimate a discrimination parameter, plus k-1 (*k= number of response categories*) thresholds (representing the θ point at which a student has 50% probability to score in that category or higher; DeMars, 2010). For nominal level data (items in which the level of correctness of each distractor is dubious or unknown) a nominal response model (NRM; Bock, 1972) can be estimated to model different discrimination (slope) and threshold (i.e., popularity; intercept) parameters for each distractor, and assess the relative correctness of each distractor.

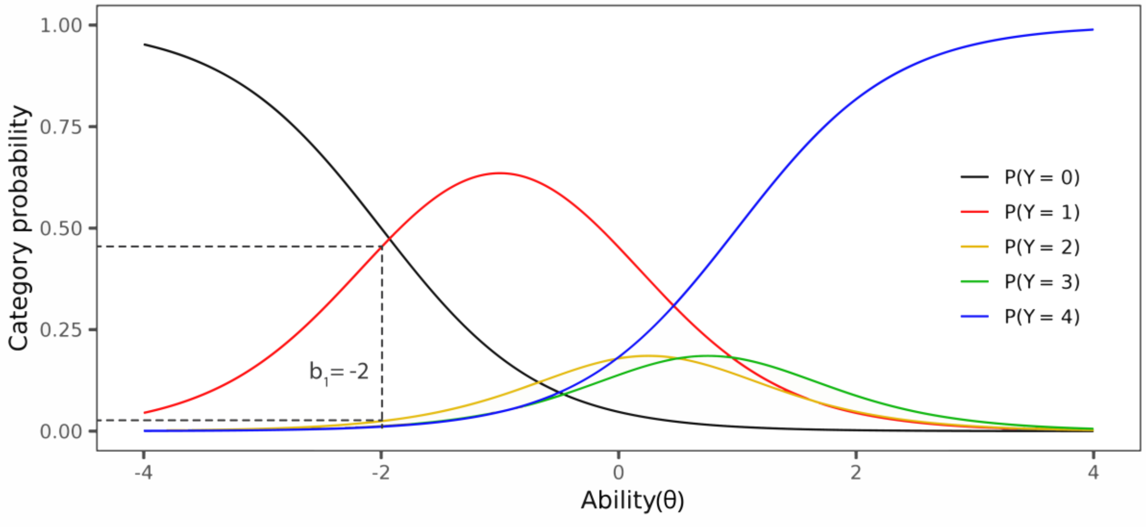


Figure 4. Option Characteristic Curve for a Graded Response Model item with a maximum score of 4 points (Y=4); a = 1.5, b_1=-2, b_2=0, b_3=0.5, b_4=1 – plot created using ShinyItemAnalysis (Martinková & Drabinová, 2019)

Note: at θ = b_1 the sum of the probabilities of obtaining 2, 3 and 4 points is 50%.

Another improvement is suggested in the form of nested logit models (NLM; Suh & Bolt, 2010), which suggest that response to a single selection item can be best characterized by a two-step hierarchical (nested) model: the first step modelling the probability of select the correct response versus the distractors through a dichotomous model (e.g., 2PL); and a second step, modelling the probabilities associated with each distractor through a NRM. These parametric IRT models will be used in Chapter 4. Its multidimensional extensions (MIRT) – which due to parameterization, will be similar in methodology to regular CFA models – will be explored in Chapter 5.

Measurement theory – formative measurement

In the interest of parsimony, formative measurement and its specifications will be described in Chapter 6's literature review, where its understanding will be germane to testing the full PPLA model in both a formative, and reflective manner.

Overview of methods

Participants

This project used four different main samples of experts, PE teachers and students throughout its studies: 1) an expert sample (N=11), 2) an initial cognitive interview students' sample (N=4), 3) a pilot testing sample from two schools ($N_{students}$ = 41, $N_{teachers}$ = 2), and 4) an initial validation sample from six different schools ($N_{students}$ = 521, $N_{teachers}$ = 22). Figure 5 provides a global flowchart for these samples. All participant classes were from grade 10 to 12 Portuguese public schools in the Lisbon metropolitan area. To avoid difficulties imposed by COVID-19 (e.g., school's refusal to participate to minimize outside contact) during the project's timeline, all schools were selected from a pool of schools with PE preservice protocol with the Faculty of Human Kinetics. Details of these sample are detailed in each main chapter, along with information regarding stratification and sample characteristics.

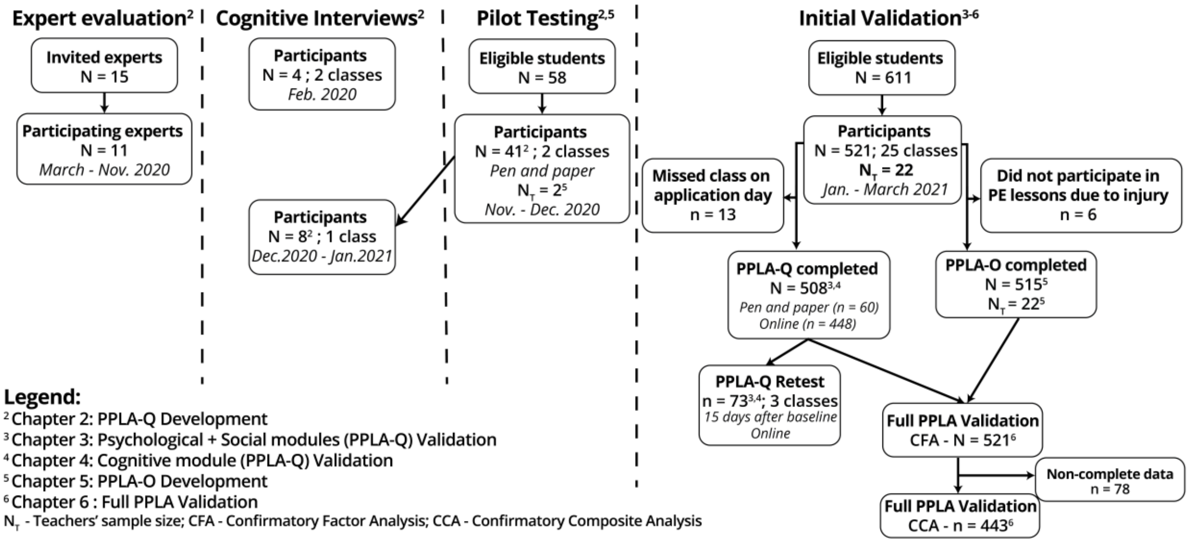

Expert evaluation[2]
Invited experts N = 15
Participating experts N = 11 March - Nov. 2020
Cognitive Interviews[2]
Participants N = 4 ; 2 classes Feb. 2020
Participants N = 8[2] ; 1 class Dec.2020 - Jan.2021
Pilot Testing[2,5]
Eligible students N = 58
Participants N = 41[2] ; 2 classes Pen and paper N_T = 2[5] Nov. - Dec. 2020
Initial Validation[3-6]
Eligible students N = 611
Participants N = 521; 25 classes N_T = 22 Jan. - March 2021
Missed class on application day n = 13
Did not participate in PE lessons due to injury n = 6
PPLA-Q completed N = 508[3,4] Pen and paper (n = 60) Online (n = 448)
PPLA-O completed N = 515[5] N_T = 22[5]
PPLA-Q Retest n = 73[3,4]; 3 classes 15 days after baseline Online
Full PPLA Validation CFA - N = 521[6]
Non-complete data n = 78
Full PPLA Validation CCA - n = 443[6]
Legend:
[2] Chapter 2: PPLA-Q Development
[3] Chapter 3: Psychological + Social modules (PPLA-Q) Validation
[4] Chapter 4: Cognitive module (PPLA-Q) Validation
[5] Chapter 5: PPLA-O Development
[6] Chapter 6 : Full PPLA Validation
N_T - Teachers' sample size; CFA - Confirmatory Factor Analysis; CCA - Confirmatory Composite Analysis

Figure 5. Portuguese Physical Literacy Assessment participants flowchart (data collection date and method in italics)

Procedures

All data collection procedures were approved by the Ethics Council of Faculty of Human Kinetics, the Portuguese Directorate-General of Education, and the directive boards of each participant school. All participating teachers and students provided a signed informed consent – underage students also provided a legal guardian's signature. Participation was confidential (i.e., only the student and respective PE teacher got access to student's results) and anonymous to the research team (i.e., each student was attributed a unique identification code throughout the study, generated by a spreadsheet sent to each teacher; this code replaced the student's name in all data collection procedures). Data collection was made by the lead investigator (i.e., the thesis author), during the timeframes detailed in italics in Figure 5. Four sets of different procedures were used, each detailed in the pertaining chapter. A summary is provided below.

1) Expert evaluation

Experts participated in two rounds of assessment of content validity of the PPLA-Q via email by filling a provided spreadsheet to rank the relevance and clarity of each item, and to comment/suggest any alteration needed. These procedures are detailed in Chapter 2.

2) Cognitive interviews

Cognitive interviews were conducted in three rounds using a semi structured format to assess the understanding and cognitive processes behind student's responses to the PPLA-Q items. There procedures are detailed in Chapter 2.

3) PPLA-Q Pilot Testing and Baseline validation

PPLA-Q was applied in self-administration format following a standardized introduction encouraging students to provide their most honest responses and reinforcing the confidentiality of data. This application was initially made in pen and paper format (Pilot Testing, and initial data collection for validation studies) and was later changed to online format due to a COVID-19 imposed lockdown. These procedures are further detailed in Chapters 2 and 3.

3b) PPLA-Q Retest validation

To assess test-retest reliability, a subsample of students completed the PPLA-Q a second time, in online format, 15 days apart from baseline. Procedures were otherwise equal to those detailed above for baseline data collection. This procedure is further detailed in Chapters 3 and 4.

4) PPLA-O Pilot Testing and Validation

Data collection for the PPLA-O was simultaneous with PPLA-Q application (i.e., teacher-reported data was sent to the lead investigator at the time of the PPLA-Q application). For this, PE teachers filled the PPLA-O in spreadsheet format with data pertaining to each participating student's a) observed levels of proficiency in the different physical activities taught during class (according to the criteria in the Portuguese PE syllabus), b) results from the FITescola® fitness protocols. Since for some classes, data collection occurred during lockdown, teachers were asked to provide the most updated information prior to lockdown. These procedures are detailed in chapter 5.

Measures

Since the focus of this thesis was the development and validation of the PPLA, theoretical frameworks for each developed measure are detailed in the chapter 2 (PPLA-Q) and chapter 5 (PPLA-O). Measures were subject to multiple revisions throughout the different studies, which are described in the main chapters. Table 1 provides an overview of measures collected, as well as number and typology of items used in the different phases. We would like to note that the International Physical Activity – Short Form (Craig et al., 2003), as used in the National Food, Nutrition and Physical Activity Survey (Lopes et al., 2017), was added to the final pilot testing and validation phases to allow for model identification detailed in Chapter 6.

Table 1. Overview of measures

Instrument (version)	Used in	PPLA -Q/O Number of items and typology per module
PPLA-Q		
PPLA-Q (0.1) – 90 items	Cognitive interviews (1st round)	**Cog**: 10 knowledge items (single and multiple selection) **Psy**: 40 Likert-type 5-point items **Soc**: 40 Likert-type 5-points items
PPLA-Q (0.2) – 90 items	Expert Evaluation (1st round)	**Cog**: *NC* **Psy**: NC **Soc**: *NC*
PPLA-Q (0.3) – 88 items	Expert Evaluation (2nd round)	**Cog**: *NC* **Psy**: *NC* **Soc**: 38 Likert-type 5-point items
PPLA-Q (0.4) – 87 items IPAQ-SF – 7 items[a] Sociodemographic items – 4 items[a] Regular PA participation – 2 items[a]	Pilot Testing + Cognitive interviews (2nd round)	**Cog**: *NC* **Psy**: *NC* **Soc**: 37 Likert-type 5-point items

Table 1. Overview of measures

Instrument (version)	Used in	PPLA -Q/O Number of items and typology per module
PPLA-Q (0.5) – 99 items	Cognitive interviews (3rd round)	**Cog**: *NC* **Psy**: 46 Likert-type 5-point items **Soc**: 43 Likert-type 5-points items
PPLA-Q (0.6) – 99 items IPAQ-SF – 7 items[a] Sociodemographic items – 4 items Regular PA participation – 2 items	Validation studies (construct validity and reliability)	**Cog**: *NC* **Psy**: *NC* **Soc**: *NC*
PPLA-O		
PPLA-O (1.0)	Pilot Testing and Initial Validation	**MCRT**: Proficiency level in 22 Physical Activities from the Portuguese PE Syllabus (Ordinal rating scale) **HRF:** 5 Health-Related Fitness protocols' results from FITescola® (Continuous and Ordinal scales)

PPLA-Q – Portuguese Physical Literacy Assessment Questionnaire; PPLA-O – PPLA Observation instrument; IPAQ-SF – International Physical Activity Questionnaire, Short Form (Craig et al., 2003); Cog – Cognitive module; Psy – Psychological module; Soc – Social module; MCRT – Movement Competence, Rules and Tactics module; HRF – Health-related fitness module; NC – number and general typology of items unchanged from last version; PA – Physical Activity; PE – Physical Education

[a] as used in the National Food, Nutrition and Physical Activity Survey (Lopes et al., 2017)

Analysis

Table 2 provides an overview of the methods used during the development and validation steps of the PPLA. Methods details, as well as criteria of assessment are thoroughly approached in each main chapter.

Table 2. Summary table of methods used in validation of the PPLA

	Domain level analysis				Higher-order analysis
	Physical[5]	**Cognitive**[2,4]	**Psychological**[2,3]	**Social**[2,3]	**PPLA**[6]
Content Validity	Content Analysis	Content Analysis Cognitive interviews Expert panel evaluation (CVI and multirater κ)			Based on previous literature review*
Construct validity					
Internal Structure / Dimensionality / Structural validity	Multidimensional Graded Response Model	Nested Logit Models + Graded Response Models	Mokken Scale Analysis[b] (H_s and H^T)		Confirmatory Factor Analysis – Bifactor models (ECV and PUC) Partial Least Squares SEM (Weight significance)
Hypothesis testing	Convergent (item parameters and bivariate correlations) Discriminant (factor correlations)	Convergent (item parameters)	Convergent (H_i) Discriminant (Disattenuated bivariate correlations)		Convergent (AVE and SMC) Discriminant (factor correlations and composite correlations)
Measurement invariance (sex)	DIF (Likelihood-ratio approach)	DIF (Likelihood-ratio approach) DTF (sDTF and uDTF)	DIF (stratified H_i) DTF (stratified H_s)		•
Criterion-related Validity					
Predictive	~	~	~		~
Concurrent	•	•	•		•
Reliability	Score-reliability (Marginal reliability)	Score-reliability (Marginal reliability, and Test Information) Test-retest (ICC, and Svenson's method)	Score-reliability (Sijstma – Molennar ρ) Test-retest (ICC)		Score-reliability (ω and ω_H)

[2-6] Chapter in which domains will be found

CVI – Content Validity Index; H_s – Coefficient H scale; H_i – Coefficient H item; H^T – Coefficient H trans; DIF – Differential Item Functioning; DTF - Differential Test Functioning; ECV – Explained Common Variance; PUC – Percent uncontaminated correlations; AVE – Average Variance Extracted; SMC – Squared Multiple Correlations (R^2); ICC – Intraclass Correlation Coefficient

[b] Monotone homogeneity model and Double monotonicity model

* (Dudley, Keegan, et al., 2017; Sport Australia, 2019)

• Not assessed; ~ Assessed as a requirement for model identification

CHAPTER 2 – PPLA-Q Development, Content Validity and Preliminary Construct Validity

Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) for adolescents (15-18 years) from grades 10-12: development, content validation and pilot testing

João Mota, João Martins, Marcos Onofre

Published on BMC Public Health: https://doi.org/10.1186/s12889-021-12230-5

[Impact Factor: 3.18]

Abstract

Background: The *Portuguese Physical Literacy Assessment* (PPLA) is a novel tool to assess high-school students' (grade 10-12; 15-18 years) Physical Literacy (PL) in Physical Education (PE); inspired by the four domains of the *Australian Physical Literacy Framework* (APLF), and the Portuguese PE syllabus. This paper describes the development, content validation, and pilot testing of the PPLA-Questionnaire *(PPLA-Q)*, one of two instruments in the PPLA, comprised of modules to assess the *psychological*, *social,* and part of the *cognitive* domain of PL.

Methods: Development was supported by previous work, analysis of the *APLF*, and literature review. We iteratively gathered evidence on content validity through two rounds of qualitative and quantitative expert validation (n= 11); three rounds of cognitive interviews with high-school students (n=12); and multiple instances of expert advisor input. A pilot study in two grade 10 classes (n=41) assessed feasibility, preliminary reliability, item difficulty and discrimination.

Results: Initial versions of the PPLA-Q gathered evidence in favor of adequate content validity at item level: most items had an Item-Content Validity Index ≥.78 and Cohen's **κ** ≥ .76. At module-level, S-CVI/Ave and UA were .87/.60, .98/.93 and .96/.84 for the cognitive, psychological, and social modules, respectively. Through the pilot study, we found evidence for feasibility, preliminary subscale and item reliability, difficulty, and discrimination. Items were reviewed through qualitative methods until saturation. Current PPLA-Q consists of 3 modules: cognitive (knowledge test with 10 items), psychological (46 Likert-type items) and social (43 Likert-type items).

Conclusion: Results of this study provide evidence for content validity, feasibility within PE setting and preliminary reliability of the PPLA-Q as an instrument to assess the psychological, social, and part of the cognitive domain of PL in grade 10 to 12 adolescents. Further validation and development are needed to establish construct validity and reliability, and study PPLA-Q's integration with the PPLA-Observation (an instrument in development to assess the remaining domains of PL) within the PPLA framework.

Keywords: physical literacy, assessment, physical education, content validity, pilot testing, high-school, adolescence.

Background

Physical Literacy (PL) is a concept based on lifelong holistic learning acquired and applied in movement and physical activity (PA) contexts (Sport Australia, 2019). Arguably, the most seminal contribution to the development of the concept in modern pedagogy have been the works of Margaret Whitehead (Whitehead, 2001, 2007, 2010), which conceptualized PL as the motivation, confidence, physical competence, understanding and knowledge to maintain physical activity throughout the life course.

Notwithstanding its lifelong development, sowing the seeds of PL during school-age seems critical, as participation in early childhood might predict adherence to active lifestyles throughout life (Telama, 2009; Telama et al., 2014), counteracting the rising levels of physical inactivity observed in adolescents and adults (Guthold et al., 2018, 2020). In this line, PL is argued as the main outcome of quality physical education (PE) in schools (UNESCO, 2015), since it provides a privileged environment – mandatory, free and qualified – for learning the life skills and values needed for active and global citizenship(Onofre, 2017); as well as being the only opportunity to participate and learn from PA for some school-aged children and adolescents (Woods et al., 2010). Thus, many authors have underlined the need to operationalize this concept in school curricula and educational policies (Corbin, 2016a; Dudley, 2015; Dudley, Cairney, et al., 2017)

Despite a general consensus on the ultimate goal of PL – sustained lifelong PA participation (Whitehead, 2013a, 2013b) –, its proposed conceptualization and constituent elements differ across sources(L. Edwards, Bryant, Keegan, Morgan, & Jones, 2017; Liu & Chen, 2021; Martins et al., 2020). These range from philosophically-driven conceptualizations, like Whitehead's PL original proposition (Whitehead, 2001) – rooted in the philosophical tenets of monism, phenomenology, and existentialism – to diametrical conceptualizations focusing solely on one of its aspects (e.g., fundamental movement skills) (D. B. Robinson et al., 2018). Although recognized as a rich theoretical concept, the former might lack pragmaticism to be implemented in practice (L. Edwards, Bryant, Keegan, Morgan, & Jones, 2017): while the later might deviate from the holistic nature of PL, compromising crucial elements like pleasure and enjoyment in taking part in PA

(Pot et al., 2018). As such, a middle-ground compromise might offer a tenable solution: providing clear and measurable outcomes, while honoring most of the philosophical-driven premises that define the concept(D. B. Robinson et al., 2018; Young et al., 2019). To this end, a team of Australia-based researchers developed the *Australian Physical Literacy Framework* (APLF) (Sport Australia, 2019), a research-based, integrated model of PL in the *physical*, *cognitive*, *psychological* and *social* domains with 30 different elements – novel in recognizing the contribute that PL might play in cultural and social participation. It provides a clear focus on a learning continuum, inspired by the *Structure of Observed Learning Outcomes* taxonomy (Biggs & Collis, 1982), designed to include individuals in different states of their PL journey: from their first steps (*pre-foundational*) to higher stages of proficiency (*transfer & empowerment*) (Keegan et al., 2017, 2019).

Physical Literacy Assessment

Given evaluation's essential role in PL implementation and practice (Corbin, 2016a) a few assessment instruments have been developed, under diverse conceptual models(L. Edwards, Bryant, Keegan, Morgan, Cooper, et al., 2017; Liu & Chen, 2021; Shearer et al., 2021). Of these , the most prolific research-wise have been the Canadian Assessment of Physical Literacy (CAPL) (Francis et al., 2016; Gunnell et al., 2018), and the Physical Literacy Assessment for Youth (Cairney et al., 2018)(PLAY). The CAPL is comprised of standardized assessments developed for children from 8 to 12 years (Longmuir, Gunnell, et al., 2018) (with preliminary testing done in 12 to 16 year-olds; Blanchard et al., 2020), to assess daily behavior, physical competence, motivation and confidence, and knowledge and understanding. The PLAY tools have been developed to assess children from 7 years up (with recommendations mainly targeted at the 7–12-year range), comprised of measures of motor competence, comprehension, and confidence. Both tools integrate observational procedures and self-report, and feature overall good feasibility in PE (Shearer et al., 2021)but lack options for older adolescents (15-18 years), a critical age range in Portugal which presents lower levels of PA(Baptista et al., 2012; Martins, Marques, et al., 2019; Matos & Equipa Aventura Social, 2018) – making them a priority target in the Portuguese PE setting.

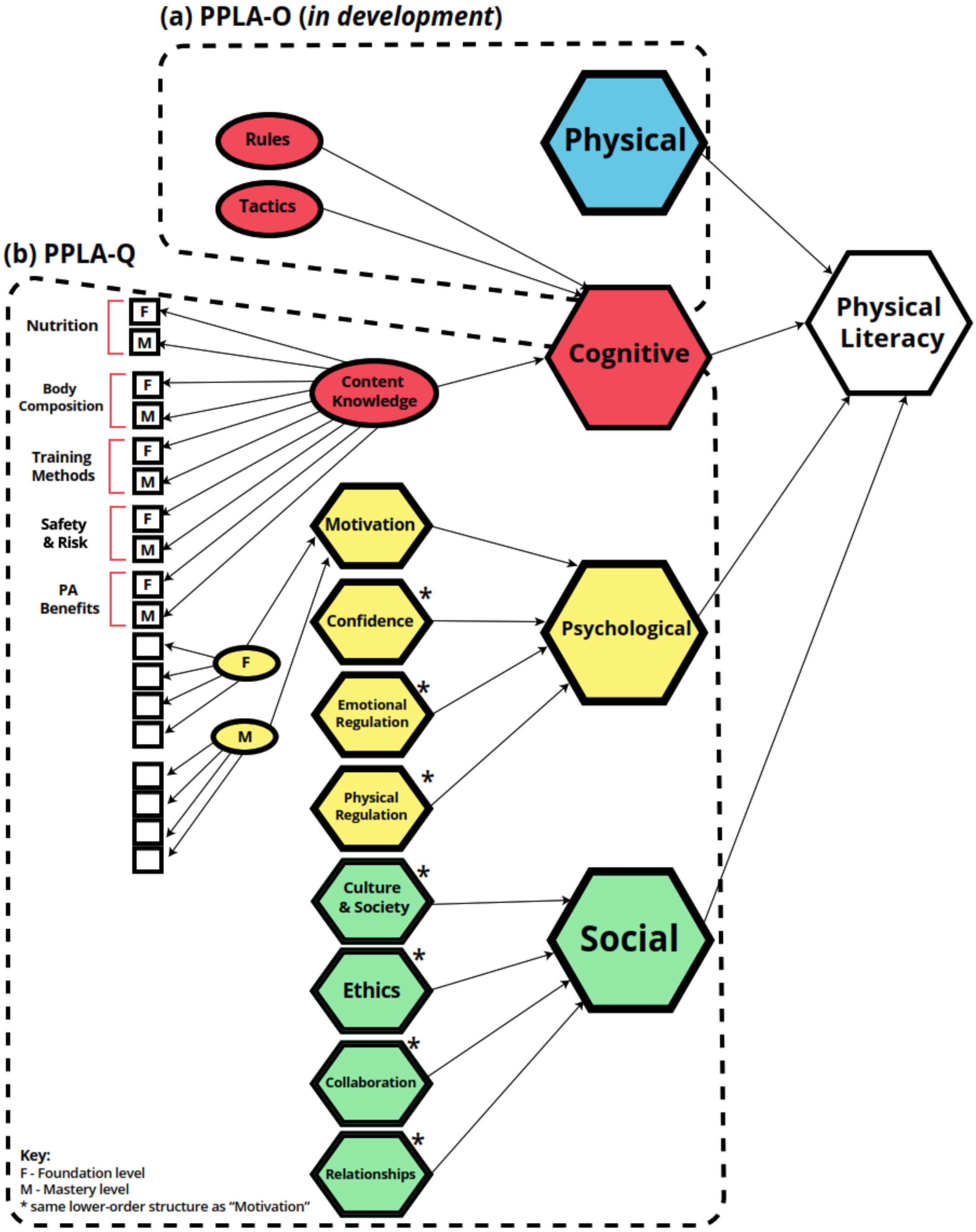


Figure 6. Portuguese Physical Literacy Assessment (PPLA) hypothesized model and instruments.

Legend: PPLA is a tool comprised of two different instruments: ***a)*** *PPLA-Observation (PPLA-O) – assesses the physical domain, and the Rules and Tactics elements of the cognitive domain of PL;* ***b)*** *PPLA-Questionnaire (PPLA-Q) – assesses the psychological, social and Content Knowledge element of the cognitive domain of PL*

Portuguese Physical Education and PL

The Portuguese PE national syllabus (PPES) was designed under the Crum's socio-critical conception of PE, contemplating integrated learning in the motor, cognitive, affective and social domains, to empower students to engage in significant PA, and actively participate in the movement culture throughout their lives (Crum, 1993); expanding beyond a restricted and instrumental participation in PA(Tinning, 2015). Although the initial development of this syllabus slightly predates Whitehead's influential works on PL (Whitehead, 2001), it implicitly aligns with the latter's ontological and epistemological premises. Akin to a phenomenological and existentialist perspective (Durden-Myers et al., 2018), it advocates pedagogical practices of differentiation, allowing a high degree of flexibility towards the achievement of curricular goals, recognizing that each individual enjoys and values different forms of movement; while using assessment as a tool to motivate and identify where every student should work to improve, in line with strategies proposed both by PL (Durden-Myers et al., 2018) and assessment specialists (Harlen, 2007).

The PPES distinguishes three learning areas: 1) Physical Activities, 2) Health-Related Fitness, 3) Knowledge. In the first area, it advocates the participation in a wide range of physical activities (sport-based team and individual activities, rhythmic and expressive activities, nature exploration activities, and traditional games), enabling students to choose from an eclectic array of physical activities throughout their life. In each of these activities, student progress is charted through 3 levels of competency – introductory, intermediate, and advanced – integrating 1) mastery of specific movement skills, 2) cognitive skills related to tactical decision, 3) knowledge and application of activity rules and 4) prosocial behavior during said activity (Onofre et al., 2020). This multilateral learning through participation in physical activities is supported by the development of health-related fitness, and the knowledge and skills needed to lead a healthy lifestyle through personal significant PA (second and third areas of the PPES, respectively).

Despite having common points with most PL definitions and models, the PPES curricular and pedagogical choices align more closely with the Australian proposal previously presented, since the latter explicitly includes the social domain as an integral part of the PL development, as well as elements pertaining to tactical and rules learning. Also, the APLF maps all development through the usage of a modified continuum based on the *Structure of Observed Learning Outcomes* taxonomy (Biggs & Collis, 1982), which recognizes that learning might differ not only in quantity (i.e., being less or more skilled/knowledgeable) but in qualitative state as well (i.e., going from a descriptive, surface knowledge to a relational understanding of a skill/knowledge); a principle mirrored in the three levels of competency in the PPES.

Considering these specificities of the PPES design and implementation, none of the presented PL assessments provide a complete picture of learning in all four domains; nor were they designed for older adolescents. As such, we developed an instrumental system – *Portuguese Physical Literacy Assessment (PPLA)* – to address this gap and use PE as a privileged mean for PL development in Portugal.

The Portuguese Physical Literacy Assessment (PPLA)

The PPLA was designed to provide a detailed and feasible assessment of each student's PL journey, and to inform pedagogical decisions (at local, regional, and national level) towards a more meaningful and targeted environment to promote PL learning of grade 10-12 (15-18 years) adolescents. The PPLA (Figure 6) is based on the PPES and integrates assessment in the four domains of the APLF, using two instruments: a) the PPLA-Observation (PPLA-O), and b) the PPLA-Questionnaire (PPLA-Q). The PPLA-O (still in development) uses observational data collected by the teachers during regular PE classes (competency levels in the different physical activities, and physical fitness levels using standardized protocols) to assess the physical and part (*Rules*, and *Tactics* elements) of the cognitive domain. The PPLA-Q, which will be the focus of this article, uses a knowledge test (with multiple-choice questions) and self-report (Likert-type scales) to assess the psychological, social and the remaining part of the cognitive domain (*Content Knowledge* element). Both these instruments were designed to be applied together to provide a holistic picture of each student's PL journey.

PPLA (Figure 6), following the APLF conceptualization of a learning continuum summarizes its five development levels (for each element of the four PL domains), into two learning levels: *Foundation* and *Mastery.* This simpler structure still captures the qualitative change in the learning experience, separating *surface* *learning* from *deep learning*, while providing a more parsimonious and feasible instrument.

The *Foundation* level represents the initial development of each element, building affective, cognitive, psychomotor and social structures that enable participation in movement and physical activities, albeit in an isolated, instrumental or externally focused manner (i.e., to obtain benefits/rewards, or conform to the norm) – akin to the *Unistructural* and *Multistructural* levels of the *Structure of Observed Learning Outcomes* taxonomy, and the foundational levels of Bloom's Revised Affective Taxonomy (Krathwohl et al., 1964).

Mastery level represents a deeper development of the element, invoking metacognitive processes, relational understanding, or internalized behaviors (i.e., integrated into the individual's sense of self) regarding participating in movement and physical activities – derived from the *Relational* and *Extended Abstract* levels of *Structure of Observed Learning Outcomes* taxonomy, and higher levels in Bloom's affective taxonomy.

As such, based on previous constructs studies of PL (Cairney et al., 2019; Gunnell et al., 2018) and the structure implied by the APLF, we hypothesize a hierarchical measurement model, with PL conceptualized as a fourth-order formative construct (Figure 6) composed by its four domains (third-order formative constructs). Each domain is then formatively composed by several elements (second-order formative constructs), in turn composed by two first-order constructs, reflexively formed by a set of manifest indicators (i.e., items).

The distinction between *formative* constructs (i.e., composites) and *reflexive* constructs (i.e., factors) is important here. While the later assumes that items (or lower-order constructs) are interchangeable – since they measure the same underlying trait (i.e. are unidimensional) – and thus are expected to covary, the former assumes the opposite: that its composing items are not interchangeable, and are not expected to covary – where an omission or deletion of an item changes the essence of the construct being measured (Andrich & Marais, 2019; Hair et al., 2017; Jarvis et al., 2003).

Based on this conceptual framework, in a series of studies, we sought to develop the *PPLA-Questionnaire (PPLA-Q)*, an instrument comprised of modules to assess grade

10-12 adolescents' *psychological*, *social,* and part of *cognitive* domains of PL; and gather evidence for its content validity, feasibility within PE setting, preliminary reliability, item difficulty and discrimination.

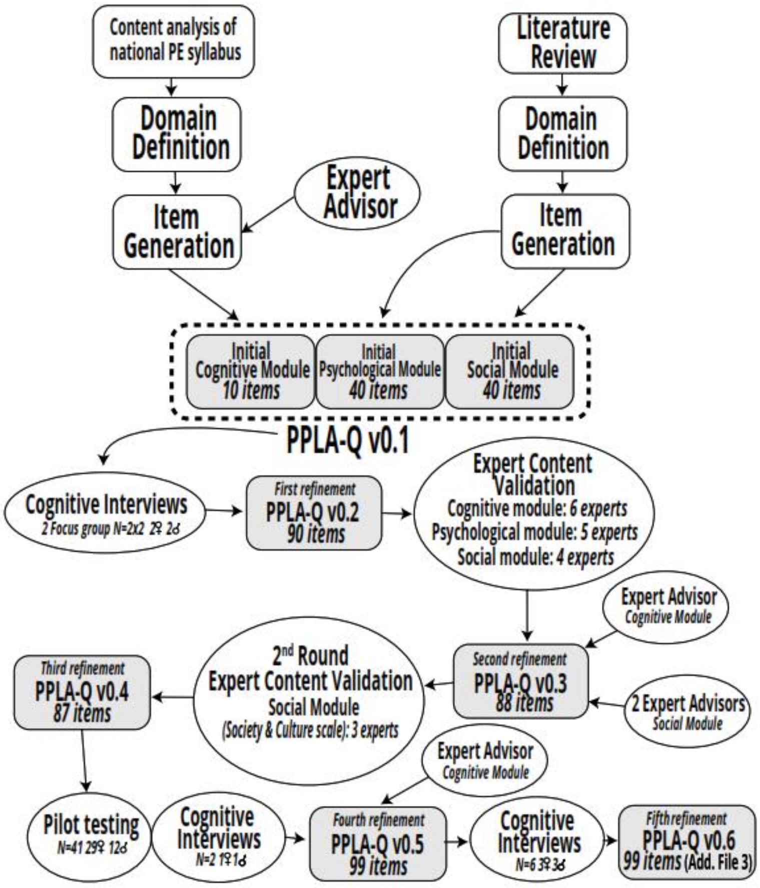


Figure 7. Overview of development studies of the Portuguese Physical Literacy Assessment – Questionnaire (PPLA-Q)

Methods

Studies overview

The development of the Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) entailed a series of studies (Figure 7), based on a multiple phase design (Armstrong et al., 2005; Boateng et al., 2018; Longmuir, Gunnell, et al., 2018), inspired by the psychological, social and cognitive domains of the PL model proposed in the APLF (Dudley, Keegan, et al., 2017; Sport Australia, 2019), and by the Portuguese PE syllabus (Ministério da Educação, 2001a, 2001b, 2018b).

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. All methods were performed in accordance with the relevant guidelines and regulations.

PPLA initial development was based on previous work done in the Erasmus+ Sport Project: PhyLit – Physical Literacy (590844-EPP-1-2017-1-UK-SPO-SSCP, January– December 2018) , where a panel of experts selected – among the 30 proposed by the APLF – relevant elements for developing and advocating PL as an essential competence for European citizenship, based on a literature review of existing conceptualizations (Martins et al., 2020).

Initial development for each of the three modules of PPLA-Q entailed domain identification and item generation; followed by an iterative process to gather judgmental evidence on content validity that included: two rounds of qualitative and quantitative expert validation; three rounds of cognitive interviews with high-school students; and multiple instances of expert advisor input. We also conducted a pilot study to assess feasibility of the questionnaire in PE and collect preliminary data on reliability and construct validity.

Domain identification

Based on literature review, we established a theoretical framework for each of the eight elements in the psychological (*Motivation, Confidence, Emotional Regulation,* and *Physical Regulation)* and social domains (*Culture & Society, Ethics, Collaboration,* and *Relationships*) (Table 3). The literature review conducted by Dudley and colleagues (Dudley, Keegan, et al., 2017), in the report preceding the creation of the APLF, was used as starting point to identify established and relevant theories for each element in the literature of motor development, physical education and/or physical activity. Then, constructs with higher conceptual proximity were chosen – caring to minimize overlap –, mapped to the two-level framework, and operational definitions derived from the APLF.

For the Cognitive Domain, we conducted a content analysis of the Portuguese PE syllabus (PPES) to identify key learning objectives coherent with the *Content Knowledge*, *Tactics* and *Rules* elements of the APLF. In this process, to ensure adequate content representation, we subdivided the *Content Knowledge* element into different content themes (*Nutrition, Body Composition, Training Methods, Safety &*

Risk, PA Benefits); each was then mapped to the two-level framework and its operational definition derived from the PPES (Table 4).

Since tactical behaviors and adherence to rules (i.e., as a participant, and as a referee or judge) are better assessed through direct observation of the student's behavior during PE, we chose to include the *Tactics* and *Rules* elements alongside the assessment of the physical domain (in the PPLA-O). As such, these elements will not be further discussed here, despite them being integral part of the Cognitive domain.

Item generation

Psychological and Social Modules

Items in the Psychological and Social domains were developed to conform to self-report measurement using Likert-type scaling, given its adequacy and versatility to measure attitudes, beliefs and self-perceived abilities (DeVellis, 2017; Price, 2017). An initial goal was set to generate a 5-item subscale per learning level (two subscales per element, four elements per module). This was a compromise between the size of the resulting questionnaire, and a larger initial item pool to provide margin for eliminating poorly performing items during testing (Clark & Watson, 1995; DeVellis, 2017); down to four per subscale – the recommended number to calculate reliability and further test measurement models (Bollen, 1989).

In an effort to use psychometrically sound items as a reference for item generation (Kyriazos & Stalikas, 2018) a non-systematic literature review was conducted using ERIC, Google Scholar, Scopus and ProQuest databases to identify a first round of eligible articles for each element, which were then used to refine further searches for articles. In these, we selected published and validated scales or subscales (in English or Portuguese), amply used in PE, sport, or PA contexts, and sampled items that adhered to each level's operational definitions (Table 3). When various identical items overlapped in content, those with higher item loading were selected.

After permission for adaptation was granted by each scale's lead author, sampled items were used as reference to generate items in Portuguese, based on the examples provided by the APLF, and technical recommendations available in the literature (Artino et al., 2014; Clark & Watson, 1995; DeCastellarnau, 2018; DeVellis, 2017; Price, 2017). When suitable reference scales were not available or failed to achieve full content representation for the element, or level, items were generated according to previous literature view.

All items used a consistent 5-points unipolar response scale, to maximize reliability and validity (DeCastellarnau, 2018; Furr, 2011) . 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?).*

Cognitive Domain

For their suitability to test cognitive ability and knowledge (Price, 2017) , and ease of application, multiple-choice questions were generated for each content theme and level (10 items), according to technical advice presented by the literature (Considine et al., 2005; Scully, 2017), and by an educational assessment expert (PhD holder with extensive experience as a PE and graduate-level college professor, as well as an employee in the Portuguese Institute for Educational Assessment).

Throughout the process in all modules, the lead author acted as item generator, while remaining authors acted as co-validators to ensure preliminary content validity.

Table 3. Domain identification for the psychological and social domains

	Theoretical framework	Operational definition (number of items)	Instruments used as reference
Psychological Domain			
Motivation	Self-determination Theory(Deci & Ryan, 2000, 2008)	**Reasons for engaging in movement and physical activity in response to internal or external factors**[1] Foundation: Controlled motivation (5 items) Mastery: Autonomous motivation (5 items)	Behavioral Regulation in Exercise Questionnaire – 3 (BREQ-3) (Markland & Tobin, 2004; Wilson et al., 2006)
Confidence	Psychological need satisfaction -Perceived competence (Deci & Ryan, 2002)	**A belief in self-worth and ability to perform in movement and physical activity** [1] Foundation: Beliefs of self-worth and ability (5 items) Mastery: Beliefs of self-worth and ability in challenging contexts (5 items)	Psychological Need Satisfaction in Exercise Scale (PNSE)(Wilson et al., 2006)
Emotional Regulation	Emotional Intelligence (Goleman, 2005)	**Ability to manage emotions and resulting behaviors in relation to movement and physical activity**[1] Foundation: Awareness of own emotions and other's (5 items) Mastery: Emotional regulation and control (5 items)	Wong and Law's Emotional Intelligence Scale (WLEIS)(Wong & Law, 2002)
Physical Regulation	NA	**Recognizing and managing physical signals such as pain, fatigue and exertion**[1] Foundation: Awareness of physical signals (5 items) Mastery: Regulation and management of physical signals (5 items)	NA
Social Domain			
Culture & Society	Sport Education(Siedentop, 1998)	**Appreciation of cultural values which exist within groups, organizations and communites**[1] Foundation: Participation in sport's cultural phenomena (5 items) Mastery: Valuing participation in sport's cultural phenomena and encouragement of others to do so (5 items)	NA
Ethics	Moral development(Gibbs, 2014; Kohlberg, 1964)	**Moral principles that govern a person's behavior, relating to fairness and justice, inclusion, equity, integrity, and respect**[1] Foundation: Respect for basic moral and ethical principles in physical activity contexts (fair-play) (5 items) Mastery: Autonomy and empowerment of others in respecting moral and ethical principles in physical activity contexts (fair-play) (5 items)	Fair Play Questionnaire in Physical Education (FPQ-PE)(Hassandra et al., 2002)
Collaboration	Personal and Social Responsibility(Hellison, 2011)	**Social skills for successful interaction with others, including: communication, cooperation, leadership and conflict resolution**[1] Foundation: Respect and cooperation with others Mastery: Caring and leading others to success	Personal and Social Responsibility Questionnaire (PSRQ) (W. Li et al., 2008)
Relationships	Psychological need satisfaction -Perceived Relatedness(Deci & Ryan, 2002)	**Building and maintaining respectful relationships that enable a person to interact effectively with others.**[1] Foundation: Interaction and relatedness with others Mastery: Management and maintaining relationships with others	Psychological Need Satisfaction in Exercise Scale (PNSE)(Wilson et al., 2006)

Table 4. Domain identification for the cognitive domain

Content	Operational definition
Nutrition	Foundation: Identify healthy food options (C1) Mastery: Evaluate impact of energetical balance in regulation of body weight (C2)
Fitness and training	Foundation: Identify main components of physical fitness (C3) Mastery: Evaluate training methods for components of physical fitness (C4)
Safety and risk	Foundation: Identify safety rules and principles in physical activities (C5) Mastery: Interpret doping's impact on health and sport ethics (C6)
PA Health Benefits	Foundation: Identify general physical activity guidelines for children, adolescents, and adults[a](C7) Mastery: Relate types of training with their benefits for health (C8)
Body composition	Foundation: Identify Body Mass Index's calculation formula (C9) Mastery: Evaluate body composition profile and make recommendations (C10)

[a]According to World Health Organization (2010)
PA – Physical Activity

Content Validity

Content validity pertains to the extent to which a set of items represents the intended construct (DeVellis, 2017). It requires evidence of content relevance, representativeness, and technical quality, assessed through evaluation by experts and population judges (Boateng et al., 2018). As such, we led an iterative process with multiple rounds (Polit et al., 2007), collecting both qualitative and quantitative evidence from both parties.

Cognitive interviews

Cognitive interviewing is a qualitative method to assess whether a survey fulfills its intended purpose, through interview of selected individuals, before, during and after pretesting (Willis & Artino, 2013). In our study, cognitive interviews were conducted in three rounds, in two different high schools in Lisbon – one with a dominantly higher socioeconomic status population, and another with a lower socioeconomic status population – involving students of the target age-group (15-18 years), through different phases of development of the PPLA-Q. Before participation, informed consent was provided by all students and their legal guardians. All interviews were conducted by the lead author during PE classes and recorded. Initial interviews were more extensive (i.e., more content, less depth), while the latter ones were progressively more intensive (i.e., narrower content, higher depth). This strategy balanced gross evaluation (e.g., format, conceptual breadth) in earlier phases with fine-tuning (e.g., wording, syntax) in later ones.

In February 2020, in each high school, a cognitive interview was conducted with a group of two students from grade 10 (aged 15) and another with two grade 11 students (aged 17). We sought to diversify these groups by 1) including, in each, a female, and a male, with different PE competency levels (according to their teacher); and 2) including students from different majors: one group from a Science, Technology, Engineering and Math major, the other from a Humanities and Arts major. Students were asked to fill in a draft version of the PPLA-Q, marking any items with ambiguous or unclear wording. Afterwards, an interview was conducted to probe for comprehension of items – focusing on the ones marked by students. Students were asked to verbally express their understanding of each and paraphrase it according to their own words. They were also questioned about general issues of the questionnaire (i.e., length and structure, layout, ease of reading, rating scales, comprehension of instructions and item stems). Average duration was 45 minutes.

In December 2020, a second round of individual cognitive interviews was conducted immediately after pilot testing (version 0.4 of PPLA-Q) with two students from grade 10 (1 female, 1 male, both aged 15) from a Humanities major class. Here, students who posed abundant questions during the questionnaire application were selected to better study the clarity of the items. Given time constraints of the project, this round enlisted less students that initially warranted. Students were asked about their comprehension of selected items – those which were the target of most of student's questions during pilot testing, as well as those previously revised. Average duration was 17 minutes.

In January 2021, a third round of individual cognitive interviews was conducted with six different students from the same grade 10 Humanities class recruited for last round (3 female, 3 males, mean age = 14.8 years). These were selected according to as different PE competency levels as possible (reported by the teacher). They were asked about their comprehension of all items changed from version 0.4 to version 0.5. Average duration was 15 minutes.

Evaluation by experts

Among the many methods available, Content Validity Index (CVI) and Cohen's coefficient *kappa* (**κ**) for interrater agreement were used to systematically assess expert consensus on content validity of an instrument (Boateng et al., 2018; Wynd et al., 2003).

Given different subject matter for each of the modules, expert selection was stratified per module to allow for more useful inferences. We intended to collect evidence from 6 experts – following recommendations of 5 (Lynn, 1986) - with relevant scientific and professional background, on each of the questionnaire's domains (i.e., psychology of physical activities/sport; sociology of sport; educational assessment/curriculum development), and ideally with experience in instrument development (Davis, 1992). According to their expertise, each expert was invited to participate either (a) in all 3 modules (n=3); (b) in 2 modules (n=1) or (c) in a single module (n=11). Further characteristics about the expert are summed up in Additional File 1 .

Experts were invited through an email presenting the project's goals and explaining the motives for selection, containing (1) instructions for intended contribution, (2) a draft version of PPLA-Q, and (3) a spreadsheet file. Operational definitions for each construct were also provided – as content validity is inextricably linked to the definition of constructs under examinations (DeVellis, 2017). In the spreadsheet file, experts were asked to: (1) rate each item on its relevance ("*How important is the item to assess the targeted construct*?") and clarity ("*Is the wording of the item clear*?"), (2) provide suggestions for item improvement, (3) provide suggestions on questionnaire structure, instructions, and rating scale. Both relevance and clarity were assessed with a 4-point Likert-type Scale (Lynn, 1986). For relevance the rating options were: 1 = *not relevant*, 2= *somewhat relevant*, 3 = *quite relevant*, and 4 = *very relevant* (Wynd et al., 2003). For clarity, the options were: 1= *not clear*, 2 = *item needs revision*, 3= *clear, but needs minor revision*, 4= *very clear* (Zamanzadeh, 2015). During analysis, both ratings were collapsed into two dichotomous categories ("content invalid" and "not clear" for ratings of 1 and 2, and "content valid" and "clear" for ratings of 3 and 4, respectively) (Lynn, 1986).

Of the invited experts, the actual first-round expert sample (n=10) consisted of 2 global experts (3 modules), 1 expert rating 2 modules, and 7 experts rating a single module. Another expert provided solely qualitative feedback (i.e., suggestions of improvements for item and questionnaire structure) on 2 of the modules, with no quantitative ratings. We had minimal missing data, with no bearing on calculations, since all adjusted for the total number of raters in each item.

All calculations used *RStudio* (RStudio Team, 2020) with *R* version 4.0.2 (R Core Team, 2020). CVI was computed both at item level (I-CVI) and module level (S-CVI/Ave and S-CVI/UA). Polit & Beck (Polit & Beck, 2006) argue that given diverse uses of CVI in the literature, one should explicit their calculations. We computed I-CVI as the proportion of experts rating each item as *content valid*. S-CVI/Ave was computed as the average of I-CVI for each module, while S-CVI/UA was computed as the proportion of items with I-CVI = 1 (i.e., *universal agreement*) for each module.

Many authors have criticized drawing content validity evidence based solely on CVI, given its susceptibility to chance agreement. They propose that Cohen's *kappa* (Cohen, 1960) – a statistic which accounts for the possibility of chance agreement of experts – be used alongside CVI (Wynd et al., 2003). For this purpose, *kappa* (**κ**) was computed using Fleiss's modified version for multiple raters (Fleiss, 1971; Polit et al., 2007) for each item:

$$k = \frac{(P_a - P_c)}{(1 - P_c)}$$

where P_a (proportion of agreement) = *I-CVI* for the item, and where P_c (probability of a chance agreement), for a random binomial variable, with one outcome:

$$P_c = \left(\frac{N!}{A!\,(N-A)!}\right) *.5^N$$

With N = number of experts, and A = number of experts rating item as content valid.

For item clarity, an identical procedure was used to calculate proportion of agreement (akin to I-CVI), and a ***κ*** statistic for each item. as the usual application of Content Validity Index (CVI) pertains to a global evaluation of the item (Polit et al., 2007), which might hide some crucial aspects of the item's quality, confounding the conceptual relevance of the item, with the clarity of its wording

We used κ to inform item level decisions, evaluating item relevance as fair (.40 to .59), good (.60 to .74) and excellent (> .74); κ lower than .40 prompted elimination of the item (Cicchetti & Sparrow, 1981; Fleiss, 1971). For clarity, the threshold increased to discriminate items needing minor revisions and ensure higher clarity throughout: we evaluated items as clear (κ > .74) and as needing revision (κ <.74).

Scale level decisions were informed by S-CVI. We used literature recommendation of .80 as an adequate level of agreement for the more stringent S-CVI/UA (Davis, 1992), and .90 for S-CVI/Ave (Waltz et al., 2010).

In the second round of expert evaluation, the same procedures were followed to gather evidence of content validity on the revised *Culture & Society* scale (version 0.3), targeting a lower number of experts (n=3, 2 of which participated in the previous round), due to time constraints in the project schedule.

Pilot Testing

Pilot testing, or *pretesting* constitutes an opportunity to (1) test the application of items in development to a representative sample of target population (American Educational Research Association et al., 2014); (2) gather feasibility evidence to plan a larger scale study (Hertzog, 2008); and (3) gather data for preliminary item analysis and estimates of reliability (Johanson & Brooks, 2010).

Although no clear-cut standard is available for sample size of pilot tests, Hertzog(Hertzog, 2008) suggests a sample size of 40 individuals for estimating preliminary data on reliability and item discrimination. As such, we pilot tested version 0.4 of PPLA-Q with a sample of 41 grade 10 students (down from an initial pool of 58 students who received the informed consent), from two classes of the different schools in Lisbon ($n_{school1}$ = 19, $n_{school2}$ = 22) aforementioned – one with a higher socioeconomic status population, another with a lower socioeconomic status population, as attested in each school's pedagogical project. This sample was composed of 29 females (71%) and had an average age of 15 (0.4) years. All students provided an informed consent signed by themselves and their legal guardian.

PPLA-Q was self-administered, in pen and paper format, both in PE gym and classroom setting – to test likely settings expected for future application – in presence of the lead author. Students were instructed to state any question regarding questionnaire's instruction, items, or rating scales. Application was timed to calculate average completion time; attrition rate was calculated as the percentage of students completing the study, among those who received the informed consent.

Preliminary Item Analysis

Psychological and Social modules

Given the novel status of any construct validation under the APLF model, as well as a complex and high number of constructs under analysis, we chose to conduct preliminary item analysis using the partial least squares – structural equations modelling (PLS-SEM) framework (Hair et al., 2019). No a priori power analysis was conducted, since our goal was to gather very rough insights into the statistical behavior of the measurement model of items. Despite this, our sample size approximated the thumb-rule of 10 times the maximum number of indicators per construct (Hair et al., 2017).

Prior to calculation, data was scanned for suspicious response patterns, and items P1 to P5 were reversed-scored – since they refer to *controlled motivation,* and thus expected to negatively load on the second-order motivation construct. Missing data was below the 5% threshold for every indicator (i.e., item), under which circumstances PLS-SEM is robust (Hair et al., 2017). *SmartPLS* 3.2 (Ringle et al., 2015) was used to calculate Cronbach's **α** , composite reliability and outer loadings (factor weighting scheme, with 300 iterations and stop criterion of $1*10^{-7}$) using a *Hierarchical Component Model* (reflective-formative) for each of the modules, with the repeated-indicator approach (Hair et al., 2017; Hair Jr. et al., 2018).

For interpretation, we followed Hair's et al. (2017) advice of using both **α** and composite reliability – as lower bound and upper bound estimates of reliability, respectively. **α** was deemed acceptable at .70 (P. Kline, 2000; Nunnaly & Bernstein, 1994), while CR was deemed acceptable at .60 (Hair et al., 2017). As for indicator reliability (outer loadings) values of .70 were deemed acceptable (Hair et al., 2017).

Cognitive Module

In order to gather preliminary evidence on construct validity for items in the cognitive module, we analyzed item's difficulty index, discrimination index, and performed a distractor analysis (Considine et al., 2005; Waltz et al., 2010) under the Classical Test Theory framework.

We had missing data for one student who did not complete this module. Item were scored using the *CTT* package (Willse, 2018) in *RStudio* (RStudio Team, 2020) with *R* version 4.0.2 (R Core Team, 2020); we used dichotomous scoring (i.e., 0 and 1) for correct answers – multiple selection items were considered correct if all correct options were selected. Difficulty and discrimination (*gULI*) indexes calculation, and distractor analysis (proportion of responses in each distractor) were calculated with the *shinyItemAnalysis* package (Martinková & Drabinová, 2019).

Item discrimination was interpreted according to cut-offs of *Very good* (>.40) ; *Reasonably good* (.30-.39); *Marginal* (.20-.29) , *Poor* (<.19) (Ebel & Frisbie, 1991; Lord, 1952). Distractors with lower than 10% of responses were considered poor functioning, to impose a stricter quality standard, although a lenient threshold of 5% is usually proposed (Towns, 2014).

Results

The following sections are organized chronologically, as to provide the reader with a detailed view of the different development phases and refinements that the PPLA-Q went through. In the Discussion section, we summarize and discuss these results according to their overarching goal (e.g., content validity).

Domain identification

Psychological Domain

Motivation

Self-Determination Theory (SDT) (Deci & Ryan, 2000) has abundant research in exercise and physical activity contexts (Teixeira et al., 2012). One of its mini-theories, Organismic Integration theory (Ryan & Deci, 2002), posits a continuum of different behavioral regulations varying according to their degree of self-determination. Among these, *external* and *introjected* are posited as more *controlled* (i.e., less autonomous) forms of extrinsic motivation; while *identified*, *integrated* and *intrinsic* are posited as more *autonomous* forms of motivation. More autonomous forms have shown positive association with increased participation in PA (Cortis et al., 2017), and with positive experiences in PE (Vasconcellos et al., 2019). We placed controlled forms of motivation in the foundational level, and more autonomous forms into the mastery level – following a two factor structure proposed in previous research (Gagné et al., 2010).

Confidence

Multiple self-concept constructs in the literature center around the belief in one's abilities to perform in PA settings; of these, (perceived) competence and self-efficacy seem to be determinants of participation in PA in children and adolescents (Babic et al., 2014; Cortis et al., 2017). Although conceptualized under different frameworks – perceived competence in the SDT tradition (as a basic psychological need driving motivation), and self-efficacy as the main construct of Social-Cognitive Theory (SCT) (Bandura, 1986) – studies have called for their integration, since they stem from the same concept of human agency (Sweet et al., 2012), and might share a common core (Hughes et al., 2011). As such, we integrated perceived competence – given its centrality to SDT, and similarity to task self-efficacy – in the foundation level, and barrier self-efficacy (Bandura, 1997) (i.e., belief in one's ability under challenging conditions) in the mastery level.

Emotional Regulation

Self-regulation is a broad concept that entails the individual's capacity to override and alter their behavior towards a standard or goal (Baumeister & Vohs, 2007). When referring to the affective domain, the construct of Emotional intelligence (i.e., ability to perceive and regulate emotion) (Goleman, 2005; Zeidner et al., 2012) has gained visible traction in research. It has been linked to PA participation, both as an outcome and as predictor (Ubago-Jiménez et al., 2019). Among its many conceptualizations we chose to adapt Wong and Law's Emotional Intelligence Scale's factorial approach(Wong & Law, 2002), mapping emotional evaluation (own and interpersonal) to the foundation level, and use and regulation of emotions to the mastery level.

Physical Regulation

Although we failed to identify a PA-specific construct that dealt with APLF's idea of regulating physiological signals and effort during PA– analogous to emotional regulation - we found it related to other affective constructs such as activity pacing (i.e., regulation of activity level towards an adaptive goal) (Nielson et al., 2013) and coping (i.e., behavioral and cognitive efforts to manage internal and external demands during stressful situations) (Lazarus & Folkman, 1984). The latter has been researched mainly in performance-oriented settings, and has showed positive association with sport commitment in adolescents (Pons et al., 2018). As such, we integrated this concept in an identical structure to that of Emotional Regulation: perception of changes in the body during exercise in the foundational level; and regulation of effort in the mastery level.

Social Domain

Culture & Society

The Culture & Society element is defined in the APLF as the appreciation of values present within communities of PA practice, however, we argue that its operationalization deals with cultural tolerance and cultural intelligence (Earley & Ang, 2003), rather than with the specific participation and appreciation of the cultural phenomenon of sport and PA. As such, we based this construct on Siedentop's call for symbolic attributes like values, rituals and traditions to be an integral part of PL (Siedentop, 1998). This ritualist facet manifests through the use of specific attire, jargon, and participation in select behaviors and habits (Mazurkiewicz, 2011); as well as through displays of fandom and sport fan passion (Vallerand et al., 2006). All these further contribute to feelings of affiliation and membership in a collective identity (Eastman & Riggs, 1994); and although literature linking this phenomena to participation in PA is sparse, it is plausible that it might play a mediator role in increasing perceived relatedness (Wallhead et al., 2013), and emotional regulation – particularly in anxiety-inducing settings (Brooks et al., 2016). We chose to map participation in cultural behaviors to the foundational level, while the mastery level represents a more involved stance in these (i.e., valuing and encouraging participation).

Ethics

Fair play, is an integral part of modern sport as its major ethical system – coherent with universal values (Bronikowska et al., 2019; Simon et al., 2015). PE plays a critical role in teaching this "inner morality of sport", which surpasses simple adherence to rules, and includes following unwritten rules and moral codes (Simon et al., 2015). Interiorization of these moral codes are concomitant with mature stages of moral development, which are known antecedents of prosocial behavior (Gibbs, 2014) (i.e., acts involving care for welfare of others) (Turiel, 2015), and might also increase intrinsic motivation in PA settings (Hassandra et al., 2003, 2007). We chose to use Gibbs' (Gibbs, 2014) model of moral development which, based on Kohlberg's work (Kohlberg, 1964), identifies two main levels in standard moral development: immature (i.e., a pragmatic, instrumental sense of morality, mapped to the foundational levels) and mature (i.e., based on social values and empathy, mapped to the mastery levels).

Collaboration

Personal and social responsibility are the main focus of Hellison's (Hellison, 2011) Teaching Personal and Social Responsibility (TPSR) model for developing prosocial behavior, providing a way to address holistic development of students in PE, and enable them with life skills for active citizenship through five levels: (1) Respect for the rights and feeling of others, (2) Effort and cooperation, (3) Self-direction, (4) Caring and helping others, (5) Transfer outside the gym. Evidence shows association of its application with many positive emotional, psychological, and social outcomes (e.g., self-efficacy, self-regulation, caring, conflict resolution) (Pozo et al., 2018). It is also suggested that students' level of personal and social responsibility are associated with intrinsic motivation in PE (W. Li et al., 2008). To avoid overlap between personal responsibility and other elements tapping into similar concepts (i.e., Ethics, Emotional and Physical regulation), we mapped TPSR's "Respect" level into the foundational level, and "Caring and Helping" into the mastery one, based on the works of Li's et al. (W. Li et al., 2008) .

Relationships

Relatedness (i.e., perceived connection with others) is another one of the basic psychological needs posited to drive motivation according to SDT. Despite its theoretical relevance, evidence has shown little to no direct association between relatedness and participation in PA, in both general (Cortis et al., 2017; Teixeira et al., 2012) and PE contextes (Taylor et al., 2010). However, some authors (Cox et al., 2009; Teixeira et al., 2012) suggest that this might be due to relatedness being highly context-dependent (i.e., affected by prevalence of solitary exercise, or lack of connection with classmates), and thus, not captured in its entirety in the researched contexts. This idea is further reinforced by evidence of peer-support associating with PA practice (Martins et al., 2015), positive outcomes in PE (Vasconcellos et al., 2019), and as mediator in other relevant outcomes as effort (Leptokaridou et al., 2015) and enjoyment (Cox et al., 2009). In our model, akin to Collaboration, we mapped a reactive role in relationships to the foundational level, while the mastery level presupposes an active role in relationship development.

Cognitive Domain

Content Knowledge

Few studies have examined the relationship between knowledge regarding PA, and outcomes in PE contexts (either affective, social, or behavioral). However, there is evidence of positive association of knowledge of PA guidelines (World Health Organization, 2010) and health benefits, both with PA participation in young adults (Abula et al., 2018; Haase et al., 2004), and physical fitness (Vaara et al., 2019). Similarly, awareness of health risks related to inactivity might predict PA participation in adults (Fredriksson et al., 2018) and adolescents (Xu et al., 2017). A consensus among aforementioned studies seems to be that knowledge of these contents is consistently low, with similar evidence in Portugal: both in PE setting (Marques et al., 2015) and in young adults (Martins, Cabral, et al., 2019).

Content Validity

Version one (v0.1): Cognitive Interviews

All students (n=4) referred to the questionnaire as having an adequate layout and length, as well as clear directions for filling in the questionnaire. Their understating of item stems and rating scales, in the psychological and social modules, matched our intention: with equivalent conceptual distance between rating scale options. The response options in the cognitive module were deemed intuitive, given their familiarity with multiple-choice items. Item content was mostly clear for all students, with some difficulties arising in discerning the meaning of many items in the Culture & Society scale; they suggested adding examples to clarify concepts like “cultural diversity” and “traditional physical activities”.

We found a quality issue with the cognitive module item C6 (i.e., doping’s impact on health and fair play): During think-aloud response, it became evident that students could extrapolate the correct response without pertinent knowledge, due to the implausibility of distractors. According to students’ comments changes were made to the questionnaire: we added examples for mentioned concepts and improved the plausibility of C6’s distractors.

Version two (v0.2): Expert evaluation – 1st round

To quantitatively assess the *relevance* and *clarity* of each item, a panel of subject matter experts were asked to rate each item on a 4-point Likert-type scale. 10 experts in total participated in this round, of these 6,5 and 4 experts rated the cognitive, psychological, and social modules, respectively. Based on their ratings, CVI (I-CVI, S-CVI/Ave, and SCI/UA) and **κ** were calculated.

Item relevance

Item CVI ranged from .33 to 1 (cf. Additional File 2): 1 item had a CVI of .33, 3 items had a CVI of .5, 10 items had a CVI of .75, 6 items had a CVI of .8 and the remaining 70 items a CVI of 1.

Κ ranged from .13 to 1, with 86 items (96%) considered either excellent (76 items) or good (10 items) (Table 5); four items were prompted for elimination – one in the cognitive module (C2, *Nutrition*) and three in the social module (S3 in *Culture & Society* scale, and S30 and S34 in Relationships scale) (Table 5).

Scale relevance

The psychological and social modules showed adequate content validity, with a S-CVI/Ave of .98 and .90 respectively (Waltz et al., 2010); while the cognitive module failed to reach the proposed adequacy threshold of .90, with an S-CVI/Ave of .87.

According to the S-CVI/UA, only the psychological module showed adequate content validity, with a value of .93 (higher than the .80 threshold) (Davis, 1992), while the cognitive and social modules did not - .60, and .68 respectively.

Item Clarity

Proportion of agreement ranged from .33 to 1 (cf. Additional File 2): 2 items had an index of .33, 2 items had an index of .50, 2 had an index of .67, 8 items had an index of .75, 7 items had an index of .80, and the remaining 69 items an index of 1.

K for clarity ranged from -.07 to 1 (Additional File 2) with 76 items (84%) considered clear and 14 items prompted for revision – the *Culture & Society* scale had the greatest number of items needing revision, followed by *Collaboration* and *Relationships*, all in the social module.

Table 5. Number of items, per scale, in each kappa category of relevance and clarity in result of expert evaluation (version 0.2 and 0.3)

	Relevance				Clarity	
	Kappa[1]			S-CVI	Kappa[4]	
Module **Element (items)**	**Elimination**	**Good**	**Excellent**	**(Ave[2]/UA[3])**	**Revision**	**Clear**
1st round (version 0.2)						
Cognitive						
Nutrition (C1 & C2)	1	—	1	.87 / .60	1	1
Fitness and training (C3 & C4)	—	—	2			2
Safety and risk (C5 & C6)	—	—	2		1	1
Health benefits of PA (C7 & C8)	—	—	2		1	1
Body composition (C9 & C10)	—	—	2		—	2
Psychological						
Motivation (P1-P9, P37)	—	—	10	.98 / .93	1	9
Confidence (P10-P18, P38)	—	—	10		—	10
Emotional Regulation (P19-P27, P39)	—	—	10		—	10
Physical Regulation (P28-P36, P40)	—	—	10		1	9
Social						
Culture & Society (S1-S9, S37)	1	4	5	.90 / .68	4	6
Ethics (S10-S18, S38)	—	—	10		—	10
Collaboration (S19-S27, S39)	—	4	6		3	7
Relationships (S28-S36, S40)	2	1	7		2	8
2nd round (version 0.3)						
Social						
Culture & Society	—	—	10	.96* / .84*	5	5

Table 5. Number of items, per scale, in each kappa category of relevance and clarity in result of expert evaluation (version 0.2 and 0.3)

	Relevance				Clarity	
	Kappa[1]			S-CVI (Ave[2]/UA[3])	Kappa[4]	
Module Element (items)	Elimination	Good	Excellent		Revision	Clear

[1] Multirater modified kappa designating agreement on relevance: κ=(I-CVI - pc)/(1 -pc), with pc (probability of a chance occurrence) computed using the formula for a binomial random variable, with one specific outcome(Polit et al., 2007);evaluation criteria for kappa(Cicchetti & Sparrow, 1981; Fleiss, 1971): Elimination <.40, Fair kappa of .40 to .59; Good kappa .60 to .74; and Excellent kappa > .74.

[2] S-SCI/ Ave - Scale CVI Average: Calculated by averaging all I-CVI in scale/module.

[3] S-SCI/ UA - Scale CVI Universal Agreement: Calculated by dividing the sum of items with I-CVI of 1.0 by module's total number of items.

[4] Modified criteria for kappa: Needs Revision < .74; Clear > .74.

*Calculation included all scales of social module.

Questionnaire refinement

Based on both qualitative and quantitative evidence from experts, two items were eliminated from the Relationships scale. It also prompted a major revision of the Culture & Society scale to increase S-CVI/Ave and S-CVI/UA of the social module to acceptable levels – informed by consultation with a subject matter expert, and one of APLF's authors.

The cognitive module underwent restructuration as most experts commented on quality issues regarding (1) implausibility of distractors, (2) syntax and (3) structure. None of the items were eliminated, as it would compromise content representation, and the two-level framework of the module. Albeit not reaching the desired threshold for S-CVI (Ave and UA), we chose not to submit the cognitive module to a formal second round of expert evaluation, given that all **κ**'s (relevance) were excellent (>.74), save from item C2. Alternatively, we consulted with an assessment expert to restructure item C2 and improve the clarity on items C6 and C8, with no changes content-wise.

Version three (v0.3): 2nd round results for Culture & Society element

We asked 3 experts to participate in a second round of evaluation of *Culture & Society* scale, given the depth of its restructuration. Same procedures and calculations applied from the 1st round.

All items in the revised *Culture & Society* scale obtained a I-CVI and **κ** of 1, indicating absolute agreement on item's relevance (Additional File 2). As such, S-CVI/UA of the social module increased to .84, entering an acceptable range (Davis, 1992).

Proportion of agreement on clarity ranged from .33 to 1 (Additional File 2): 1 item with .33, 4 items with .67 and the remaining 4 with 1; **κ** ranged from -.07 to 1, with 5 items considered clear and 5 prompted for revision.

Questionnaire refinement

5 items in the Culture & Society scale – with clarity κ lower than .74 – were revised (S3 – S5, S7 and S9), and S6 was eliminated, since expert's comments pointed to it being more representative of general cultural tolerance than adherence to sport's culture.

Version four (v0.4): Pilot Testing & Cognitive Interviews

Feasibility

Of the 58 students who got the informed consent, 41 completed the PPLA-Q, resulting in an attrition rate of about 30%. These 41 students (71% female) studied in grade 10 of two different schools, with two different majors (19 students from a Science, Technology, Engineering and Math course, 22 from Humanistic course), mean age 15 (0.4) years.

Completion time was gathered to assess the questionnaire's feasibility during PE classes. Average completion time was 27 (7) minutes (n=34, with the remaining 7 students failing to fill in the beginning and ending time). Questionnaire application in the gym allowed for ample space between students, which restricted talking; however, application in a crowded classroom promoted student's sharing ideas about the items, and their correct option(s) (in the cognitive module). No response errors or any suspicious response patterns were identified on the responses (e.g., straight or diagonal lining, or alternating poles; Hair et al., 2017).

Preliminary reliability (Psychological and Social modules)

Preliminary reliability for each subscale, as well as each item's outer loading (indicator reliability) on its intended construct are summarized in Table 6.

10 of the subscales (63%) attained acceptable reliability according to both **α** and CR (>= .70, and >=.60, respectively); 2 subscales only attained acceptable values in the

upper bound estimate (i.e., composite reliability). Out of the remaining 6, the *Ethics* element had the lowest reliability on its two subscales. We noticed a discrepancy in **α** and CR's expected behavior (i.e., **α** lower than composite reliability) in the *Motivation* foundation, and *Physical Regulation* foundation subscales.

We found that 42 items (56%) had acceptable individual item reliability (outer loading >.70). 11 items had unexpected negative loadings - as they were intended to relate positively with their constructs; these were, however, mostly negatively worded items found in *Motivation*, *Physical Regulation* and *Ethics* foundation level subscales.

Table 6. Preliminary item and subscale reliability of Psychological and Social modules (n=41; PPLA-Q version 0.4)

Psychological Module				Social Module			
Element (Subscale)	**Item**	**Outer Loading**	**Subscale Reliability[1]**	**Element (Subscale)**	**Item**	**Outer Loading**	**Subscale Reliability[1]**
Motivation (Foundation)	P1	.32	**.76**/ .22	**Culture (Foundation)**	S1	.69	.66/ **.79**
	P2	-.81			S2	.79	
	P3	-.77			S3	.34	
	P4	.37			S4	.91	
	P5	-.10					
Motivation (Mastery)	P6	.85	**.87/ .91**	**Culture (Mastery)**	S5	.88	**.86/ .90**
	P7	.86			S6	.87	
	P8	.88			S7	.81	
	P9	.80			S8	.89	
	P37	.67			S34	.54	
Confidence (Foundation)	P10	.88	**.93/ .95**	**Ethics (Foundation)**	S9	-.14	.57/ .59
	P11	.87			S10	-.95	
	P12	.92			S11	-.88	
	P13	.91			S12	-.28	
	P14	.86			S13	.11	
Confidence (Mastery)	P15	.81	**.70/ .80**	**Ethics (Mastery)**	S14	.27	.36/ .53
	P16	.84			S15	.81	
	P17	.21			S16	-.51	
	P18	.81			S17	.72	
	P38	.59			S35	.61	
Emotional Regulation (Foundation)	P19	.70	**.75/ .76**	**Collaboration (Foundation)**	S18	.77	**.81/ .87**
	P20	.66			S19	.91	
	P21	.44			S20	.58	
	P22	.74			S21	.69	
	P23	.55			S22	.80	
Emotional Regulation (Mastery)	P24	.68	.61/ **.78**	**Collaboration (Mastery)**	S23	.83	**.80/ .87**
	P25	.75			S24	.84	
	P26	.65			S25	.88	
	P27	.85			S26	.80	
	P39	.21			S36	.35	
Physical Regulation (Foundation)	P28	-.28	.62/ .17	**Relationships (Foundation)**	S27	.73	**.85/ .90**
	P29	-.26			S28	.79	
	P30	-.14			S29	.90	
	P31	.78			S30	.91	
	P32	.76					
Physical Regulation (Mastery)	P33	.80	**.76/ .84**	**Relationships (Mastery)**	S31	.91	**.74/ .91**
	P34	.38			S32	.68	
	P35	.85			S33	.79	
	P36	.83			S37	.60	
	P40	.69					

[1]Statistics presented: Cronbach's **α** / Composite Reliability.
Note: Results higher than .70 (outer loading and **α**) and .60 (composite reliability) are bolded (acceptability threshold).

Item analysis (Cognitive module)

Table 7 summarizes the preliminary item analysis of the cognitive module of the PPLA-Q. We found a mismatch between intended complexity of the item and its difficulty in 2 of the 5 content groups (i.e., foundational items being answered incorrectly more often that mastery items for the same content); as well as an overall low success in foundational items. Additionally, average difficulty of the items in the module was .50, representing a more difficult test than ideal for maximizing discrimination – .70 to .74, for a test with four, five and six options multiple-choice items (Lord, 1952). Notwithstanding, 6 items showed good or very good discrimination between lower-knowledge and higher-knowledge students (D > .30). Distractor analysis revealed that 16 (57%) were low functioning distractors (i.e., ≤ 10% of total responses for the item); these were mostly in easier items.

Table 7. Difficulty, discrimination, and distractor analysis of items in the Cognitive module (n=40; PPLA-Q version 0.4)

						Distractor analysis (%)[4]					
Content	**Level**	**Item**	***p***[1]	***D***[2]	**Evaluation**[3]	**Response Option**					
						a	**b**	**c**	**d**	**e**	**f**
Nutrition	Foundation	C1	.95	.08	- -	0	95	5	0	—	—
	Mastery	C2	.78	.31	+	3	5	13	80	—	—
Fitness and training	Foundation	C3	.45	.54	++	48	45	5	3	—	—
	Mastery	C4	.32	.77	++	33	10	15	43	—	—
Safety and risk	Foundation	C5*	.40	.54	++	98	0	80	15	90	45
	Mastery	C6	.82	.31	+	3	3	85	10	—	—
PA's Health Benefits	Foundation	C7	.32	.46	++	33	20	15	33	—	—
	Mastery	C8	.80	.23	-	10	8	3	80	—	—
Body Composition	Foundation	C9	.15	.23	-	15	18	53	15	—	—
	Mastery	C10*	.10	.23	-	43	13	90	5	53	—
		Average *p*	.50								

* Multiple selection items ("choose all that apply").

[1] p - Difficulty index: number of correct responses / total number of responses – higher number means easier item.

[2] D - Discrimination index (generalized ULI): difference in ratio of correct answers in upper and lower third of students.

[3] Evaluation cutoffs for discrimination index(Ebel & Frisbie, 1991): >.40 Very good (++); .30-.39 Reasonably good (+); .20-.29 Marginal (-), <.19 Poor (- -).

[4] Percentage of students choosing option – correct options are bolded.

Cognitive Interviews – 2nd round

Further individual cognitive interviews (n=2) were conducted to probe student's understating of changes made to the items in the last 3 versions of the PPLA-Q, as well as in items which raised frequent requests for clarification during pilot application. Interviewed students showed good comprehension of the items. Additionally, a minor change was suggested in one of the distractors of the item pertaining to basic safety procedures during PA (C5): substitute "Always drink water" for "Drink water regularly".

Questionnaire refinement

Results of preliminary reliability analysis prompted a detailed analysis of every item and subscale in the psychological and social modules. Based on this, negatively stated items were changed into positively stated ones to improve comprehension, and subsequently, validity and reliability. Minor changes were made to item stems as well, to improve clarity.

Additionally, 11 global assessment items (e.g., "I'm motivated to practice PA") were introduced into the psychological and social modules to allow for convergent validity assessment through redundancy analysis of the second, third and fourth-order formative constructs (Cheah et al., 2018; Hair et al., 2017) in further stages of PPLA-Q development. Of these, 8 targeted each of the elements, 2 targeted the general psychological and social domains, and 1 targeted general PL. Their content followed the respective operational definition stated by the APLF (see Table 3), while adhering to the same structure and rating scale as the remaining items.

Informed by the results of the preliminary item analysis, items in the cognitive module were revised to better conform to the expected difficulty levels (i.e., mastery items harder than foundation ones). We revised low functioning distractors, to make them more plausible to students. C6 was modified from a single selection multiple-choice to *cloze-type* item, with no changes to intended outcome. All revisions in this module were made in consultation with a subject matter expert to ensure technical adequacy and content validity.

Before the next iteration of cognitive interviews, all items were co-validated by non-generating authors to guarantee that clarity was improved, and content validity was left unchanged.

Version five (v0.5): Cognitive Interviews

To assess the clarity of items changed between version four (v0.4) and five (v0.5) of the PPLA-Q, 6 students were interviewed. Most Likert-type items were clear and coincident with their intended meaning, except those regarding justice (e.g., "I try to be just"), which led to interpretations related with collaboration and teamwork, instead of the intended meaning regarding ethics/fair-play. In the revised cloze-type item in the cognitive module (C6), one of the students failed to respond to the item according to instructions (i.e., filled in the spaces, instead of circling options that would fill each space), revealing a need to clarify its instructions. According to this data, we fine-tuned all pertaining items. Similarly, informed by the pilot test, we created two different versions of the cognitive module by mirroring the arrangement of options – in the second version, A became D, B became C – hoping to discourage students to share their answers during application and reduce subsequent measurement error.

Discussion

This article followed the development, content validation, and pilot testing of the first of two instruments that comprise the *Portuguese Physical Literacy Assessment* (PPLA)*: the PPLA-Questionnaire (PPLA-Q)*, it assesses the *psychological*, *social* and part of the *cognitive* domains of PL, inspired by the *Australian Physical Literacy Framework* (APLF) and Portuguese PE syllabus (PPES). Its primary target are high-school students (grade 10 through 12) in PE context. Older adolescents are a critical intervention group – especially in Portugal – given that they possess lower PA levels than their younger peers (Baptista et al., 2012; Martins, Marques, et al., 2019; Matos & Equipa Aventura Social, 2018); and will cease to have mandatory and free access to professional guidance in PA and movement, eventually becoming dependent on their PL to participate in meaningful PA, and further advance on their journey.

Content Validity

We gathered evidence on content validity using an iterative process with experts in each subject matter domain (i.e., for each of the modules), and target population. The number of experts per module was considered acceptable and ranged from 4 to 6. Although literature recommends 5-7 experts to rate content validity (Boateng et al., 2018), a minimum of 3 is acceptable for content areas in which expert recruitment might prove difficult (Lynn, 1986) – as we argue was the case in this study, given constraints imposed by the COVID-19 pandemic.

PPLA-Q showed evidence for adequate content validity at item level improved throughout multiple revisions. In version 0.2, using κ, 96% of the items were rated as good or excellent (>.74) (Cicchetti & Sparrow, 1981; Fleiss, 1971) regarding relevance, and 84% considered clear (>.74). Module-wise, a S-CVI/Ave of .90 is considered adequate (Waltz et al., 2010), a cut-off that decreases to .80 for S-

CVI/UA, given that it requires universal agreement between all raters (Davis, 1992). While the psychological module attained an adequate S-CVI on both accounts (.98/.93), the social module did so only on S-CVI/Ave (.90 /.68), and the cognitive failed to achieve both standards (.87/.60); further analysis identified that most items with lower I-CVI in the two latter modules were those generated without a conceptual reference to an existing instrument (i.e., *Culture & Society*). Qualitative suggestions from the experts and advisors augmented quantitative data, targeting concepts in need of rewording or clarification. We then revised and eliminated items to improve content validity across all modules. A targeted revision of the *Culture & Society* scale increased overall social module's S-CVI to .90/.84 (Ave/UA) on a second round of expert evaluation aimed solely at it (version 0.3).

Multiple rounds of qualitative cognitive interviews were conducted until saturation was achieved (i.e., no new suggestions emerged) (I. B. Rodrigues et al., 2017) with a heterogenous sample of high-school students (n=12), using different versions of the PPLA-Q. These informed improvement on item wording and syntax, to effectively target the intended concepts and reduce ambiguity. During initial stages, students noted lack of clarity in abstract concepts like those from the *Culture & Society* (values, rituals, and traditions of sport/PA), and *Ethics* (justice, honesty, fair play) scales; notwithstanding evidence that iterative revisions clarified these items, further validation efforts should scrutinize their performance. Similarly, despite obtaining evidence for item-level content validity (except for item C2), and subsequent reviews in consultation with a test and assessment expert – based on the qualitative comments of experts and students – we advise further quantitative scrutiny of the cognitive module to establish its module-wise content validity.

Feasibility

Average completion time for the PPLA-Q was 27 minutes. Although it might impose a substantial burden upon respondents, diversity of constructs and items used throughout the different modules might have effectively reduced it. Notwithstanding, depuration of subscales in the modules – during the next steps in development – will certainly reduce this time and further improve feasibility.

We had no response errors and low levels of missing data during pilot testing, which might stem from student's routine exposure to questionnaires using the same item format (i.e., multiple-choice items and Likert-type scales). We also argue that application of the questionnaire during PE class, with the lead investigator present, to clarify any question, might have played a determinant role in this. In one of the application settings (i.e., classroom) it was notorious the student's urge to copy or share their answers from/with colleagues, especially in the cognitive module. The similarity of this module with usual summative evaluation instruments used in school setting might partially explain this occurrence; non the less, we expect that future use of the two differently arranged versions of the cognitive module (i.e., mirrored distractors) might reduce this.

We experienced a high rate of attrition (≈30%). Constraints imposed by the COVID-19 pandemic might have reduced the number of students completing the questionnaire: both by reducing their willingness to participate, as well as the possibility to be present during application (due to prophylactic lockdown). This number shall inform the sample size calculations in further phases of development, as it is expected that these conditions might endure during next phases.

Preliminary reliability and item analysis

Results of reliability analysis in the psychological and social modules established preliminary evidence of adequate reliability in 10 out of 16 subscales (**α** >.70 and composite reliability > .60) (Hair et al., 2017; Nunnaly & Bernstein, 1994). Analysis of item reliability highlighted items that were contributing negatively to subscale reliability (outer loading <.70) of the remaining 6 subscales: Upon careful inspection, most of these were negatively worded. Although the use of negative wording might filter out unwarranted responding patterns (e.g. acquiescence), they have the potential to confuse students and compromise validity and reliability (DeVellis, 2017) by, for example, creating an artificial subconstruct within the intended subscale (Sonderen et al., 2013). As such, these items were altered and then tested for comprehension during subsequent cognitive interviews, with positive results. Further reliability testing is warranted with a bigger sample size, to gather more definite evidence on the reliability of these subscales.

Regarding item analysis of the cognitive module, item difficulty ranged from .10 (very hard) to .95 (very easy), with an average difficulty of .50. Initial evaluation of its 10 items identified 6 good or very good discriminating items (D > .30) (Ebel & Frisbie, 1991; Lord, 1952) (i.e., capable of differentiating knowledge levels among students).

We expected items designed for in the mastery level (i.e., pertinent to deeper learning) to be more difficult than those in the foundation level, within the same content; however, pilot data does not fully support this idea, as foundational items were more difficult than their mastery counterpart in 2 content pairs (C5 & C6, C7 & C8). We identified low-functioning distractors in the mastery level's C6 & C8 (non-plausible), that increased likelihood of a correct answer, even without full knowledge of the content. Conversely, C5 and C7 (foundation) had characteristics which inflated its difficulty: one of C5's (multiple selection item about safety during PA) intended "correct" options contained absolute language ("[one should] hydrate during *all the duration* of the activity"), steering respondents away from it; while C7 measured factual knowledge of the recommendations for PA in children and adults, which has been previously shown to be low among adolescents (Marques et al., 2015) and young adults (Haase et al., 2004; Martins, Cabral, et al., 2019). A similar phenomenon emerged with C9, which asked respondents to select the Body Mass Index calculation formula – although students might be familiar with the concept, they might not recall its formula. Informed by this data, distractors were thoroughly revised.

We would like to acknowledge, that although the methods used here to preliminarily assess the quality of the items followed the *Classical Test Theory* framework , *Item Response Theory* and Rasch models might play a role in further validation efforts, since they expressly integrate the notion of item difficulty (as well, as other possible parameters like discrimination and guessing) into the calculation of student's scores (Andrich & Marais, 2019); this would allow precise student scoring along the learning continuum posited in the development of PPLA. These were not used in this pilot study, given their requirement of larger sample sizes (Haladyna, 2004).

PPLA as a whole is intended to assess the integrated physical, cognitive, psychological, and social variables that are posited to underpin PL; both to direct the pedagogical action at local, regional and national level in proving a PL-supporting environment, and to inform self-directed changes by the students. Even though it pertains to attitudes, skills and knowledge applied in general PA settings, further adaptation is warranted if it is to be applied to younger students and/or outside of PE. Moreover, we argue that although culture might play a defining role in the representation of PL – as stated by Whitehead (Whitehead, 2010) – and that the PPLA-Q was designed with this peculiarity in mind, most of its indicators (i.e., items) might be easily adapted to other cultural contexts.

Strengths and Limitations

To our knowledge, this study is the first report of content validity for a measurement instrument of PL designed for grade 10 to 12 adolescents. The content in the tool was inspired by the APLF and the PPES, informed by previous decisions of consortium of experts during a European project (PhyLit). Its development used an iterative process of content validation, using both subject matter experts in each knowledge domain (i.e., cognitive, psychological, and social), as well as target population, resulting in many revisions to improve its clarity and validity.

Although great care was taken to create a heterogeneous sample for the cognitive interviews and pilot test, all participants were nonetheless from a convenience sample from Lisbon's metropolitan area. Similarly, we could not reach our goal of 6 experts participating in every module. The effects of the COVID-19 pandemic might have had an overarching effect on expert availability to participate in the project, and students' participation rate – through previously discussed constraints. However, we did not collect enough information to extrapolate specific causes for attrition, which could provide additional insights to prepare future studies and further improve feasibility.

Given that only preliminary testing was done regarding reliability and construct validity, further work is warranted and is currently ongoing to establish evidence in this regard, with a statistically adequate sample size.

PPLA inherits the complex nomological network of APLF, as such, some theoretical constructs underwent adjustments in other to be fully integrated into the same model; as such, further robust construct validation needs to ensure adequate dimensionality of each construct chosen, as well as the accuracy, validity, and practical usefulness of the usage of the learning continuum posited through the *foundation* and *mastery* levels. Further studies should also evaluate PPLA-Q's integration with PPLA-O (in development) to provide a holistic, integrated assessment, as warranted.

Similarly, this effort might allow for depuration of the instrument, contributing to a more parsimonious and shorter version; further improving its feasibility in PE contexts. As the PPLA-Q only targets older adolescents now, future adaptation into earlier age ranges might provide a clearer picture of PL development throughout all school-age.

Conclusion

This study details the iterative development process of the PPLA-Q as an instrument to assess the psychological, social, and part of the cognitive domain of PL in grade 10 to 12 adolescents (15-18 years). It also provides evidence for adequate content validity at item level, and, except for the cognitive module, at module level. It was improved through multiple rounds of expert and target-population consultation. This instrument has also shown good feasibility within PE settings and gathered preliminary evidence in favor of its reliability for application in older adolescents. Further validation efforts are needed to reinforce these conclusions, establish evidence of construct validity, and study PPLA-Q's integration with the PPLA-O (an instrument in development to assess the remaining domains of PL) within the PPLA framework to provide feedback to support older adolescents in their PL journey.

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. All methods were performed in accordance with the relevant guidelines and regulations.

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

Supporting data is not available as participants of this study did not explicitly agree to share their data publicly. The latest development version of the questionnaire used - *Portuguese Physical Literacy Assessment Questionnaire* (version 0.6) - is available in Additional File 3, both in its original Portuguese version, and translated to English for reader's convenience; interested readers might procure an updated version with the lead author.

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, all the experts which participated in the content validation, and of all the tireless PE teachers and students. Secondly, we would like to thank the HALO Research Group for their work on the Canadian Assessment of Physical Literacy which inspired this, and probably others' effort to measure Physical Literacy. This work could not be completed without the seminal work of the Australian Physical Literacy Framework team, and the genius vision of Margaret Whitehead. The lead author would also like to thank his co-authors for their constant support and inspiration throughout the whole project, and the reviewers of the initial manuscript for their minute attention and professionalism.

CHAPTER 3 – PPLA-Q Psychological and Social Modules Construct Validity and Reliability

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, João Martins, Marcos Onofre

Submitted to Assessment [IF: 4.67, Q1] in shortened form

Preprint available at: https://doi.org/10.21203/rs.3.rs-1458709/v3

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 in 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 .83 to .94); regarding test-retest reliability: 3 scales had good to excellent reliability ($ICC_{95\%CI}$ ranging from .72 to .95), and 5 scales presented moderate to good reliability ($ICC_{95\%CI}$ ranging from .51 to .85). 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 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.

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

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 the endeavor to track and understand progress (Corbin, 2016b).

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, in development) developed for use in PE to provide a feasible and integrated 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 (W. 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 of measurement invariance across sexes at item and scale-level (Gamerman et al., 2019; Moorer et al., 2001; Teresi et al., 2008). Similarly, test-retest reliability is crucial in distinguishing random short-term scores differences from true change (Polit, 2014), allowing reliable tracking of learning in these elements over time.

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 (baseline)

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 prior 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 Math (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 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 Mokken Scale Analysis (MSA) (Mokkink et al., 2018; Straat et al., 2014).

Retest study phase

A subsample of 73 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

Characteristic	**Baseline** **N = 508**[1]	**Retest** **N = 73**[1]
Sex		
Female	299 (59%)	41 (56%)
Male	209 (41%)	32 (44%)
Age	16 (1)	16 (1)
Grade		
10	204 (40%)	-
11	137 (27%)	73 (100%)
12	167 (33%)	-
Major		
Economics	75 (15%)	-
Humanities	165 (32%)	
STEM	268 (53%)	73 (100%)
School		
School 1	39 (7.7%)	-
School 2	61 (12%)	-
School 3	21 (4.1%)	-
School 4	69 (14%)	-
School 5	207 (41%)	73 (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 (general stem: "*How much do these statements describe you?*").

Procedures

Main study (baseline)

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 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 psychological and social modules, respectively.

Retest study phase

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 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 analyzed 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 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% to 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 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 of 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 followed the same cutoffs as those of Cronbach's *α* (Cronbach, 1951): with ρ > .70 considered as acceptable (Nunnaly & Bernstein, 1994) and ρ >.80 considered as good (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 at both time points (Liljequist et al., 2019). ICC values of .90, .75, .50 were used, respectively, as thresholds for excellent, good, and moderate test-retest reliability (Koo & Li, 2016).

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 of each variable pair 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 magnitudes 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.

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.

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
		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
		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 these 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
		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
		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 these statements describe you?*"; [R] Reverse-coded item

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

In the social module, 6 items were unscalable: 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). All 15 items were removed in a stepwise manner, ensuring that 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*), within the same specific trait (e.g., P9 and P11 with the same motivational regulation) or had similar wording. Within each pair, an item was chosen to be removed according to content coverage of 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 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 their 95%CI, 2 scales formed medium to strong (*Motivation*, and *Physical Regulation*), while the remaining 6 formed strong Mokken hierarchical scales (H lower bound > .50, 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 (Table 19), thus, such ordering might be too inaccurate for practical purposes (Ligtvoet et al., 2010) – students might perceive neighbor 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 Scaling

Motivation

We noticed a pattern of borderline $CI_{95\%}$ lower bound values for H_i in items P2 and P5 in the *Motivation scale*; and high scalability between items P3 and P4. 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, after removal of items flagged in local dependence pairs, they formed two strong Mokken scales (since Cluster 2 was composed of only two items, it was not considered) conforming with the DMM: Cluster 1 (H= .61, H^T = .50) and Cluster 3 (H= .60, H^T = .56).

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)			
		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

The clustering pattern observed in the *Emotional Regulation* scale in items P24, P26 and P27 led us to perform the exploratory 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 returned the same results, clustering these 3 items. The two clusters 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 *a priori* scale for this element, the Cluster 2 is formed by items assessing whether the student 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 its interpretability, this last solution was used for confirmatory analysis: 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 a PLD pair (Table 14).

Measurement invariance

To assess whether items presented differential item functioning (DIF) according to sex, we calculated 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 - 18, column 4) – P5 with higher scalability for males, and the others 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 scalability (H difference <.10; no 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).

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 like *rho*, and all above .80 as well (ranging from .83 to .91).

Test-retest reliability

The *Motivation*, *Confidence*, and *Culture* scales showed good to excellent reliability ($ICC_{95\%CI}$ lower bound ranging from .72 to .87, and upper bound from .89 to .95; Table 19, column, 6); while the remaining scales showed moderate to good reliability. Mean scores were stable across time points, with slight decreases in four of the eight scales (Table 19, columns 7 and 8).

Discriminant and convergent validity

Estimated disattenuated correlations among scales-scores within the Psychological domain ranged from .31 to .83 (Table 20Table 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. Estimated disattenuated correlations within the Social module ranged from .21 to .69 (Table 20Table 20). Of these, *Culture* was the lowest common correlate; with other scales correlating moderately to strongly. Correlations across domains were low, except for *Culture* and *Relationships*, which showed moderate disattenuated correlations with scale-scores in the Psychological domain.

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

	Dimensionality		**Reliability**				
			Test-score		**Test-Retest 15-day interval N = 73**		
Subscale – number of items	**Scalability H [95% CI]**	**Invariant Item Ordering H^T (item ordering)[1]**	**Molennar-Sijtsma ρ**	**Cronbach's α**	**ICC$_{2.1}$ [95% CI][2]**	**Mean Scores Baseline [95%CI]**	**Mean Scores Retest [95%CI]**
Psychological							
Motivation – 7 items	.47 [.43, .51]	.33 (P5, P8, P2, P1, P7, P43, P11)	.83	.83	.82 [.72, .89]	19.8 [18.7, 20.9]	19.0 [18.1, 20.0]
Confidence – 9 items	.66 [.62, .69]	.08 (P13, P17, P44, P16, P19, P14, P21, P22, P20)	.94	.93	.92 [.87, .95]	22.2 [20.6, 23.8]	22.2 [20.6, 23.8]
Emotional Regulation - 7 items	.58 [.53, .62]	.19 (P25, P28, P30, P31, P23, P32, P29)	.90	.88	.77 [.66, .85]	17.5 [16.3, 18.6]	17.6 [16.5, 18.7]
Physical Regulation – 8 items	.46 [.41, .50]	.41 (P35, P36, P38, P42, P37, P33, P39, P41)	.84	.84	.66 [.51, .77]	22.3 [21.2, 23.4]	21.7 [20.6, 22.7]
Social							
Culture – 7 items	.62 [.59, .66]	.32 (S6, S5, S3, S7, S2, S8, S9)	.91	.91	.88 [.82, .92]	14.0 [12.6, 15.5]	14.4 [12.8, 16.0]
Ethics – 6 items	.54 [.49, .59]	.52 (S13, S14, S17, S21, S22, S12)	.86	.85	.71 [.58, .81]	18.7 [18.0, 19.5]	19.1 [18.4, 19.8]
Collaboration – 6 items	.64 [.60, .69]	.26 (S26, S27, S23, S24, S28, S30)	.88	.87	.70 [.56, .80]	19.4 [18.6, 20.1]	18.8 [18.1, 19.5]
Relationships – 6 items	.61 [.57, .66]	.22 (S32, S33, S37, S36, S34, S38)	.88	.88	.68 [.53, .78]	17.0 [16.1, 17.9]	16.4 [15.5, 17.3]

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)

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*), within the same specific trait (e.g., P9 and P11 with the same motivational regulation) or 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 onwards) 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 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 the *Culture* scale 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), these global items figured in the middle-upper range of the difficulty continuum (i.e., lower mean score); this was to be expected, as these were based on the operational definition of each element (Mota et al., 2021) stating 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 pairs, except those in the *Motivation* and *Confidence* scales. This might be due to: 1) endorsement of these items being highly dependent on the capacity of the student 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 might represent a different skill altogether; 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 of the *Motivation*, *Emotional Regulation* and *Physical Regulation* scales – using the Automated Item Selection Procedures (AISP) and Genetic Algorithms (GA) at lower-bound c = .45) – 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, along with the clustering of autonomous regulations are coherent with previous results with adolescents (Navarro et al., 2021; Vasconcellos et al., 2019).

For the *Emotional Regulation* scale, clustering patterns suggested that 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, coherent with the *a priori* construction of *Foundation* and *Mastery* levels. Although these alternative dimensionality structures could be supported for these scales, we recommend their use as single scales within the PPLA framework as their total scalability coefficients evidence enough unidimensionality to locate an individual in each of these latent traits. Other research applications (e.g., theory development) might benefit from use and exploration of these alternative structures.

Regarding IIO, 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. Use of parametric IRT models might support this effort, although their restrictive assumptions regarding item response functions' shape might not fit those 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. Since 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 (relating to concepts of physical signs and fatigue during PA). As such, we advise caution on the interpretation and comparison of between-sexes score differences in this scale. Since previous literature in this construct is sparse, further investigation and refinement of this construct and items is recommended through complementary quantitative (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 total 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 also drew evidence of moderate to excellent test-retest reliability of the scales. Since sample mean scores were stable, variation across time points might have been due to individual differences, a plausible consequence of lockdown and school closure in Portugal – concurrent with data collection – especially in constructs related with social interactions (*Ethics*, *Collaboration*, and *Relationships*), likely hampered during this period. Despite evidence of scale adequacy for drawing evidence of reliable change in constructs over time, further research using IRT methods (e.g., growth models) in a setting outside COVID-19 impositions might reinforce these findings.

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 the same construct. However, previous research has identified a moderate to strong correlation between similar constructs (r = .64; Sweet et al., 2012), not differing much from the our estimated raw correlation. As such, these findings should tentatively bear on further studies, since disattenuated correlations might over inflate estimates (Murphy & Davidshofer, 2005). More robust interpretations would also be possible through integration of these scales as indicators of a higher-order latent variable along with other Psychological scales, as generally posited in our PPLA model (Mota et al., 2021). Refinement of items as previously suggested and further replications might also shed light on this correlation, evaluating the tenability of collapsing these scales to improve feasibility of the questionnaire.

Correlations among the *Relationships* scale-score and both *Confidence* and *Motivation* were coherent in magnitude with those observed in previous studies (Sweet et al., 2012), providing 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). Similarly, the *Collaboration – Motivation* correlation was also supported by similar previous results (W. 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 also be further supported by integrated 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. For this, we used MSA, a non-parametric scaling technique that models these scales using a cumulative model, allowing items to 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, since 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 of this study, 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 changes over time based on these scales.

Also, despite being a useful and powerful method with increasing traction in instrument development, we 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 reliability (test-score and test-retest) 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 4) 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 can be integrated into the PPLA framework and used to provide a feasible and integrated assessment of the individual journey of each grade 10-12 (15-18 years) student in Portuguese PE.

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.

CHAPTER 4 – PPLA-Q Cognitive Module Construct Validity and Reliability

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

João Mota, João Martins, Marcos Onofre

In preparation for submission

Preprint available at: https://doi.org/10.21203/rs.3.rs-1488826/v1

Abstract

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

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

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

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

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

Background

Physical literacy corresponds to the skills and attributes that individuals demonstrate through physical activity (PA) and movement throughout their lives, enabling them to lead healthy and meaningful lifestyles (Physical Literacy for Life, 2021). It is a key competence that can be developed during quality physical education (PE) (UNESCO, 2015). Despite the wide array of specific definitions available in the literature (L. Edwards, Bryant, Keegan, Morgan, & Jones, 2017; Martins et al., 2020), all integrate a reference to some form of knowledge and understanding (Whitehead, 2010) or content knowledge (Sport Australia, 2019). Similarly, learning outcomes related to knowledge pertaining to PA and movement contexts are imbued into the PE curriculum of many countries (e.g., Society of Health and Physical Educators (SHAPE) America, 2014), including Portugal (Ministério da Educação [Ministry of Education], 2001, 2018).

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

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

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

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

This study is part of a series of studies to gather evidence in support of validity and reliability of the PPLA (Mota et al., 2021, 2022c) – a tool composed of two instruments, the PPLA-Q, and the PPLA – Observation tool (in development). In it, we sought to gather evidence to support construct validity (internal structure and measurement invariance across sexes) and reliability (score reliability and test-retest) of the cognitive module of the PPLA-Q (*Content Knowledge* test) through the lens of IRT. Secondary aims of this study were to assess a) whether modelling data from distractors posed an advantage in locating students in the latent continuum;

b) whether the sum-score possessed enough accuracy for practical-oriented settings.

Methods

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

Participants

Main study (baseline)

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

Retest study phase

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

Measures

The cognitive module of the PPLA-Q is part of a questionnaire developed to assess the psychological, social and part of the cognitive domains of Physical Literacy, inspired by the Australian Physical Literacy Framework (Sport Australia, 2019) and the Portuguese PE syllabus. A previous content validity study resulted in Scale-Content Validity Indexes of .87 (average) and .60 (universal agreement), as evaluated by experts; and highlighted adequate cognitive elicitation of students (Mota et al., 2021). This module measures content knowledge in physical activity and movement settings and was comprised of 10 items: 7 single selection questions, 2 multiple selection questions, and a close-type question; these items are subdivided into 5 different themes, with 2 items each (one designed to assess lower, foundational knowledge, the other designed to assess deeper knowledge application; Table 21).

Procedures

Main study (baseline)

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

Retest study phase

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

Analysis

All statistical analysis used RStudio 1.4.1106 (RStudio Team, 2020), with R 4.0.1(R Core Team, 2020). We had marginal missing data (< 1%) in each item (Table 21, column, 3). Prior to parametric IRT analysis, two datasets were derived: 1) answers coded as dichotomous (i.e., correct, or incorrect); 2) answer coded as polytomous (ordinal for items 5,6 and 10, and nominal for the remainder). Ordinal coding of multiple selection items used a penalization for each incorrect choice (i.e., -1 point) to reduce chances of students obtaining maximum score through selection of all options. The four blocks of item 6 (cloze-type) were collapsed into a single variable to suppress the possibility of local dependency between blocks; students who selected two options in each block were coded as missing (2 cases).

Table 21. Responses and missing data for the PPLA-Q content knowledge test (N=508)

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

F – Foundation level; M – Mastery level
Note: correct responses are bolded
[1]Multiple selection items; [2]Close-type item; [3] Equivalent to difficulty index (*p*) x 100

Model estimation

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

Model fit and selection

Limited-information statistic C2 (Cai & Monroe, 2014) and corresponding p-value were used to assess absolute fit of each estimated model to the data. A .05 significance level was used, with non-significant p-values indicating good absolute fit. The Root Mean Square Error of Approximation based on C2 (RMSEA_{C2}) with a threshold of .06 (C. R. Li, 2019) was used as indicative of adequate approximate fit. Comparative Fit Index (CFI) was used as an auxiliary indicator of model fit with a tentative threshold of .95, since to our knowledge, no research has established the adequacy of this index based on the C2 statistic. Comparison between the nested 1PL and 2PL models used the likelihood-ratio test (LRT; e.g., Finch & French, 2015) based on the -2LL statistic for each model, with a significance level of .05, to assess whether adding parameters significantly improved the fit of the model. Comparison between non-nested models (NRM, 2PNL, 2PNL + GRM) used both the Akaike informatic criterion (AIC; Akaike, 1998) and Bayesian information criterion (BIC; Schwarz, 1978), with lower values indicating better model fit.

Relative efficiency, calculated as the ratio between the total information available in a more complex model versus that of a less complex model (Finch & French, 2015; Lord, 1980), was used to assess the trade-off between information and model complexity. Item fit was assessed through the significance (p-value <.05) of the *S.X*2 statistic (Orlando & Thissen, 2000, 2003) and its accompanying RMSEA. For concision's sake, item parameter estimates, and item/option characteristic curves plot are presented only for the best fitting model. Person fit was assessed through the *Zh* statistic (Drasgow et al., 1985), with a threshold of |1.96| for the final model, using *mirt*.

Score Reliability

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

Assumptions

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

Score correlations

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

Differential Item and Test Functioning

We analyzed differential functioning at item (DIF) and test level (DTF). DIF analysis was performed between sexes using a two-stage approach. First, a mixed-format multiple-group IRT model with no equality constraints across-groups was used as reference to run the DIF function in *mirt* – which adds, and tests via LRT, equality constraints for one item at a time, returning multiplicity-controlled (Benjamini & Hochberg, 1995) p-values. Three items with highest p-values were selected as anchors (i.e., assumed invariant) and a final addictive sequential analysis was run on the anchored model, with freely estimated means and variances. DTF analysis were performed on the final anchored model via the sDTF statistic with 1000 draws (Chalmers et al., 2016) using the females as reference group – before this could happen, a case had to be removed to equalize the number of categories used in item 1 across groups. sDTF represents the number of points on the test, on average, that the reference group will score higher (Chalmers et al., 2016)

Test-retest reliability

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

Results

Model fit

The 2PNL+GRM model showed the best absolute fit (C2(21) = 23.92, p = .30) and approximate fit (RMSEA_{C2} = .017 [0, .043]) to the data, out of all the models (Table

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

As expected, the amount of information offered by each of the models increased as the number of estimated parameters increased: the addition of the discrimination parameter in 2PL model offered a 9% increase in information against the 1PL; modelling data as nominal (NRM) increased the information further by 67%, a value which increased by 6% with the nested 2PNL model. With the mixed-format model (three items estimated using the GRM), the total information decreased by 3% versus all items estimated using the 2PLN (Table 22) – this decrease happens in the lower range of θ, approaching similar levels of information around -1.5 (Figure 11).

Table 22. Model fit for Item Response Theory models

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

GRM – Graded Response Model; RMSEA – Root Mean Square Error of Approximation; CFI – Comparative Fit Index; AIC – Akaike's Information Criteria; BIC – Bayesian Information Criteria

[a] Log-likelihood ratio test 1PL – 2PL model: $\Delta \chi^2 = 28.80$, $\Delta \chi^2$ df = 9, p = .001

[1] Borderline item with p = .05, and $RMSEA_{x2}$= .03

[2] Information previous model / information of current model (Lord, 1980)

According to the S.X2 statistic, item fit was sequentially improved from the 1PL model to the 2PNL (Table 22). In the 2PNL + GRM, item 7 displayed a borderline p-value (.05), albeit with a low RMSEA value (.03). We identified 9 students whose *Zh* statistic was higher that |1.96| in the final mixed model with values ranging from -

4.33 to -2.10. Analysis of their response pattern revealed that their removal would invalidate re-estimation of the mixed model (which requires 3 unique categories per item) to assess their impact on item parameters, and so we chose to keep them given their likely low impact.

Score Reliability

Parallel to the increase in information from the 1PL model to the 2PNL model, marginal reliability showed an increase throughout these models, decreasing slightly in the 2PNL + GRM model (Table 22). Conditional reliability analysis throughout different values of θ, revealed that the 2PNL model attained acceptable levels of reliability (i.e., r_{xx} = .70) from -3 to around -1, while the mixed-format model only did so from around -2 to -1 (Figure 12); the NRM model got closer to the threshold in the -3 to -2 range, providing equivalent reliability to the mixed model until θ≥0, where it underperforms comparatively to both aforementioned models. For comparison purposes, CTT reliability coefficients were estimated as Cronbach's α= .48, and McDonald's ω= .49, and were like the marginal reliability of the 1PL model.

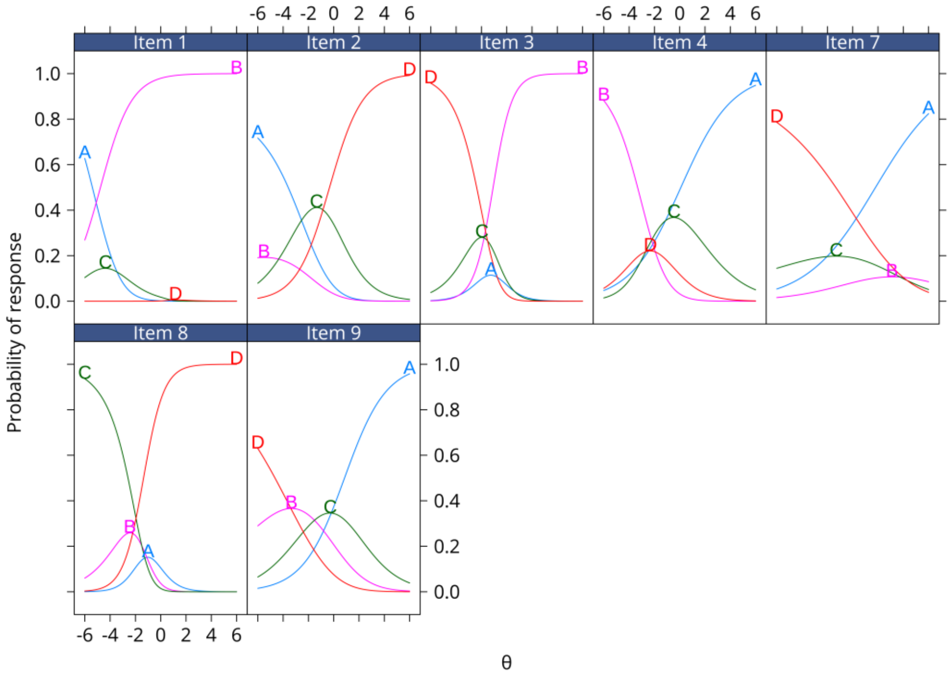


Figure 8. Option Characteristic Curves for items 1-4, 7-9 (2-Parameter Nested Logit + Graded Response Model); A-D – response options

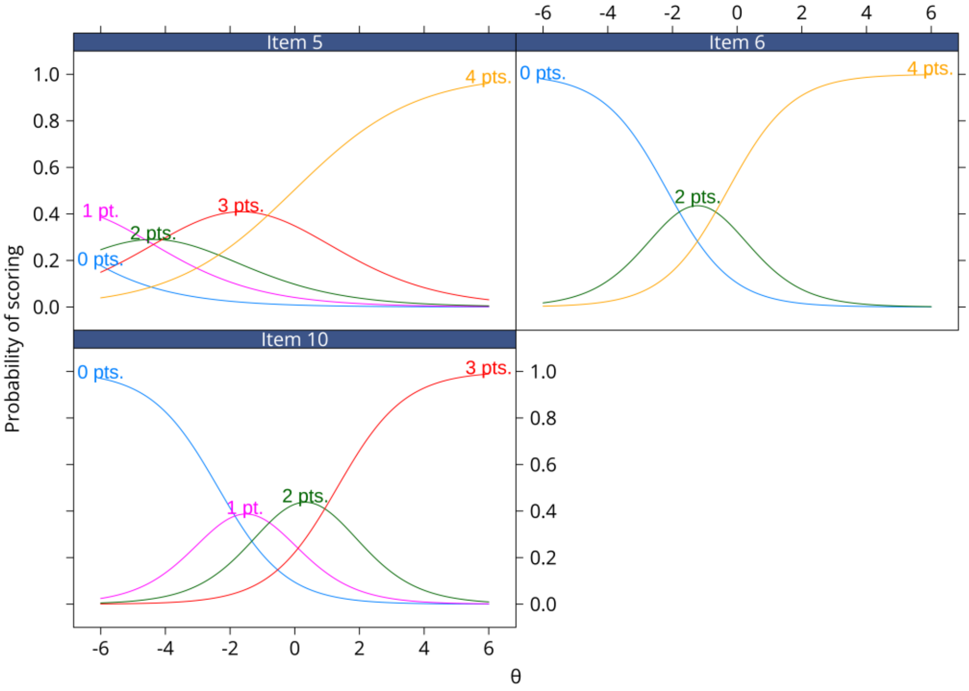


Figure 9. Option Characteristic Curves for items 5,6, 10 (2-parameter Nested Logit + Graded Response Model

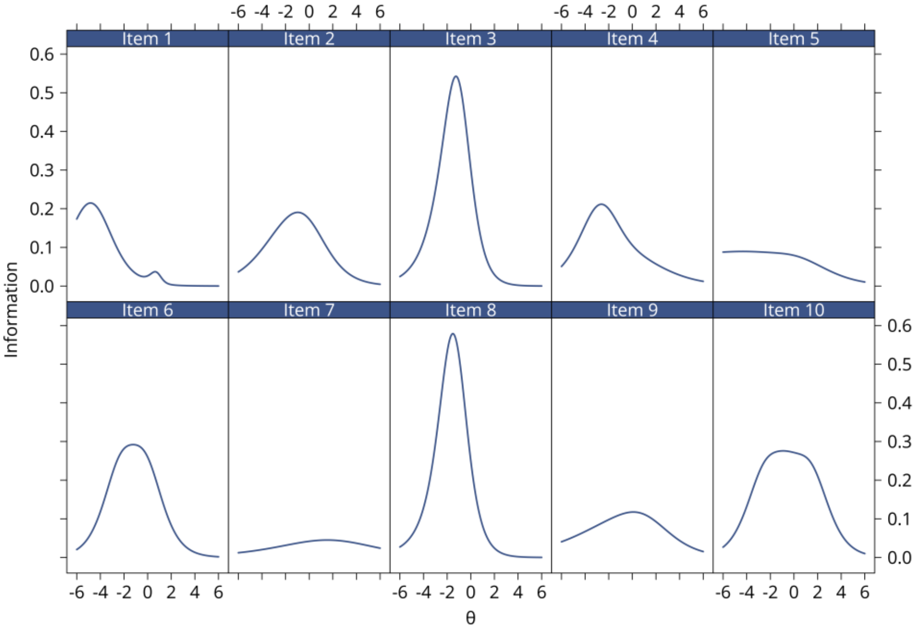


Figure 10. Item information curves for the 2-Parameter Nested Logit + Graded Response Model

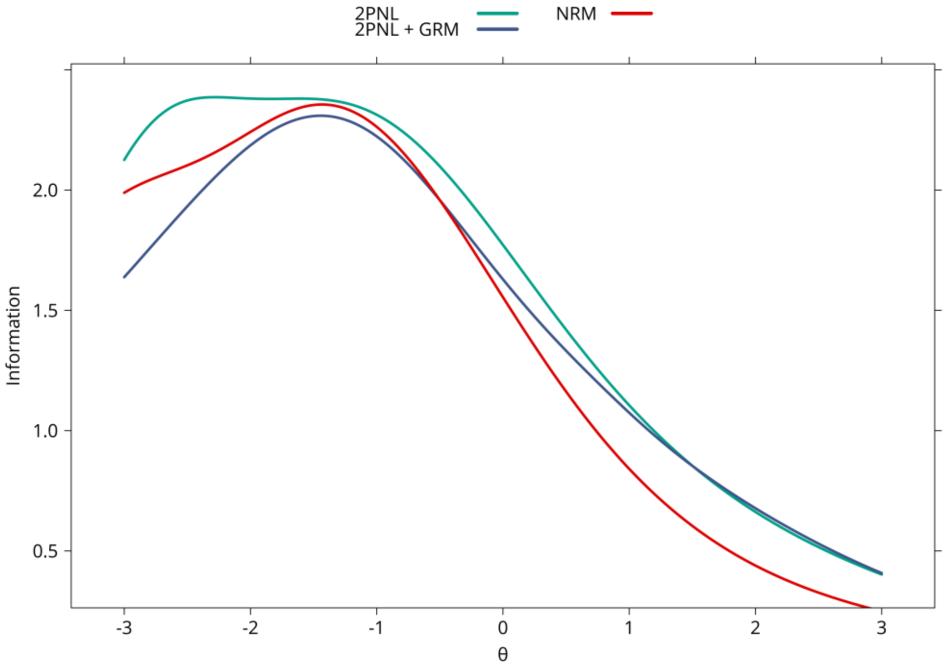


Figure 11. Test Information comparison between the Nominal Response Model (NRM), 2-parameter nested logit (2PLN) and 2-parameter nested logit + graded response model (2PLN+GRM)

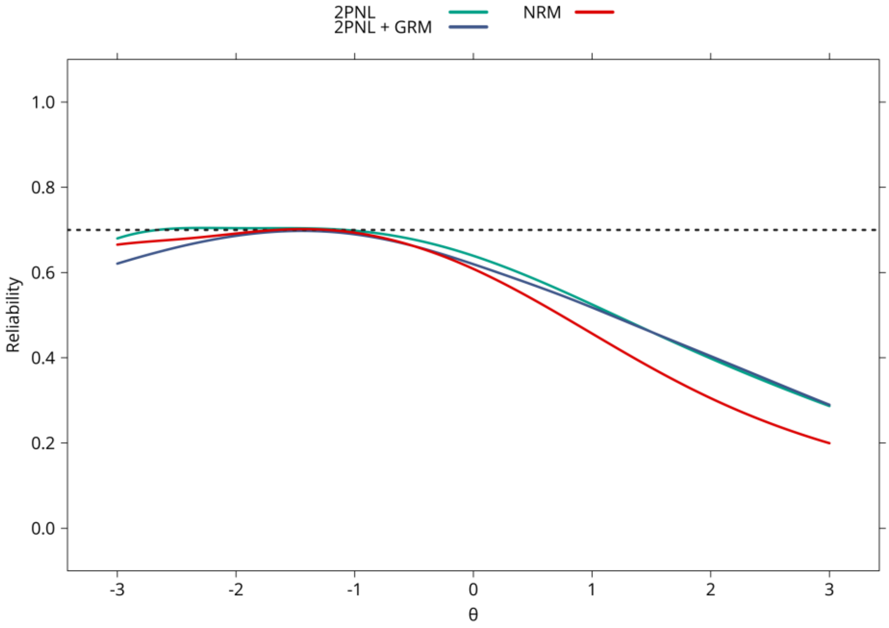


Figure 12. Conditional reliability plot comparison between the Nominal Response Model (NRM), 2-parameter nested logit (2PLN) and 2-parameter nested logit + graded response model (2PLN+GRM)

Assumptions

Preceding item parameter interpretation, we tested the unidimensionality, and local independence assumptions of the best fitting model (2PNL+GRM). During unidimensionality assessment, an exploratory two-factor model fitted better than its one-factor counterpart (significant LRT test), at the cost of parsimony (higher BIC statistic; Table 23). Analysis of item loadings on factors revealed no interpretable pattern. Implications of this finding for interpretation of this test will be further discussed in the Discussion section. We used Yen's Q3 to check for any large violations of local independence. After controlling for θ, no items showed a pairwise residual correlation higher than |.20|. Absolute values ranged from .06 to .16 (not shown), with no discernable pattern of residual correlation among content duplets (e.g., item 1 and 2).

Table 23. Unidimensionality assumption testing for the 2-parameter nested logit+ graded response model

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

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

Item parameters

For concision's sake, we display only item parameters (Table 24) and option characteristic curves (Figure 8 and Figure 9) for the 2PNL+GRM model. According to the CTT difficulty index (*p*), some items designed to be harder for the same content were found to be easier instead (item 5 and 6, and 7 and 8; Table 21); IRT parameters estimate the same relative pattern for these duplets, however, suggest that item 5 is only more difficult than item 6 at maximum score (Table 24). IRT model difficulty parameters also propose a different relative ordering of item's difficulty, with item 7 being the hardest in the test (*b*= 1.805), instead of item 10 (*p* = .26). Discrimination parameters for the correct response ranged from 0.368 (item 7) to 1.332 (item 3); as such, items with lower discrimination parameters, also display flat information trace lines (Figure 8), providing low amounts of information across the whole range of θ.

Some items modelled as 2PNL displayed flat distractor trace lines: item 7's B and C; item 1's C and D, item 2's B, item 3's A distractors were not very discriminative nor very popular (i.e., low a and γ, respectively; Figure 8 and Table 24). Distractors' order, according to their *a* parameter (De Ayala, 2009), was coherent with the theoretical correctness (i.e., based on item's content) of each distractor.

Scores correlation

The scores estimated using the 1PL model correlated perfectly with the sum-score (Table 25). Scores estimated using other models displayed strong, albeit decreasing, correlation with the sum-score according to the degree of parameterization of the model (with the 2PNL being the most parameterized model, and lowest correlated, *r*= .89, [.87, .91]). There was a close to perfect correlation between the scores estimated using the 2PNL and the mixed-format 2PNL+GRM – different estimates mostly in the -1 to -2 θ range (Figure 13).

Table 24. Item parameters for the 2-parameter nested logit + graded response model

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

CTT – Classical Test Theory

[1] Difficulty order

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

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

Note: letters in parentheses indicate the category's label

Table 25. Pearson correlations [95%CI] between estimated scores of each model

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

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

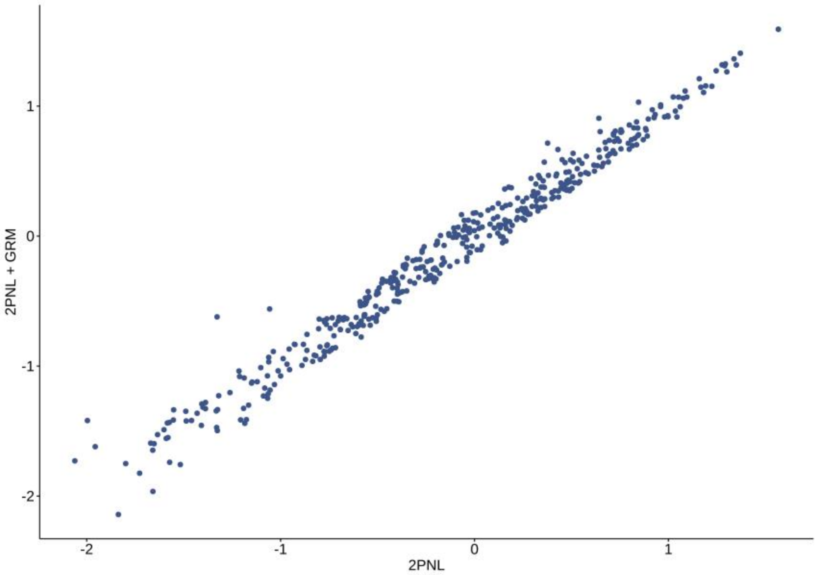


Figure 13. Scatter plot of estimated scores using 2-parameter logit model (2PNL) and 2-parameter nested logit + graded response model (2PLN+GRM)

Differential Item and Test Functioning

We found evidence of DIF in item 1 (p = 0.018, X2 (2) = 8.04). Analysis of the item parameters and OCC (Figure 14) highlighted the existence of non-uniform DIF (Finch & French, 2019) – item is easier and has lower discrimination for boys than for girls (*b* = -3.865 *versus b* = -3.028, and *a* = 0.875 *versus a* = 1.952). Also, distractor parameters suggest different functioning at distractor-level.

To analyze whether the detected DIF would translate in DTF, we calculated *sDTF* statistic (using females as the reference group). Results suggests non-existence of significant DTF (*sDTF* = -0.06 [-0.65, 0.58]; *sDTF%* = -0.14%, [-1.67, 1.49], p = .86). This would mean that, on average, boys would score an estimated 0.14% (0.06

points) higher than girls. Graphical analysis (Figure 15) shows that this is the case mostly around the -2 to -1 θ range.

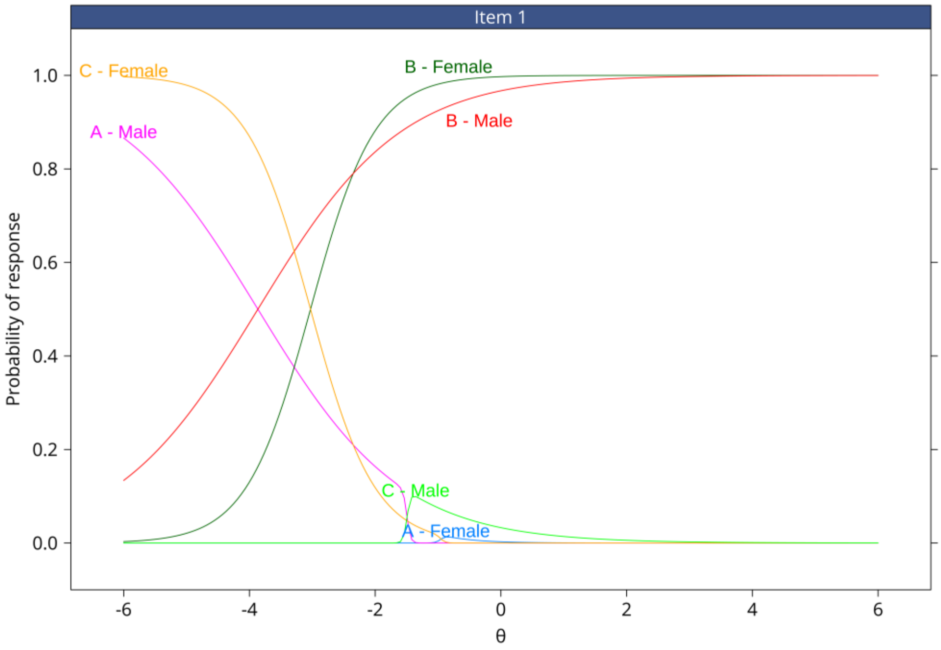


Figure 14. Option Characteristic Curve stratified by sex for Item 1 (Differential Item Functioning)

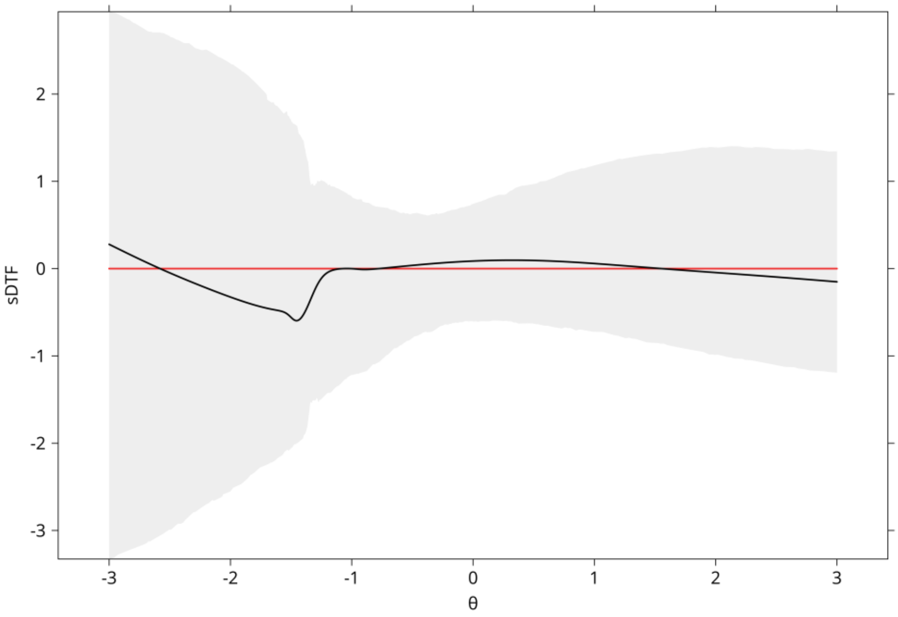


Figure 15. Signed Differential Test Functioning pl−t – Reference group (red) Female; (black) Male

Test-retest reliability

To assess the test-retest reliability of the scores estimated using the 2PNL + GRM model, we computed the Intraclass Correlation Coefficient (ICC) for two applications spaced 15 days in 73 students. There was poor to moderate/good test-retest reliability in these scores (ICC = .56, [.38, .70]; not shown) (Koo & Li, 2016). Follow-up analysis at item-level using Svensson's method scoring on dichotomously scored items, suggested that 6 items showed an acceptable percentage of agreement (>.70; Table 26). All other 4 items displayed signs of small individual variability (significant RV ranging from .04, to .11); additionally, item 4 displayed a small downwards systematic disagreement (RP < 0), while item 10 displayed a small upwards systematic disagreement (RP > 0).

Table 26. Svenson's agreement based on ordinal paired data and Classical Test Theory difficulty at both time points (N=73)

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

PA – Proportion of agreement; RP – Relative position; RV – Relative rank variance
Note: values in which the 95%CI does not include 0 are bolded

Discussion

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

Model fit

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

Dimensionality analysis under this model suggested the existence of a possible second factor. Given the complexity and number of cognitive and personality factors at play during item response, it is usually the case that tests are not strictly unidimensional (Hambleton et al., 1991), and that the substantive consequences of a violation of this assumption must be analyzed according to the intended application of the test (Wells & Faulkner-Bond, 2016): in practice, small degrees of multidimensionality might not distort item parameters and score estimates as long as essential unidimensionality is assured (Harrison, 1986). Analysis of the residual correlation between items did not suggest any significant clustering pattern (> |.20|) between content duplets which could happen due to sampling from the same specific subdomain (i.e., content theme). Some residual correlation (|.10 - .16|) did happen between item 2, 8 and 10 which, we surmise, could be due to similarity of the cognitive processes involved in response (i.e., analysis), or due to closer relationship between content domain for these items (energy balance, health benefits of different types of training, and body composition and its effect on health). As such, these results seem compatible with a parsimonious stance: that a single essential latent trait is being measured in grade 10 to 12 students – general content knowledge in the context of PA. Nonetheless, further studies should test this idea using other methods (e.g., bifactorial IRT modelling), as well as different stances on measurement – assuming that content knowledge could be surmised under a composite-formative model (Stadler et al., 2021).

Score reliability & correlations

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

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

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

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

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

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

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

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

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

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

Item parameters

Regarding item's estimated difficulty versus their intended difficulty, 3 out of 5 duplets behaved as expected (i.e., item evoking higher-order cognitive abilities as harder, than their lower-order counterparts) with item pairs 5 and 6, and 7 and 8 not adhering to this. In the first case, both are scored as multiple selection items, and our data suggests that item 5 is only more difficult than item 6 at maximum score (Table 4), while it is easier for intermediate scores (i.e., scoring points in the latter requires higher ability, than in the former, except for maximum score). This could be result of higher plausibility of distractors in item 5 (selected by ~60 to 65% of respondents; Table 21) and scoring penalization to wrong selection inflating the difficulty of achieving maximum score. It is also plausible that our decisions regarding coding of multiple selection items (5 and 6) might have introduced a degree of bias in the results by restricting the range of combinations (i.e., 2 points in item 5 could be obtained by multiple combinations of right and wrong answers). In the future, different coding schemes might be considered and compared.

In the second's duplet case, multiple factors might be at play: a) the ability to recall information which is not used daily (i.e., recommendations for physical activity, in item 7) might be confounding the intended difficulty as students were not aware that they were going to be tested; b) despite being based on a lower-order cognitive ability (memorization), these guidelines require a specific knowledge that cannot be inferred using an understanding of biology, or general health literacy, and as such, need to be taught explicitly during PE classes. This data is in accordance with previous research (Marques et al., 2015) that suggests that Portuguese students do not know the PA guidelines for health promotion. Nonetheless, a careful look at the distractor's popularity (δ; Table 24) suggests that they seem to be aware of the guidelines for children and adolescents, while not knowing the specific ones for adults (distractor D). This implies that more attention should be dedicated to explicitly teaching these guidelines, with more emphasis on those for adults, since arguably, they will be of most importance in the near-future of high-school students.

DIF and DTF

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

Test -retest reliability

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

Strengths and limitations

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

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

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

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

Conclusion

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

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

Declarations

Ethics approval and consent to participate

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

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

Consent for publication

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

Availability of data and materials

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

Competing interests

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

Funding

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

Authors' contribution

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

Acknowledgements

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

CHAPTER 5 – PPLA-O Development, Construct Validity and Reliability

Portuguese Physical Literacy Assessment Observation (PPLA-O) for adolescents (15-18 years) from grades 10-12: development and initial validation through Item Response Theory

João Mota, João Martins, Marcos Onofre

In preparation for submission

Preprint available at: https://doi.org/10.21203/rs.3.rs-1488826/v1

Abstract

Background: Aims of these studies were to develop the PPLA-Observation instrument (PPLA-O) to assess the physical and part of cognitive domain of Physical Literacy through data collected routinely by Physical Education (PE) teachers; and assess the construct validity (dimensionality, measurement invariance, and convergent and discriminant validity) and score reliability of one of its modules (Movement Competence, Rules and Tactics [MCRT]).

Methods: Content analysis of the Portuguese PE syllabus and literature review were used for domain identification of the PPLA-O. Multidimensional Item Response Theory (MIRT) models were used to assess construct validity and reliability, along with bivariate correlations in a sample of 515 Portuguese grade 10-12 students (M_{age} = 16, SD =1).

Results: PPLA-O development resulted in an instrument with two modules: MCRT (22 physical activities) and Health-Related Fitness (5 protocols); both assessed with teacher-reported data entered in a spreadsheet. A two correlated dimensions Graded Response Model (Manipulative-based Activities [MA], and Stability-based Activities [SA]) showed best fit to the MCRT data, suggesting measurement invariance across sexes, and adequate to good score reliabilities (MA = .89, and SA = .73). There was a moderate to high correlation (r = .68) between dimensions, and boys had higher scores in both dimensions. Correlations among MCRT scores and HRF variables were similar in magnitude to previous reports in meta-analysis and systematic reviews.

Conclusions: The resulting PPLA-O is composed of two modules that integrate observational data collected by PE teachers into a common frame of criterion-referenced PL assessment. While the HRF module makes use of data collected through widely validated FITescola® assessment protocols, the MCRT makes use of teacher-reported data collected in a wide range of activities and movement pursuits to measure movement competence and inherent cognitive skills (*Tactics* and *Rules*). We also gathered initial evidence supporting construct validity and score reliability of the MCRT module. This highly feasible instrument can be used to provide Portuguese grade 10-12 (15-18 years) PE students with feedback on their PL journey, along with the other instrument of PPLA (PPLA-Q). Further studies should focus on assessment of inter and intra-rater reliability and criterion-related validity of its two modules.

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

Background

Physical literacy (PL) is a holistic concept composed of four inter-related domains (physical, emotional/psychological, cognitive and social) referring to the skills and attributes that individuals demonstrate through physical activity (PA) and movement throughout their lives (Physical Literacy for Life, 2021; Sport Australia, 2019). This concept is also at the heart of recommendations posed towards quality Physical Education (PE) for school-aged children and adolescents (Roetert & MacDonald, 2015; UNESCO, 2015).

Two crucial elements within the physical domain of PL are movement competence (MC) and health-related fitness (HRF), as they are conceptualized as part of a spiral of engagement that leads to increased PA participation in children, that might strengthen into adolescence (Stodden et al., 2008, 2009) – a stage in life in which we will focus, given their concerning low levels of PA (Guthold et al., 2020). However, if the goal is a *meaningful* and involved PA participation, its decision-making and tactical aspects (elements of the cognitive domain of PL) need to be also considered (Bunker & Thorpe, 1982; Dudley, 2015; Sport Australia, 2019; Whitehead, 2001).

Development of these elements is an explicit, or implicit part of some PE syllabus (e.g., Society of Health and Physical Educators (SHAPE) America, 2014), as is the case in Portugal (Ministério da Educação, 2001a, 2001b, 2018b). Here, data on MC – through an authentic assessment lens, that integrates movement and decision-making skills (Slade, 2010) – and HRF of students is routinely collected by PE teachers. These are qualified movement professionals, that observe students in many settings (Essiet et al., 2021; Faught et al., 2008), and may be in a privileged position to assess multiple aspects of student's development (Harlen, 2005, 2009). Nonetheless, while HRF assessment makes uses of standardized protocols (FITescola®; Direção-Geral da Educação & Faculdade de Motricidade Humana, 2015) that produce generalizable and interpretable data for educational and research stakeholders, within and outside of schools, this has not been the case for assessment of MC.

One option to solve this issue would be the use of available motor skill assessment batteries; however these suffer from multiple drawbacks: 1) they require additional training and/or lesson time for correct application (Shearer et al., 2021), and so lower their feasibility in PE settings; 2) they focus mostly on children (Hulteen et al., 2020); 3) those available for adolescents are generally product-oriented (Tidén et al., 2015), providing assessment only in discrete, low-generalization tasks (Giblin et al., 2014) that lack the needed ecological validity (Stodden et al., 2008) to understand engagement in advanced physical experiences in a variety of domains and environmental constraints (Burton & Rodgerson, 2001; Giblin et al., 2014) – a characteristic that defines motor development in adolescence (Gallahue, 1996; Goodway et al., 2020); and, 4) they neglect the decision-making aspects previously mentioned, requiring separate use of other instruments, that are however limited to formalized games (Gréhaigne et al., 1997; Oslin et al., 1998).

This problematic motivated the development of a criterion-referenced instrument that could frame observational data collected by teachers in the physical and cognitive domains into the Portuguese Physical Literacy Assessment (PPLA) tool, which already counts with measures to assess all other domains of PL in adolescents (aged 15-18)(Mota et al., 2021, 2022c, 2022b).

Our aims for the following studies were to a) develop the PPLA-Observation based on review of relevant conceptual frameworks and the Portuguese PE syllabus – resulting in two modules, the Movement Competence, Rules and Tactics (MCRT) module and the Health-Related Fitness (HRF) module; b) investigate the dimensionality structure of MCRT module through Item Response Theory (IRT) methods; c) test this structure for differential item functioning (DIF) according to sex, as comparisons between sexes are likely in the future, due to suggested differences in object-controlling/manipulative skills (Barnett et al., 2016); d) establish support for convergent and discriminant validity, and score reliability for this module. A secondary aim was to draw inferences for scoring and criterion-referenced cut-scores mechanisms. We did not focus on validation of the HRF module as it is comprised of measures (i.e., results obtained through FITescola®

protocols) that have already published evidence in support of validity and reliability – further details in the Results section.

Methods

Overview

The development and testing of the *Portuguese Physical Literacy Assessment Observation* (PPLA-O) followed a common philosophy and multiple phase methodology to that of the other part of PPLA (Questionnaire; Mota et al., 2021). It was inspired by the physical and cognitive domains of the PL model proposed in the APLF (Dudley, Keegan, et al., 2017; Sport Australia, 2019), and by the Portuguese PE syllabus (Ministério da Educação, 2ª01a, 2001b, 2018b).

These studies entailed domain identification and measure selection, resulting in an instrument with two modules: HRF and MCRT (Table 28).; followed by content analysis of the PPES according to chosen taxonomies to ensure content validity. A pilot test evaluated feasibility of the data entry for PE teachers. Finally, we assessed the dimensionality and reliability of the *Movement Competency, Rules and Tactics* module. Since the HRF module is grounded on widely used and reported protocols (i.e., FITescola©; Direção-Geral da Educação & Faculdade de Motricidade Humana, 2015), no validation was done. In all phases, adherence to standards for instrument development and validation was sought (American Educational Research Association et al., 2014; Mokkink et al., 2018).

Domain identification and measure selection

Similar to the procedures conducted for the development of the PPLA-Q (Mota et al., 2021), a theoretical framework was established for each of the nine selected elements in the physical and cognitive domains based on literature review of relevant theories in the fields of motor development, physical fitness and PE; supported by previous review efforts done by the APLF team (Dudley, Keegan, et al., 2017), and analysis of the Portuguese PE syllabus (PPES; Ministério da Educação, 2ª01a, 2001b, 2018b). Afterwards, each selected element was mapped into the two-level PPLA framework (Mota et al., 2021). This framework establishes a *Foundation* (initial development that enables participation in movement and PA) and *Mastery* level (relational understanding and application of skills) of development for each element, based on the original APLF work, and the structure of observed learning outcomes taxonomy (SOLO; Biggs & Collis, 1982). Operational definitions per element and level were based on the APLF (Sport Australia, 2019) (Table 28). Then, based on the PPES and its assessment norms, measures or instruments for each element were selected to maximize feasibility and ecological validity.

Since, as we will detail in the Results section, the PPES uses an integrated criterion-referenced assessment of movement competencies, along with rules' knowledge and tactical development, a summative content analysis of the syllabus was conducted (Hsieh & Shannon, 2005) to study possible factorial structures that would allow to disentangle these various elements from each other. Coding was made by the lead investigator, using a deductive categorization (e.g., Elo & Kyngäs, 2008) with categories extracted from the respective theories or models; as no specific taxonomy existed for the *Rules* element, a inductive approach was taken. For the *Movement Competence* skills, sport/specialized skills in each chosen activities were assessed for the diversity of movement skills required in its execution, based on Gallahue's (1996) taxonomy of Locomotion, Manipulative and Stability movement skills, along with Dudley's (Dudley, 2015) taxonomy for Moving with equipment (or Object Locomotion). For the *Tactics* element, diversity of tactical actions were counted according to the Game Performance Assessment System (Oslin et al., 1998).

Pilot testing

Concurrent with the pilot test of the PPLA-Q (Mota et al., 2021) in November 2020, two PE teachers from the involved classes were asked to complete the resulting PPLA-O from the previous phase. PPLA-O took the form of a spreadsheet file (Additional File 5) where teachers could enter all results from the selected 1) proficiency levels for MCRT – ordinal coded; and 2) HRF protocols – continuously coded, except for Shoulder Stretch, which was coded as a binary variable; along with demographic information for each student. Feasibility was assessed through qualitive comments on the clarity of the provided instructions for data insertion, and identification of *bugs* in the automated spreadsheet files used to generated unique codes for each student (to assure anonymity) and insert data.

IRT Analysis of the Movement Competence, Rules and Tactics module

Participants

This study used the same sample as previous PPLA-Q validation studies. Sampling procedures are fully described in Mota et al. (2022c). Briefly, a convenience sample of 521 grade 10-12 students from 25 classes in 6 public schools in Lisbon's metropolitan area was used. Recruitment was stratified by grade, and course major according to population percentage quotas. Schools from diverse socioeconomic backgrounds were chosen to increase sample representativeness. Student sample characteristics are summed up in Table 27. Data about students was reported by 22 PE teachers. Sample size conformed to recommendations for multidimensional graded response models (GRM) (Jiang et al., 2016).

Measures and Procedures

PPLA-O was completed by the PE teachers (N=22) of each class from January to March 2021. Data collection for this tool was concurrent with the one for PPLA-Q validation studies (Mota et al., 2022c, 2022b). As such, upon acceptance to participate, teachers were sent the PPLA-O matrix and were asked to return the latter upon data collection of the PPLA-Q. Since a lockdown was in effect due to the COVID-19 pandemic for most of data collection, teachers were asked to provide the most recent data prior to lockdown, according to the levels provided in the PPES and protocols of the FITescola®. Despite not being part of the PPLA-O, height and weight information were collected to calculate body mass index (BMI) for each student. This measure would be used for testing relevant correlations with measures in the MCRT module.

Analysis

All analysis were performed in RStudio (RStudio Team, 2020) with R 4.1.0 (R Core Team, 2020). Partial PE proficiency levels (e.g., partial Elementary level) were collapsed into the adjacent lower category to equalize assessment across schools – since its common for each school to define their own criteria for these partial levels as a mean to motivate students.

Table 27. Student sample characteristics

Characteristic	N = 521[1]
Sex (n miss. = 2)	
Female	303 (58%)
Male	216 (42%)
Age	16 (1)
Grade	
10	208 (40%)
11	144 (28%)
12	169 (32%)
Major	
Economics	76 (15%)
Humanities	166 (32%)
STEM	279 (54%)
School	
School 1	40 (8%)
School 2	67 (13%)
School 3	21 (4%)
School 4	71 (14%)
School 5	208 (40%)
School 6	114 (22%)

STEM – Sciences, Technology, Engineering and Math
[1] *Statistic presented:* n (%); M(SD)

Descriptive statistics (Table 30) were generated using the *psych* (Revelle, 2021), *naniar* (Tierney et al., 2021) and *summarytools* (Comtois, 2021) packages. Students without any collected data (n =6; non-participation in PE due to injury) were then removed from the dataset. Little's test was used to assess tenability of data missing completely at random (MCAR; Little & Rubin, 2020). Results of χ^2 (766) = 1681, p <.001 (with missing patterns = 91) provided evidence against MCAR. The assumption of missing at random (MAR) was plausible based on results of sensitivity analysis of missing data grouped by class. Two items (Rhythmic Gymnastics, and Modern Dance) were eliminated prior to further analysis due to low observed frequency (n=1, and 0, respectively).

Dimensionality

All IRT models were estimated using Marginal Maximum Likelihood with the expected-maximization algorithm in *mirt* (version 1.34.11, Chalmers, 2012), robust to high degrees of missing data (Bernaards & Sijtsma, 1999). A two-stage analysis was performed. In a first stage, sequentially more complex models were estimated until there was no improvement in model-data fit, or convergence issues occurred due to over factoring. As such, we fitted a 1) unidimensional partial credit model (1d-PCM), i1) unidimensional graded response model (1d-GRM), and ii1) exploratory multidimensional correlated GRM (2d-GRM and 3d-GRM). Comparison between models used the likelihood-ratio test (LRT; e.g., Finch & French, 2015) based on the

-2LL statistic for each model (significance level of .05) to assess whether adding parameters (i.e., discrimination) and extra dimensions improved the fit of the model. The Akaike informatic criterion (AIC; Akaike, 1998) and sample-adjusted Bayesian information criterion (SABIC; Schwarz, 1978) provided additional insights, with lower values indicating better model fit.

After an optimal exploratory solution was attained, its standardized loadings (*oblimin* rotated) were assessed to identify non-salient items: with a threshold of $\lambda < .30$ (e.g., Reise & Revicki, 2015) or communality < .40. Cross-loadings were assessed using a variance explained ratio (λ_1^2/ λ_2^2), with values lower than 1.5 (Hair Jr. et al., 2019) considered for elimination depending on factor interpretability. These items were then removed one by one (with model re-estimation) until simple structure was achieved. For the second stage, all previous models were rerun to detect whether the sequential improvement in fit held after removal of items. Finally, item loadings were constrained to load on its salient factor, and a confirmatory GRM model was fit.

In this final solution, the magnitude of standardized loadings and discrimination (slope) parameters were assessed: a) loadings were interpreted as excellent, very good, good, fair or poor when higher than .71, .63, .55, .45 and .32, respectively (Comrey & Lee, 1992); b) discriminations were interpreted as very high, high, moderate, low, and very low when higher than 1.70, 1.35, 0.65, 0.35 and 0.01, respectively (Baker & Kim, 2017).

Differential Item Functioning (DIF)

Before DIF analysis, five cases had to be removed to equalize categories in the Throws and Jumps (both from Athletics) activities. DIF analysis was performed between sexes using a two-stage approach. First, a multiple-group IRT version of the final model was fit with no equality constraints across-groups and used as reference to run the DIF function in *mirt* – which adds, and tests via LRT, equality constraints for one item at a time, returning multiplicity-controlled (Benjamini & Hochberg, 1995) p-values. Three items with highest p-values were selected as anchors (i.e., assumed invariant) and a final addictive sequential analysis was run on the anchored model (i.e., three invariant items constrained to equality), with freely estimated means and variances. Adjusted p-values < .05 were used as threshold for existence of DIF.

Discriminant and convergent validity

Bivariate Pearson and polyserial correlations (and 95% CI) were calculated using the *polycor* (Fox, 2019) and *piercer* (Pierce, 2021) packages using all pairwise complete observations. These were used to evaluate discriminant validity (threshold of r =.85 to discern whether resulting variables were statistically different) and convergent validity based on magnitude reported in similar studies. Magnitudes were interpreted as : very high, high, moderate, and low correlations, when r >.90, >.70, >.50, >.30, respectively (Hinkle et al., 2003). Inter-factor discriminant validity was assessed via correlation in the final MCRT model, using the same .85 threshold.

Reliability and scoring

Marginal reliability (Green et al., 1984) using Expected a-posterior (EAP) (Embretson & Reise, 2000) scores was calculated to quantify average reliability across the θ continuum. These were evaluated as acceptable (ρ_{xx}>.70 ; Nunnaly & Bernstein, 1994), and as good (ρ_{xx}>.80 ; Price, 2017). Thresholds for each item (d_k, or intercept parameter) were transformed into *difficulty* parameters (b_k) using $b_k = -(d_k / a_k)$ (Reckase, 2009) for easier interpretation.

Results

Given the initial focus on the development of the PPLA-O, this section will first describe the results of domain identification and measure selection – including relevant definitions, and a summary literature review of its theoretical framework and relationships with PA participation or other relevant outcomes. It will then present the results of the remaining studies: content analysis, pilot testing and IRT analysis of the MCRT module.

Domain Identification and measure selection

Health-Related Fitness (HRF) module

Physical fitness can be interpreted as the capacity to perform PA and/or physical exercise that integrates most bodily functions involved in movement. (Martínez-Vizcaíno & Sánchez-López, 2008). Some authors suggests that it might be a predictor of PA in youth (Britton et al., 2020; Stodden et al., 2008), with active youth presenting healthier physical fitness profiles (Boreham & Riddoch, 2001). However, this is disputed by other authors (Kemper & Koppes, 2006; Martínez-Vizcaíno & Sánchez-López, 2008).

More robust evidence, however, correlates fitness with various health outcomes throughout the life span (Bushman & American College of Sports Medicine, 2017). Among these, cardiovascular endurance is linked with diverse metabolic markers (Committee on Fitness Measures and Health Outcomes in Youth et al., 2012), mental health (Janssen et al., 2020; Ortega et al., 2008), and cognitive benefits including academic performance (Chaddock-Heyman et al., 2014; Scudder et al., 2014). Musculoskeletal fitness is liked with increased bone density (Committee on Fitness Measures and Health Outcomes in Youth et al., 2012) and positive self-perceptions (Lubans & Cliff, 2011). And, despite there being no compelling link between flexibility and health, the former is suggested to be central to correct posture and increased functional capacity (The Cooper Institute, 2017).

Given its prominent role in a healthy and active life, HRF is an integral part of the PPES, as one of its three major areas, along with physical activities, and knowledge. Its assessment is operationalized through the FITescola© test battery (Direção-Geral da Educação & Faculdade de Motricidade Humana, 2015). This battery, analogous to FitnessGram© (The Cooper Institute, 2017), offers a set of protocols to assess whether children and adolescents meet evidence-based criteria for health-related benefits. From these, we selected the most disseminated ones in PE teacher's practice, that simultaneously adhere to international recommendations (Committee on Fitness Measures and Health Outcomes in Youth et al., 2012; Plowman & Meredith, 2013) (Table 28, column 5), and have extensive validity and reliability evidence (Artero et al., 2011; Lubans et al., 2011; Mayorga-Vega et al., 2014, 2015; Patterson et al., 2001; Vanhelst et al., 2016). The obtention of the Healthy Fitness Zone was mapped as the transition point between *Foundation* and Mastery level for elements in this module, with the Athletic Profile values used as reference for maximum points. The latter is a zone designed to assess athletic potential in youth (Henriques-Neto et al., 2020).

Movement Competence, Rules and Tactics module

Movement competence (MC) can be defined as the development of sufficient movement skills to assure successful performance in a variety of physical activities, be that work or play (Bisi et al., 2017; Burton & Rodgerson, 2001). This concept is employed by Whitehead (Whitehead, 2010) in allusion to a "bank" that enable individuals to respond automatically and meaningful to movement situations. Most commonly, these skills are divided into 1) fundamental movement skills, and 2) specialized movement skills (Gallahue, 1996). Fundamental movement skills are organized series of basic movements that involve combinations of two or more body segments (Gallahue, 1996), and form the building block for specialized movement skills (Logan et al., 2018), which represent application of these fundamental movement skills to specific physical activity or sport contexts with increased refinement (e.g., fielding a groundball; D. L. Gallahue & Donnelly, 2002; Goodway et al., 2020). Different, yet analogous taxonomies include the subdivision into general, refined and specific movement patterns (Durden-Myers et al., 2018). All these movement skills can be categorized into different movement skills sets according to their function (Burton & Rodgerson, 2001) as *locomotor*, *stability* or *manipulative* movement skills (Gallahue, 1996), and present multiple phases and stages of development throughout the lifespan. Other sources add a fourth category that includes movement skills with equipment (e.g., bike, surfboard, skate rollers; Dudley, 2015; Sport Australia, 2019).

MC has a suspected cause-effect relationship with PA (Holfelder & Schott, 2014), with multiple reviews identifying a positive association between the two across childhood (L. E. Robinson et al., 2015). This association also seems to be higher with object control/manipulative movement skills (Barnett et al., 2009; Lubans et al., 2010).

However, few studies have examined this correlation in adolescents (L. E. Robinson et al., 2015). Similarly, positive correlations have been identified with perceived competence (Babic et al., 2014), and health-related fitness (Cattuzzo et al., 2016; Stodden et al., 2009).

In the PPES, MC is developed within the physical activities area, which includes subareas for diverse physical activities (i.e., Team-sports, Gymnastics, Athletics, Racquets, Combat, Roller Skating, Swimming, Rhythmic-Expressive, Traditional games, and Nature exploration). In each of these subareas, multiple physical activities (to which we will refer simply as *activities*, from now on) are used as means of development and assessment of each student through three levels: Introductory, Elementary and Advanced. The Introductory level frames multiple foundational skills and knowledge needed for participation in each activity – usually deals with reduced or constrained gameplay, or pedagogical progressions leading to the formal setting of the activity; the Elementary level refers to the mastery of the main elements of each activity – deals with the full formal setting of the activity; the Advanced level establishes skills and knowledge needed for higher-degree participation in the activities (e.g., performance-settings). Assessment uses a set of rubrics that establish 1) the skill, knowledge, or attitude to be observed, 2) the context (e.g., 2x2 reduced gameplay of volleyball, or a gymnastics sequence composed of predetermined movements, and c) multiple qualitative criteria that describe the action. Given the above frame, we corresponded the Introductory and Elementary levels in these activities with the *Foundation* and *Mastery* levels of the PPLA in all elements of movement competence (i.e., locomotion, manipulative, stability, moving with equipment).

Table 28. Domain identification for the Physical and Cognitive Domain of the PPLA-Observation instrument (PPLA-O)

	Theoretical framework	Operational definition	Definition per level	Instruments / Measures	PPLA-O Module
Physical Domain					
Health-related Fitness					
Cardiorespiratory Endurance	FITescola©	**Ability of the heart and lungs to deliver oxygen to working muscle**	**Foundation:** Building health-related fitness that allows for a functional lifestyle and health-related benefits	PACER/20-meter shuttle run[b]	Health-Related Fitness (HRF)
Muscular Endurance		**Ability of muscle(s) to repeatedly exert force over a sustained period**	**Mastery:** Building health-related physical fitness necessary for excelling in performance-driven settings	Curl-ups (core endurance)[b] 90º push-ups (upper-body endurance) [b]	
Flexibility		**Capacity of a joint or muscle to move through its full range of motion**		Backsaver Sit-and-reach (lower flexibility)[b] Shoulder Stretch (upper flexibility)[b]	
Movement Competence					
Locomotion	(Gallahue, 1996; Gallahue et al., 2020)	**Movement skills that allow a person to move from one place to another (on multiple environments)[a]**	**Foundation:** application of baseline skills and techniques in reduced settings (exercises, reduced or constrained gameplay) (Introductory level in the PPES)	Teacher-reported proficiency levels in Physical Activities in PE	Movement Competence, Rules, and Tactics (MCRT)
Object Manipulation		**Movement skills that use a body part to move or manipulate an object**			
Stability / Balance		**Skills involving balance and weight transfer[a]**	**Mastery:** application in settings representing the physical activity (global, formal level of participation) (Elementary level in the PPES)		
Moving with equipment	(Dudley, 2015)	**Movement skills used to move on, in or with, equipment from one place to another**			
Cognitive Domain					
Rules	(Dudley, 2015; Sport Australia, 2019)	**Explicit or understood regulations and principles governing conduct or procedure with movement and PA**	**Foundation:** Knowledge and compliance with safety rules and regulations of activities **Mastery:** Active participation on the enforcement or adaptation of rules	Teacher-reported proficiency levels in Physical Activities in PE	
Tactics	(Bunker & Thorpe, 1982; Dudley, 2015; Oslin et al., 1998)	**Planed and ad hoc decisions and actions, employed in the moment for the pursuit of goals**	**Foundation:** Accumulation and application of simple tactics to solve a problem (single constraints) **Mastery:** Relational application of tactics in response to multiple constraints		

PA – Physical activity; PACER – Progressive Aerobic Cardiovascular Endurance Run; PE – Physical Education; PPES – Portuguese PE Syllabus

[a] According to the Australian Physical Literacy Framework (Sport Australia, 2019)

[b] FITescola® (Direção-Geral da Educação & Faculdade de Motricidade Humana, 2015)

Rules

Although conceived within the realm of team-sports and games, most literature on rules provide an easy generalization to other movement contexts. Rules provide a structure that manages and guides practitioners' action (Gréhaigne & Godbout, 2013). These can be considered primary or fundamental when they act as constraints that regulate and apply restrictions on the mode of action available to the individual (e.g., scoring rules); or as secondary when they constitute written or unwritten rules that facilitate participation (e.g., safety and ethical rules of organized PA; Dudley, 2015). Both contribute to the *form* of the activity as we know it (Slade, 2010). Understanding of rules and their application is therefore an essential part of every activity – something that Bunker and Thorpe frame as "Game Appreciation" (1982).

Within the PPES, rules' knowledge and understanding are integrated holistically within each activity proficiency level previously mentioned. Thus, all activities promote the learning of safety codes and equipment management, while activities like team-sports and athletics allow learning of more closed scoring and playing rules. These outcomes are framed into the *Foundation* level of this element. At higher level (mostly Advanced), students are asked to be officials and referees, which works as a powerful learning tool to reinforce rule knowledge and conditional application of all aspects of the activity (Slade, 2010). This skill is proposed as part of the *Mastery* level.

Tactics

Tactics can be generally framed as time-sensitive responses to problems posed in movement and PA contexts, be that inherent to game participation (i.e., gaining advantage), or informal PA (i.e., maximizing quality and efficiency) (Dudley, 2015; Gréhaigne et al., 2005). These contexts act as eventful *dynamic systems* (Gréhaigne & Godbout, 2014) that require participants to develop and apply higher-level cognitive skills (e.g., comparing, contrasting, analyzing, evaluation) required for thoughtful decision making (McBride & Xiang, 2004), in interaction with others and the environment (Dudley, 2015). Despite being separated here into two different elements, tactical knowledge and application is mostly conceived as the next (higher-order) level of rules' knowledge, in a learning continuum that frames decision-making within PA (Bunker & Thorpe, 1982; Dudley, 2015; Gréhaigne &

Godbout, 2013): Only after participants can identify the constraints imposed by rules, can they begin to acknowledge degrees of freedom available to act.

Game sense approaches that propose the teaching of PA through reduced or adapted forms of the formal activity (e.g., Teaching Games for Understanding [TGfU]; Bunker & Thorpe, 1982), recognize that the learning of specific skill and tactics constrain each other (Butler & Griffin, 2010); while *traditional*, skill-centered approaches (i.e., analytical) focus on the former as the main constrainer of the capacity to participate in PA. The TGfU approach recognizes the similarity between tactical actions among the many games by categorize them in 1) target games, 2) net/wall games, 3) striking/fielding, 4) invasion games (Bunker & Thorpe, 1982). Based on this taxonomy, the *Game Perfomance Assessment Instrument* typifies tactical action these into six transversal categories : 1) decision making, 2) adjust, 3) cover, 4) support, 5) guard/mark, 6) base (cf. Memmert & Harvey, 2008; Oslin et al., 1998) – skill execution excluded.

Benefits of using these approaches might include increased engagement, enjoyment and motivation in PE classes (Díaz-Cueto et al., 2010). Also, authors argue that awareness and decision making skills might transfer to contexts outside of movement (Dudley, 2015; Sport Australia, 2019), being central to critical thinking, as a general education outcome (McBride & Xiang, 2004).

As aforementioned, the PPES frames tactical skills within the learning of activities and into the diverse levels of learning. Assessment is made in-context, through combination of skills and decision making, coherent with principles of *authentic assessment* (Slade, 2010; Wiggins, 1990). We framed a more constrained application of tactics (i.e., reproduction of descriptive tactics) to the *Foundation* level, while a more critical, relational stance on decision-making was framed in the *Mastery* level.

Given the integrated nature of the *Movement Competence*, *Rules*, and *Tactics* elements, the specification levels for each activity were selected as holistic, process-oriented measures of these elements. A set of 22 physical activities that represent the full breadth of subareas within the syllabus were chosen, with the possibility for teachers to include any other activity assessed. Chosen activities spanned all movement forms (Durden-Myers et al., 2018; Murdoch & Whitehead, 2010) and two of the four types of games according to the TGfU classification (Table 29). Target and striking games are not commonly developed in Portuguese PE and were not included.

Content analysis

Table 29 presents the summary of the content analysis of the PPES. Higher levels of proficiency in each activity entailed a higher diversity of movement skills in all typologies; however, this tendency only emerged between the Introductory and Elementary levels, with almost no new movement skills required when transitioning to the Advanced level. Locomotor skills were required with similar diversity across all types of activities, with two clusters emerging according to manipulative skills (mostly team-sports) and stability (gymnastics and Roller Skating) movement skills: while team-sports required mostly dynamic balancing, twisting, turning, landing, and dodging movement skills, gymnastics uniquely required skills combining inverted supports, rolling, and diverse bending and stretching movement skills. Tactics-wise, a similar pattern was noted with increasing levels requiring a higher diversity of tactical action – without the plateau observed for movement skills. As expected, tactical actions were mostly requested by Team-Sports and Racquets activities.

Finally, regarding rules, four general categories emerged from the analysis. Knowledge and application of safety rules and specific activity rules was mostly observed in the Introductory levels; while identification of referee signals, and officiating were mostly skills required for Elementary and Advanced levels, respectively.

Pilot Testing

Participating teachers had no difficulties with data insertion and regarded the instructions as clear. As expected, data collection on activities and HRF protocols was already part of their lessons, implying no additional effort. They highlighted errors in the code generator spreadsheet and PPLA-O spreadsheet, which were corrected for the next phase of these studies.

Table 29. Content Analysis of the Portuguese Physical Education (PE) Syllabus

	Classification			Content Analysis																		
Physical Activity	Movement Form (Durden-Myers et al., 2018; Murdoch & Whitehead, 2010)	Portuguese PE Syllabus (Ministério da Educação, 2ª01a, 2005, 2ª18a)	TGfU / GCS (Bunker & Thorpe, 1982; Werner et al., 1996)	Locomotion Skills[a] max. points 8			Manipulative Skills[a] max. points 13			Stability / Balance Skills[a] max. points 10			Moving with equipament[b] max. points 6			Tactics[c] max. points 5			Rules[d]			
				I	E	A	I	E	A	I	E	A	I	E	A	I	E	A	Safety rules	Specific rules	Referee signals	Participations as referee / judge
Races (Athletics)	Athletic	Athletics		2	2	2				2	3	3							I			A
Throws (Athletics)	Athletic	Athletics			1	2	1	1	1	1	4	4							I			A
Jumps (Athletics)	Athletic	Athletics		2	3	3				4	5	5							I			A
Wrestling	Competitive	Combat		1	1	1		1	1	5	5	5				2	3	3	I	I	E	E
Judo	Competitive	Combat		1	1	1	1	1	1	6	6	6				2	3	3	I	I	I	
Floor Gymnastics	Athletic	Gymnastics			2	2				5	8	8							E			
Artistic Gymnastics	Athletic	Gymnastics		2	2	2				4	6	8							I			
Acrobatic Gymnastics	Athletic	Gymnastics		3	3	3				7	8	8							I	I		
Rhythmic Gymnastics	Aesthetic and Expressive	Gymnastics		2	3	3	3	4	4	3	8	8							I			A
Handball	Competitive	Team Sports	Invasion	3	3	4	3	3	3	3	5	5				2	5	5	I	I	E	A
Football	Competitive	Team Sports	Invasion	2	3	3	5	5	5	3	5	5				3	3	4	I	I	E	A
Basketball	Competitive	Team Sports	Invasion	4	4	4	4	4	4	4	6	6				2	4	4	I	I	E	A
Rugby	Competitive	Team Sports	Invasion	2	3	3	4	5	5	4	5	5				3	5	5	I	I	E	
Orienteering	Adventure	Nature Exploration		1	1	1				2	2	2				1	1	1	I	I		
Climbing	Adventure	Nature Exploration		1	1	1				3	4	4				1	1	1	I	E		
Roller Skating[e]	Athletic / Aesthetic and Expressive	Roller Skating			2	2				4	6	6	1	1	1				I			A
Table Tennis	Competitive	Racquets	Net							3	3	3				1	1	4		I		A
Badminton	Competitive	Racquets	Net	1	2	2				5	6	6				2	2	4		I		A
Volleyball	Competitive	Team Sports	Net	2	3	3				4	5	6					3	4	I	I	E	
Dance (Modern)	Aesthetic and Expressive	Rhythmic and Expressive		6	7	7				4	5	5										A
Dance (Social)	Interpersonal / Relational	Rhythmic and Expressive		1	1	3				3	3	3							I			
Aerobics	Fitness & Health	Rhythmic and Expressive		3	6	6				2	6	6										

TGfU – Teaching Games for Understanding; GS – Game Sense; I – Introductory proficiency level; E – Elementary proficiency level; A – Advanced proficiency level

[a] Based on (Gallahue, 1996); [b] Based on (Dudley, 2015); [c] Based on Game Performance Assessment Instrument items (Oslin et al., 1998), extended to general decision-making in all activities; [d] Level at which item appears; [e] After Introductory level, Roller Skating takes the form of i) Roller skates Racing, ii) Artistic Roller Skating, or iii) Hockey on Roller skates – analysis presented here refers to ii)

IRT Analysis of the Movement Competence, Rules and Tactics module

Preliminary analysis

Seven activities had lower than 90% assessment rate (Modern Dance, Rhythmic Gymnastics, Rugby, Wrestling, Judo, Acrobatic Gymnastics, and Tennis; Table 30). The most prevalent level of proficiency was Introductory, with the Advanced level attaining only residual prevalence (0 to 5.1% of assessed students). Flexibility protocols had lower percentages of assessed students compared to other protocols (Table 31).

Table 30. Descriptive statistics for teacher-reported proficiency levels in physical activities - Movement Competence, Rules and Tactics Module (N=515)

Physical Activity	Missing cases (%)	Observed Proficiency Levels			
		Non-Introductory[a]	Introductory	Elementary	Advanced
Races (Athletics)	187 (36.3%)	22 (6.7%)	180 (54.9%)	126 (38.4%)	
Throws (Athletics)	346 (67.2%)	2 (1.2%)	87 (51.5%)	80 (47.3%)	
Jumps (Athletics)	392 (76.1%)	5 (4.1%)	87 (70.7%)	31 (25.2%)	
Wrestling	491 (95.3%)	10 (41.7%)	14 (58.3%)		
Judo	490 (95.1%)	3 (12%)	22 (88%)		
Floor Gymnastics	32 (6.2%)	91 (18.8%)	320 (66.3%)	72 (14.9%)	
Artistic Gymnastics	53 (10.3%)	85 (18.4%)	271 (58.7%)	104 (22.5%)	2 (0.4%)
Acrobatic Gymnastics	475 (92.2%)	14 (35%)	26 (65%)		
Rhythmic Gymnastics	514 (99.8%)	1 (100%)			
Handball	114 (22.1%)	78 (19.5%)	212 (52.9%)	111 (27.7%)	
Football	64 (12.4%)	116 (25.7%)	179 (39.7%)	133 (29.5%)	23 (5.1%)
Basketball	43 (8.3%)	84 (17.8%)	265 (56.1%)	123 (26.1%)	
Rugby	500 (97.1%)	8 (53.3%)	7 (46.7%)		
Orienteering	345 (67%)	1 (0.6%)	82 (48.2%)	87 (51.2%)	
Climbing	410 (79.6%)	11 (10.5%)	61 (58.1%)	33 (31.4%)	
Roller Skating	338 (65.6%)	84 (47.5%)	76 (42.9%)	17 (9.6%)	
Table Tennis	297 (57.7%)	23 (10.6%)	141 (64.7%)	54 (24.8%)	
Badminton	8 (1.6%)	56 (11%)	264 (52.1%)	163 (32.1%)	24 (4.7%)
Volleyball	5 (1%)	40 (7.8%)	295 (57.8%)	163 (32%)	12 (2.4%)
Dance (Modern)	515 (100%)				
Dance (Social)	204 (39.6%)	53 (17%)	208 (66.9%)	48 (15.4%)	2 (0.6%)
Aerobics	395 (76.7%)	4 (3.3%)	96 (80%)	20 (16.7%)	
Tennis	469 (91.1%)	2 (4.3%)	38 (82.6%)	6 (13%)	

[a] Non introductory level refers to students that have yet to achieve the standards for the Introductory level

Table 31. Descriptive statistics for teacher-reported results for Health-Related Fitness module (N=515)

Health-Related Fitness Measures	Missing cases (%)	M (SD)	Median
PACER	22 (4.2%)	49.5 (22)	44.0
Push-ups	26 (5.0%)	18.1 (9.6)	18.0
Curl-ups	23 (4.4%)	48.6 (21.7)	45.0
Shoulder Stretch (% of achievement)			
Right	83 (15.9%)	95%	
Left	83 (15.9%)	89%	
Sit and Reach (cm)			
Right	85 (16.3%)	30.7 (8.3)	31.0
Left	84 (16.1%)	30.2 (8.2)	31.0

Dimensionality

In the first stage of analysis, the 2d-GRM presented the best fit according to information criteria (AIC, SABIC and -2LL; Table 32). According to the likelihood ratio test (LRT), freely estimating discrimination (slope) parameters improved the fit from the 1d-PCM to the 1d-GRM; and estimating an additional dimension also improved fit from the 1d-GRM to the 2d-GRM. A 3d-GRM was estimated, however its information matrix could not be inverted, signaling an empirically unidentified model; as such, its estimates are not presented.

Item standardized loadings and parameters were analyzed based on the 2d-GRM exploratory solution. Reasons for item removal are presented in Table 32. As a note, Wrestling item had a borderline variance ratio (1.66), and we opted initially for non-removal based on its added value as unique item concerning Combats activities. However, estimation of the following second stage confirmatory 2d-GRM (with items constrained to load on its salient factor) did not converge. Removal of this item allowed the solution to converge.

Second stage consisted of sequential re-estimation of all models, without removed items, to assess if the results obtained in the first stage were robust. Improvement in fit between models was equivalent to those observed during first stage. Finally, a confirmatory 2d-GRM was fit, with decrease in fit (according to all indices) *versus* its exploratory counterpart, which was expected since the former imposes more constraints to item loadings (cross-loadings constrained to 0).

Table 32. Model fit indices and statistics for the Movement Competence, Rules, and Tactics module

	AIC	SABIC	-2LL	LRT	Removed items (reasons)
First stage					
1d-PCM	8360.60	8407.67	8272.59		
1d-GRM	8026.04	8094.51	7898.03	Δ $\chi^2(20) = 374.56$, $p < .001$	
2d-GI(E)	7889.53	7979.41	7721.53	Δ $\chi^2(20) = 176.50$, $p < .001$	**Aerobics, Tennis, Social Dance** (non-salient loadings) **Orienteering** (low communalities) **Judo, Rugby** (SE larger than slope parameters) **Acrobatic Gymnastic, Wrestling** (problematic cross-loadings)
Second stage					
1d- PCM	7112.58	7145.75	7050.58		
1d- GRM	6928.36	6974.37	6842.36	$\Delta\chi^2(12) = 208.22$, $p < .001$	
2dIRM (E)	6788.18	6847.03	6678.18	$\Delta\chi^2(12) = 164.18$, $p < .001$	
2d- IT GRM (C)	6861.48	6908.55	6476.16	$\Delta\chi^2(11) = 95.29$, $p < .001$[a]	

1d – unidimensional: 2d – multidimensional model with 2 correlated factors; (E) -Exploratory; (C) – confirmatory; AIC – Akaike's Information Criteria; SABIC – Sample Adjusted Bayesian Information Criteria ; -2LL - -2* Log-Likelihood ; LRT – Likelihood Ratio Test; [a]In favor of the exploratory model

Loadings in the final confirmatory solution ranged from very good to excellent (.75 to .92, and .64 to .91), for dimension 1 and 2, respectively (Table 33, Figure 16). An equivalent pattern of moderate (a> .65) to very good (a > 1.70) discrimination parameters (Baker & Kim, 2017) indicates that items are performing correctly in their respective dimension (i.e., providing information to separate students with different levels of θ). Interpretation of these two moderately (r = .68) correlated dimensions is coherent with items (i.e., PA) being better measures of either Manipulative skills, or Stability skills, as such we named these dimensions as *Manipulative-based Activities* (MA), and *Stability-based Activities* (SA), respectively (Table 33). Usage of Locomotion skills are likely highly prevalent and common across all activities, and thus no third factor emerged based on it. Surprisingly, all Athletics disciplines had higher loadings on the Manipulative factor than on the Stability factor; also, loadings patterns do not suggest that tactical skills might be a source of covariation among tactical-alike activities (e.g., Handball and Basketball).

Differential Item Functioning (DIF)

In the first stage of the analysis, the Throws (Athletics), Climbing and Roller Skating indicators were selected as anchors (adjusted p-values= 1.00). Subsequent sequential analysis with these indicators constrained to equality across groups revealed no DIF according to sex.

Discriminant and convergent validity

Inter-factor correlation between MA and SA was moderate to high (r=.68; Table 33). Table 35 displays the bivariate correlations between all variables in both PPLA-O modules, along with an additional BMI variable.

Table 33. Item parameters, inter-factor correlations and reliability for 2-dimensional graded response model

	Exploratory			Confirmatory							
				Standardized loadings			Slope parameters		Intercept parameters		
	Manipulative – based Activities	Stability-based Activities	Communalities	Manipulative – based Activities	Stability-based Activities	Communalities	a_1 (SE)	a_2 (SE)	d_1 (SE)	d_2 (SE)	d_3 (SE)
Races (Athletics)	**.66**	.34	.74	.84		.71	2.66 (.33)		4.47 (.46)	-1.24 (.23)	
Throws (Athletics)	**.87**	-.18	.66	.75		.57	1.95 (.42)		4.34 (.76)	-1.16 (.30)	
Jumps (Athletics)	**70**	.24	69	.83		69	2.53 (.52)		4.41 (.70)	-2.46 (.46)	
Handball	**.74**	.31	.84	.91		.83	3.78 (.45)		2.95 (.37)	-2.89 (.36)	
Football	**.81**	.09	.73	.86		.73	2.77 (.25)		2.00 (.23)	-1.44 (.21)	-5.40 (.43)
Basketball	**.85**	.14	.84	.92		.84	3.94 (.46)		3.91 (.45)	-2.79 (.35)	
Table Tennis	**.95**	-.25	.76	.78		.61	2.12 (.31)		3.77 (.44)	-1.54 (.25)	
Badminton	**.94**	-.07	.83	.88		.78	3.20 (.30)		4.51(.40)	-1.25 (.22)	-5.94 (.50)
Volleyball	**.70**	.23	.68	.82		.67	2.41 (.22)		4.27 (.33)	-1.20 (.18)	-5.95 (.47)
Floor Gymnastics	-.10	**.83**	.64		.64	.41		1.41 (.17)	1.97 (.18)	-2.29 (.19)	
Artistic Gymnastics	.28	**.67**	.62		.91	.83		3.75 (.95)	3.72 (.80)	-3.06 (.67)	-11.09 (2.43)
Climbing	.07	**.61**			.55	30		1.12 (.37)	2.53 (.41)	-1.01 (.27)	
Roller Skating	.18	**.64**	.48		.67	.45		1.53 (.32)	.76 (.25)	-2.48 (.33)	
Marginal Reliability	.88	.67		.89	.73						
Correlation	.43			.68							

SE – standard error
Note: salient loadings in each factor are bolded in the exploratory model

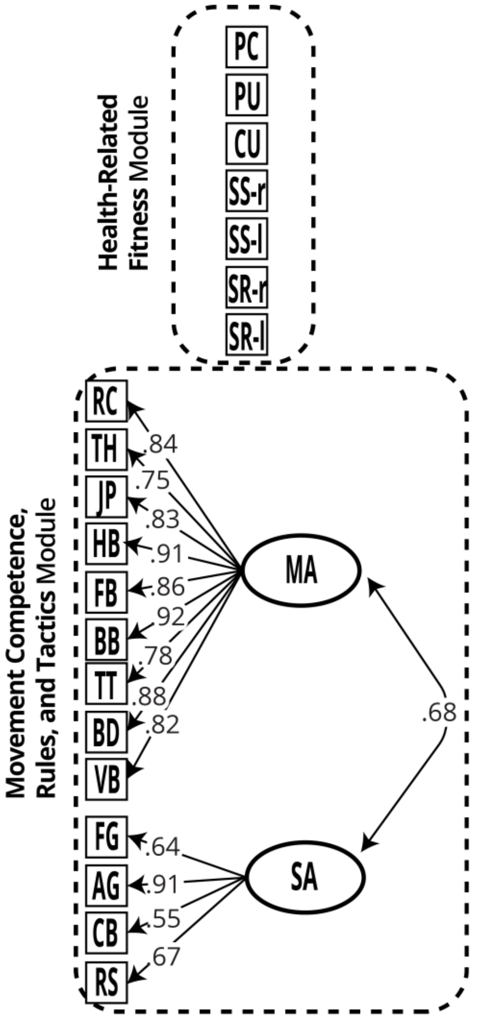


Figure 16. Portuguese Physical Literacy Assessment - Observation (PPLA-O) two modules, with estimated parameters for the Movement Competence, Rules and Tactics module (2-dimensional graded response model)

Legend: PC – Pacer, PU – Push-ups, CU- Curl-ups, SS-r – Shoulder Stretch (right), SS-l – Shoulder Stretch (left), SR-r – Backsaver Sit and Reach (right), SR-l – Backsaver Sit and Reach (left), RC – Races (athletics), TH – Throws (athletics), JP – Jumps (athletics), HB – Handball, FB – Football, BB – Basketball, TT - Table Tennis, BD – Badminton, VB – Volleyball, FG – Floor Gymnastics, AG – Artistic Gymnastics, CB – Climbing, RS – Roller Skating, MA – Manipulative-based Activities, SA – Stability-based Activities

Reliability and scoring

Both dimensions of the MCRT attained acceptable marginal reliability in the final solution (ρ_{xx}= .89 and .73, respectively; Table 33). Table 34 presents transformed intercept parameters (category threshold) which can be interpreted as transition-points between levels of proficiency for each activity (i.e., θ point at which there is a 50% probability to be scored in that category or higher; DeMars, 2013). Median values represent a heuristic cut-score between general proficiency levels (θ) in each dimension. I.e., a student with θ = -1.68 is likely to be transitioning from Non-Introductory to Introductory level on most Manipulative activities.

Table 34. Difficulty of each physical activity proficiency level transition point (threshold)

	b (difficulty)		
	Non-Introductory to Introductory	**Introductory to Elementary**	**Elementary to Advanced**
Manipulative – based Activities			
Races (Athletics)	-1.68	0.47	
Throws (Athletics)	-2.23	0.59	
Jumps (Athletics)	-1.74	0.97	
Handball	-0.78	0.76	
Football	-0.72	0.52	1.95
Basketball	-0.99	0.71	
Table Tennis	-1.78	0.73	
Badminton	-1.41	0.39	1.86
Volleyball	-1.77	0.50	2.47
Median	**-1.68**	**0.59**	**1.95**
Stability-based Activities			
Floor Gymnastics	-1.40	1.62	
Artistic Gymnastics	-0.99	0.82	2.96
Climbing	-2.26	0.90	
Roller skating	-0.50	1.62	
Median	**-1.19**	**1.26**	**2.96**

Table 35. Pearson and polyserial bivariate correlation matrix for PPLA-O variables

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**
1. Age										
2. MA	.23 [.15, .31]									
3. SA	.18 [.09, .26]	.79 [.75, .82]								
4. BMI	.05 [-.05, .14]	-.04 [-.13, .06]	-.13 [-.22, -.03]							
5. PACER	-.06 [-.14, .03]	.37 [.29, .44]	.31 [.23, .39]	-.25 [-.34, -.16]						
6. 90° Push-ups	.03 [-.05, .12]	.43 [.35, .50]	.35 [.27, .43]	-.18 [-.27, -.09]	.61 [.55, .66]					
7. Curl-ups	-.04 [-.13, .05]	.34 [.26, .42]	.27 [.19, .35]	-.19 [-.28, -.10]	.44 [.37, .51]	.41 [.33, .48]				
8. Shoulder Stretch (Right)[a]	-.06 [-.20, .09]	-.40 [-.51, -.27]	-.33 [-.45, .20]	-.18 [-.31, -.03]	-.04 [-.19, .11]	.00 [-.15, .15]	.05 [-.10, .19]			
9. Shoulder Stretch (Left)[a]	-.20 [-.31, -.07]	-.36 [-.47, -.25]	-.28 [-.39, -.16]	-.26 [-.37, -.14]	-.03 [-.16, .10]	-.05 [-.18, .08]	-.05 [-.17, .08]	.71 [.62, .78]		
10. Backsaver sit-and-reach (Right)	.00 [-.09, .09]	-.24[-.33, -.15]	-.04 [-.13, .05]	.01 [-.08, .11]	-.14 [-.23, -.05]	-.08 [-.17, .02]	-.05 [-.14, .05]	.28 [.14, .41]	.30 [.17, .41]	
11. Backsaver sit-and-reach (Left)	.00 [-.09, .10]	-.22[-.30, -.13]	-.01 [-.10, .08]	-.01 [-.10, .09]	-.14 [-.23, -.05]	-.05 [-.14, .05]	-.01 [-.11, .08]	.29 [.15, .42]	.29 [.16, .40]	.93 [.92, .95]

MA – Manipulative-based Activities; SA – Stability-based Activities; BMI – Body mass index; PACER – Progressive Aerobic Cardiovascular Endurance Run

[a] Polyserial correlations in these rows

Table 36. Movement Competence, Rules and Tactics mean scores stratified by sex for Manipulative-based Activities (MA) and Stability-based Activities (SA)

	θ (SD)			Transformed scores (SD)		
	Female	Male	Total	Female	Male	Total
MA	-0.30 (0.91)	0.41 (0.84)	0 (0.95)	48.1 (20.9)	64.4 (19.4)	54.9 (21.8)
SA	-0.15 (0.85)	0.21 (0.81)	0 (0.86)	40.6 (16.3)	47.5 (15.5)	43.4 (16.4)

Discussion

Our aims for the following studies were to a) develop the PPLA-Observation based on review of relevant conceptual frameworks and Portuguese PE syllabus practices; b) investigate the dimensionality structure of one of its modules - Movement Competence, Rules and Tactics module - through Item Response Theory (IRT) methods; c) test this structure for differential item functioning according to sex; d) establish support for convergent and discriminant validity, and score reliability for this module. A secondary aim was to draw inferences for scoring and criterion-referenced cut-scores mechanisms.

IRT Analysis of the Movement Competence, Rules and Tactics module

Dimensionality

Our results based on exploratory and confirmatory IRT analysis provide evidence in favor of a two correlated factor solution for assessing Movement Competence, Rules and Tactics, with evidence of measurement invariance (no-DIF) across sexes. This is somehow contrary to our initial conceptualization that proposed that six latent variables could be responsible for the variance in observed proficiency levels of activities: Locomotion, Manipulative, Stability, and Movement skills using Object, Rules, and Tactics. Items (activities) did not cluster according to different tactical typologies, movement forms, or subareas. Instead, the obtained solution suggest that their variance is driven according to competence in two different types of movement skills, namely, Manipulative movement skills, or Stability movement skills. Competence in Locomotor movement skills did not emerge as a latent factor explaining variance. This might be due to locomotor skills being transversally required in specialized skills in all evaluated activities in both dimensions (e.g., slide to hit a falling shuttlecock, or running and then jumping onto a trampoline) – as can also be seen in our content analysis of movement skills (Table 29).

Another unexpected finding was that two Athletics disciplines that were expected to load on the SA dimension (i.e., Running, and Jumps) – as specific skills for these activities are mostly locomotor and stability-based – presented higher loadings on MA. This might originate from how this group of activities (Athletics) is conceived and assessed within the PPES: rubrics for all disciplines are grouped together and assessed as a single activity, however, throughout the syllabus documents (Ministério da Educação, 2001a), the three disciplines appear mentioned as different activities. This might have led, inadvertently, to teachers reporting according to different standards. This requires scrutiny and caution in further developments of this tool.

Regarding *Tactics*, content analysis of the PPES revealed that up until the Elementary proficiency level, both movement skills and tactical requisites increase simultaneously, while it is during the transition to the Advanced level that tactical indicators take precedence (Table 29). As such, it is plausible that skill and tactical factors covary closely until the Elementary level, and only when students transition into Advanced levels is the tactical factor singularly driving variance in items – since movement skills factors cease or lower their effect at this level. However, in our sample most students were at, our below, the Elementary level in all activities (Table 30), which could preclude the mentioned disentanglement of variance between these factors. Also, since most tactical-heavy activities happen to be those that require manipulative skills, the MA factor might likely be accounting for variance due to tactical knowledge and application. Further studies with large-scale samples, with higher proportion of students in Advanced stages could evaluate these hypotheses and offer insights into this factorial structure.

Regarding *Rules*, variance caused by differing degrees of rule knowledge and application might be similarly overshadowed by movement skills and tactics. That is, a student might know and apply all rules from an activity, but absence of required skill and tactical factors might preclude him from advancing in proficiency level. Albeit aligned with an authentic assessment perspective, this invalidates measurement of this element using only the observed activities levels, and will likely require an external instrument (e.g., scale) to isolate.

Differential Item Functioning (DIF)

Items seem to function similarly for both sexes (i.e., no DIF), and as such, results can be meaningfully compared; despite suggestions in the literature pointing to the presence of bias when teachers observe MC (Faught et al., 2008; Hay & Donnelly, 1996), with the tendency to consider girl's competence in PA to be below average compared to boys of the same age.

Discriminant and convergent validity

The moderate to high correlation between MA and SA (*r*=.68; Table 33) is like that obtained by other batteries evaluating movement skills with the same conceptualization in older children and adolescents in a Portuguese sample (*r* = .64; Luz et al., 2016). Due to the strength of this correlation, a general motor ability underlying results in both factors is tenable (Burton & Rodgerson, 2001), and could be further investigated through second-order or bifactorial modelling (Brown, 2015; Reise, 2012). Despite this, discriminant validity is still ensured, with inter-factor correlations below .85 (Brown, 2015).

In general, correlations observed in our study among MA and SA, and correlates like sex, age, BMI and fitness (Table 35) were coherent with those found in the literature regarding movement skills in adolescents, strengthening the evidence for construct validity of the MCRT. Boys had higher scores than girls in both dimensions (Table 36), with the difference being smaller in stability skills (Luz et al., 2017; L. P. Rodrigues et al., 2019). Values for the correlation of age and scores on both dimensions (*r* = .23 [.15, .31], and *r*= .18 [.09, 26] , MA and SA, respectively) were similar to those reported in a meta-analysis by (Barnett et al., 2009). Also, similarly, to results reported in Barnett and colleague's review, we found an inverse correlation for BMI and SA scores (*r* = -.13 [-.22, -.03]). Cardiovascular and muscular endurance were also correlated with both scores, in similar magnitude as in previous studies (Luz et al., 2017; L. E. Robinson et al., 2015). Finally, despite inconclusive results in reviews (Cattuzzo et al., 2016; L. E. Robinson et al., 2015), we also observed negative correlation between all flexibility indicators and scores in both dimensions; this correlation was lower in regards to SA which is plausible with the idea that stability-based activities require higher ranges of motions. The role of flexibility warrants further scrutiny since our results pointed to a mostly negative correlation with other fitness indicators; especially the sit-and-reach indicators might be collapsed, since their correlation suggested that they are statistically equivalent ($r > .85$).

Reliability and scoring

Use of a sub-score for each of the identified dimensions of the MCRT seems plausible given the evidence of sub-score reliability (ρ_{xx} = .89, and .73). We suggest a transformation so that these scores provide an easy 0 to 100 interpretation – like other scores in PPLA. For this transformation, the median θ score estimated for the transition from Elementary to Advanced level (θ = 1.95, and 2.96, respectively; Table 34) can be used as the upper bound, and the estimated θ score for a student with the lowest possible levels in all activities as a lower bound (θ_{MA} = -2.38, and θ_{SA} = -2.27, not shown). As an example,

$$X_{MA} = \frac{\theta + 2.38}{(1.95 + 2.38)} x100$$

with X being the new 0-100 score, and θ the estimated θ_{MA} score.

Since these scores require complex computations, the effectiveness and precision of simpler options (e.g., sum-scores) should be investigated in the future, given our concern for feasibility.

Reliability has been widely established for the HRF module protocols. We suggest that results from each protocol should be similarly transformed using the values reported by FITescola®'s Athletic Profile, based on sex and age, as upper bound. In this manner, a 0 to 100 criterion-referenced score can also be obtained.

Strengths and limitations

One of the major strengths of the PPLA-O is that it uses data routinely collected by PE teachers to frame the evaluated elements into a common reference frame of Physical Literacy. Its content validity is also maximized by making use of 1) HRF protocols that have been chosen and adapted with the PE context in mind (FITescola®), and 2) data referent to proficiency levels in diverse physical activities that were chosen to figure in the Portuguese syllabus by curriculum design experts. It also evaluates movement skills – and inherent tactical actions – within tasks and environmental constraints that will be common to activities practice outside of PE, providing a chance for an authentic, ecologically valid, and highly feasible assessment. Nonetheless, further efforts could study content and face validity with students, and other educational stakeholders, as well as with motor development specialists to provide another layer of validity evidence.

Another strength rests in the use of IRT methodologies to analyze construct validity and reliability. Due to the same ecological approach mentioned above, missing data will always assume large proportions, since different students' needs will dictate that each class will work on and assess different activities. IRT's algorithms were specifically designed to work with categorical data and are robust to missing data, using all information available to estimates parameters that also have higher degrees of invariance from sample to sample (Bock & Gibbons, 2021; Reise & Revicki, 2015). In this way, students with just a few assessed activities will still be able to be scored. However, large amounts of missing data still posed a limitation regarding assessment of absolute fit of the models through statistical tests equivalent to chi-square (i.e., C2; Cai & Monroe, 2014) and derived relative fit indexes (root mean square error of approximation).

One limitation of this study lies in the unknown inter and intra-observer reliability of PE teachers while assessing both the fitness protocols and activities levels. We would argue that many factors could contribute to higher reliability, including 1) extensive training during initial teacher's education, 2) clear and task-specific rubrics for each activity and level available in the syllabus (Brookhart, 2013), 3) specific fitness protocols with detailed instructions and resource for application, 4) collaborative training and observation opportunities within schools, and 5) each assessment is based on multiple in-context observations. Despite this, these inferences require further scrutiny and empirical validation, since process-oriented assessments are more susceptible to bias caused by different levels of observer's expertise (e.g., Griffiths et al., 2018; Schoemaker et al., 2012). As part of this effort, demographic data on PE teachers, along with teaching experience and other relevant variables should also be collected to better understand assessment patterns, which we did not do during these studies.

A final, more general limitation is concerned with the timeframe of this study. All data collection was done amongst lockdowns imposed by the COVID-19 pandemic. This limited the number and quality of activities assessed by PE teachers (especially those involving physical contact like wrestling or acrobatic gymnastics); and might have imposed additional unforeseen limitations on these results. As such, these results should be replicated in a larger, more representative sample of students in more normal PE circumstances, which will enable a deeper insight in the *Tactics* element as we discussed.

Conclusion

Throughout this article, we detailed the development of the PPLA-O, an instrument that assess the physical and part of the cognitive domains of PL in grade 10 to 12 adolescents (15-18 years). It is composed of two modules, 1) Health-Related Fitness (HRF), and the 2) Movement Competence, Rules and Tactics (MCRT), that integrate observational data from PE teachers into a common frame of criterion-referenced PL (Figure 16). The former makes use of data collected through widely validated FITescola® assessment protocols, while the latter makes use of teacher-reported data collected in a wide range of activities and movement pursuits to measure movement competence and inherent cognitive skills (*Tactics* and *Rules*). We also gathered initial evidence supporting construct validity and score reliability of the MCRT module through IRT multidimensional models, with a final two-dimensional solution (Manipulative-based Activities, and Stability-based Activities). Further studies should focus on assessment of inter and intra-rater reliability and criterion-related validity. This highly feasible instrument can be used to provide students with feedback on their PL journey, along with the other instrument of PPLA (PPLA-Q).

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. All methods were performed in accordance with the relevant guidelines and regulations. 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 instrument 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, Dean Dudley, and of all the tireless PE teachers and students which participated in these studies. We also extend our gratitude to all R developers and community, who continuously devout their time for the benefits of others and science. The lead author would also like to thank his co-authors for their ever-present guidance and support during his PhD project.

CHAPTER 6 – Full PPLA Construct Validity and Reliability

Portuguese Physical Literacy Assessment for adolescents (15-18 years) from grades 10-12: validation using Confirmatory Factor Analysis and Confirmatory Composite Analysis

João Mota, João Martins, Marcos Onofre

In preparation for submission

Abstract

Background: Aims of this study were to assess the construct validity and reliability of the full Portuguese Physical Literacy Assessment (PPLA) integrating both of its instruments (a questionnaire, and a tool using teacher-reported data) to measure the four domains of Physical Literacy (PL); this also included the assessment of the adequacy of a total PL-score and respective subscales by domain. We also sought to discuss conceptual and practical implications of reflective *versus* formative measurement for PL.

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. Bifactor model-based indices ($\omega_,$ 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 CFI= 97, Robust RMSEA = .05 [.04,.06], SRMR =.04), while CCA resulted in best absolute fit for single first-order composite models (d_G, d_L, and SRMR below or borderline of their 95% critical value, in both the optimal and unit weighted models). The tenability of both paradigms to assess PL is discussed. Through a reflective paradigm, a total PL score should not be used in isolation (ECV = .49, ω_H = .71, lower than recommended .80); subscales for each PL domain attained acceptable score reliability except for the cognitive one (ω_S = .76, .82, .80, and .60, for the physical, psychological, social, and cognitive sub-scores, respectively), and dimensional uniqueness, except for the psychological one (ECV_{SS} = .71, .23, .61, .98).

Conclusions: Present results provide evidence that a general trait of PL is responsible for a considerable amount of variance in all indicators – albeit with insufficient strength to be interpret–ted in isolation - with demarked domain-specific variance. We advise 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 to quickly compare different classes and schools in low-stakes settings, the latter allows for more meaningful interpretation of students PL profiles and needs. Caution is advised on using the psychological sub-score in high-stakes settings, as most of its variance is absorbed by the PL general trait, which warrants further scrutiny. The use of a bifactor measurement model for further research efforts using the PPLA is recommended. We encourage further research into the tenability and implications of PL measurement using both the reflective and formative paradigms, as both seemed tenable according to our results. Overall, evidence supported the construct validity and reliability of the PPLA for its intended use an integrated tool to measure Physical Literacy as a multidimensional construct in 15 to 18 years old Portuguese students in a PE setting.

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

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). This multidimensional concept is argued as the foundation for the physical education, sport, and public health agendas (Organization for Economic Co-operation and Development, 2018; UNESCO, 2015; World Health Organization, 2020; Ydo, 2021). Given this prominent role, in the last few years multiple efforts have been made to develop and refine measuring tools that support learning towards a meaningful movement journey (Barnett et al., 2020; Cairney et al., 2018; S.-T. Chen et al., 2020; Gandrieau et al., 2021; Gunnell et al., 2018; Mohammadzadeh et al., 2021; Physical Literacy for Life Consortium, 2021; Sum et al., 2018).

Similarly, the Portuguese Physical Literacy Assessment (PPLA) was developed as a tool composed by 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 the PL of grade 10 to 12 (15-18 years) students. It was inspired by the Australian Physical Literacy Framework (APLF; Sport Australia, 2019) – a conceptual model of PL learning composed of 30 elements across 4 domains (physical, psychological, social and cognitive) – and by the outcomes and didactic philosophy of the Portuguese PE syllabus (Ministry of Education [Ministério da Educação], 2001a, 2001b).

Both instruments have previously gathered evidence supporting construct validity and reliability at element-level using Item Response Theory models (Mota et al., 2022c, 2022b, 2022a). However if PL is to be understood as a holistic framework to understand human movement and PA participation, construct validation and reliability evidence must be gathered to support the intended integrated interpretation of all four domains (Finch & French, 2019).

The dimensionality of this tool can be assessed through the means of structural equation modelling (SEM), where two main approaches can be taken depending on the auxiliary theories assumed to underlie measurement (J. R. Edwards & Bagozzi, 2000; Henseler, 2021; Sarstedt et al., 2016): a) reflective measurement, or b) formative measurement. Previous studies on PL measurement have always implicitly assumed a multidimensional reflective view, modelling PL as a) correlated factors (Cairney et al., 2018; Longmuir et al., 2015; Sum et al., 2018), or b) higher-order factor (Cairney et al., 2019; Gunnell et al., 2018). In a following section, we will provide an overview of ontological and conceptual issues inherent to both approaches.

Similarly, before scores from this test are used to inform student's and teacher's practice, or in research settings, a concurrent effort must be made to assess the validity of the derived scores according to this intended use (American Educational Research Association et al., 2014). When SEM is used, estimated scores can be directly used to explain antecedents or outcome variables of interest (Henseler, 2021; R. B. Kline, 2016). However, for applied used in PE, it becomes essential that timely, easy to obtain scores are available. These can either be refined, more precise, scores derived from confirmatory analysis; or coarser, easier to obtain scores using summation (i.e., sum-scores) (DiStefano et al., 2009; Grice, 2001). Since each measuring approach presents slightly different scoring implications, we address these issues simultaneously in the next section.

Our initial model for PPLA (Mota et al., 2021) 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 in each of the domains would be plausibly independent from each other. E.g., one could conceive that an increase in Cognitive-related skills or knowledge would not be simultaneous with an increase in Psychological-related attitudes or perceptions. 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, the aims of this paper were to a) establish evidence supporting construct validity and reliability of the PPLA, integrating measures derived from the PPLA-Q and PPLA-O, by comparing results drawn from factor-based (reflective) methods, and composite-based (composite-formative) methods; and based on this b) assess adequacy of using a PL total-score, and respective subscales. As secondary research aim, we aimed to discuss implications of the different methods on the wider conceptual understanding of PL.

Literature Review

Measuring model conceptualization

While most construct validation studies and scale development done in education and social sciences has been under the reflective paradigm (Bollen, 2002), many disciplines are now starting to explore the formative paradigm (Brown, 2015; Diamantopoulos, 2008). This has generated an ongoing debate and research for the latter's impact and adequacy for measurement (J. R. Edwards, 2011; Evermann & Rönkkö, 2021; Henseler, 2018; Henseler et al., 2014; Rigdon, 2016; Rigdon et al., 2017). Each of these conceptualizations bears different ontological and practical premises.

Reflective measurement

Ontological and conceptual issues

Measurement under the reflective conceptualization assumes a realist perspective (Borsboom et al., 2003), wherein the construct represent a real entity tangible through its measures. It is represented through a common factor underlying all observed measures (effect indicators) and responsible for their covariance (Lord & Novick, 1968). This is the approach at the heart of Classical Test Theory (CTT) (also named True Score Theory) (McDonald, 1999) and Item Response Theory (Bollen & Bauldry, 2011; Embretson, 1996). According to this, a causal relationship is assumed between the construct and its measures, with a change in the construct expected to cause equivalent changes (after accounting for error) in all its measures (Brown, 2015). The notion of measurement error is explicitly modeled at indicator-level (i.e., residuals) which is then used in factor-based Structural Equation Modelling (SEM) methodologies to provide more accurate depictions of correlation among constructs (i.e., through disattenuation; R. B. Kline, 2016)

Factors (i.e., latent variables) are assumed unidimensional entities, composed of multiple interchangeable measures of the same phenomenon; and elimination of any of these does not alter the meaning of the factor (Brown, 2015). Testing of this dimensionality structure is usually performed using covariance-based methods like Confirmatory Factor Analysis (CFA), by comparing different *a priori* specifications based on substantive theory (Brown, 2015). Constructs, however, are unlikely to adhere to strict unidimensionality, whether intentionally - because the researchers wanted to capture different facets of the construct through to the use of subscales – or unintentionally – due to method factors (Brown, 2015), responses bias (DeVellis, 2017) or even item difficulty/popularity (Sijtsma & Ark, 2021). As such, multidimensional model can be estimated, with the most common one being the a) correlated first-order factor models, and b) higher-order models (Cho, 2016; Rindskopf & Rose, 1988). Other options include the estimation of correlated residuals between indicators to accommodate covariation unrelated to the construct at hand (Brown, 2015).

Correlated factor models freely estimate correlations among construct composed of interchangeable indicators (Figure 17). If the pattern and magnitude of correlations is substantial, and there is a strong conceptual justification , it is then tenable to estimate a higher-order model to account for this correlation (R. B. Kline, 2016; Law et al., 1998). The two most common options for this are the hierarchical model and the bifactor model.

In both cases, the different first-order factors are conceptualized as being interchangeable indicators of the same, more abstract, construct (N. Lee & Cadogan, 2013). One of the main differences lies in the way the effect of the higher-order factor is conceived upon observed indicators: in the hierarchical model, the first-order factor fully mediates the effect of the higher-order construct into the indicators (Figure 17); while in the bifactor model, direct effects are estimated from the higher-order construct into the indicators (F. F. Chen et al., 2006; Yung et al., 1999). Bifactor model allows the analysis of group-factors (equivalent to the error terms on the hierarchical model, denominated disturbances, Cho, 2016) which represent variance that is specific to a set of indicators, over and beyond that of the general-factor (i.e., the higher-order construct); and modelling of independent relationships of the group and general factors on outcomes of interest (F. F. Chen et al., 2006; Ward et al., 2015). Because of these nuances, the bifactor model enables a more robust study of dimensionality and on the interpretation of scores, which we will describe below.

Further iterations of the bifactor model have been developed as this approach gains traction within measurement assessment of many disciplines; namely non-symmetrical bifactor models (Figure 17) - bifactor S-1 or S·I-1 models (Eid, 2020; Eid et al., 2017). These can be used when the group factors are not interchangeable indicators (i.e., *random effects*), and are structurally different (i.e., *fixed effects*) – as is usually the case when researchers conceptualize facets or domains. These allow for estimation of correlation among group factors – which in the canonical bifactor model are constrained to zero by definition (Holzinger & Swineford, 1937) – and use a group factor or general measures of a construct to establish a reference frame based on previous theory, instead of allowing for data's intricacies to do so by collapsing specific indicators or entire group factors – anomalies which are usual in the estimation of canonical bifactor models (Eid et al., 2017).

Scoring implications

Summation of indicators to obtain an approximation of the position of each individual on the latent variable is one of the most common uses for scales and measurement instruments in applied settings (c.f. Avila et al., 2015; McNeish & Wolf, 2020; Nunnaly & Bernstein, 1994). This assumes that all indicators are equally weighted (i.e., unit-weighted), and represents one of many unrefined approaches to derive weights (c.f., DiStefano et al., 2009). Other options include optimal weighting (i.e., using loadings, or factor coefficients derived from CFA), and refined approaches which use all available information to produce factor scores. However, due to specific nuances of the common factor model (i.e., factor indeterminacy, Grice, 2001), extracted factor scores might be not be adequate for individual-level analysis or interpretation. Despite being a simple, viable and robust alternative in many settings (Dana & Dawes, 2004; McDonald, 1999), sum-scores' adequacy rests on the dimensionality of the scale or instrument in question (Reise, Bonifay, et al., 2013).

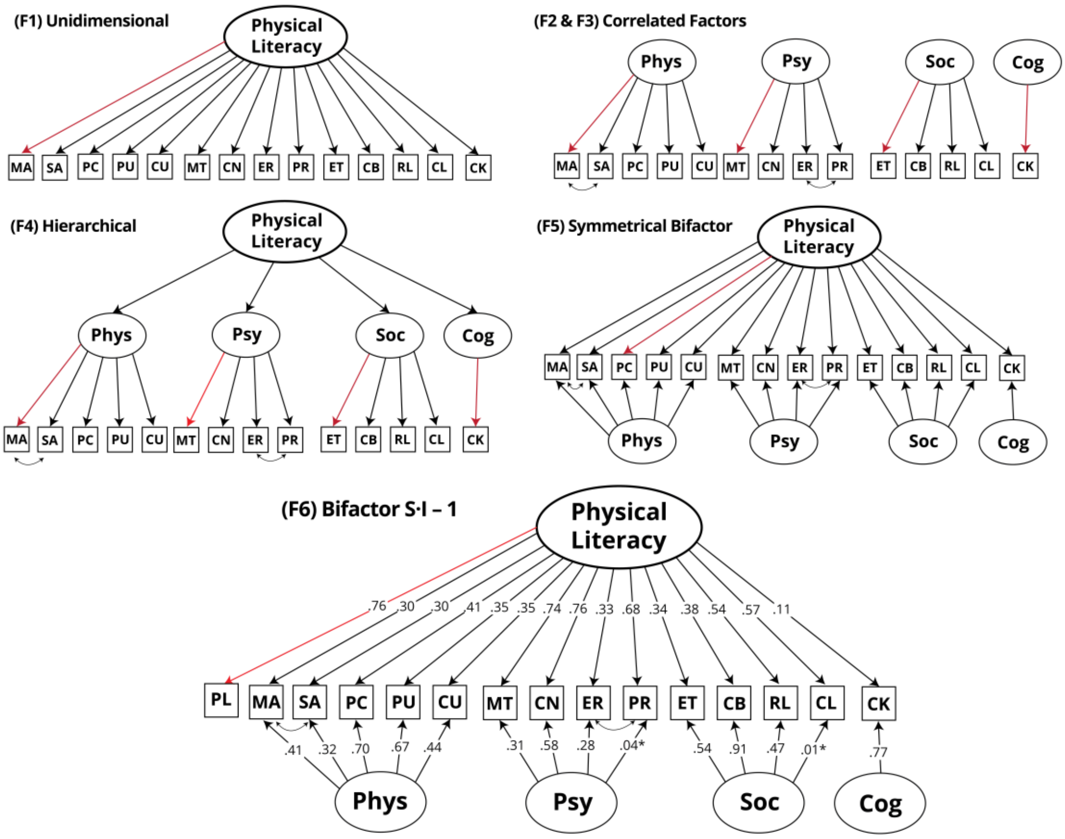


Figure 17. Factor-based models estimated in the study, standardized factor loadings are presented in F6

Legend: PL – Physical Literacy, MA – Manipulative-Based Activities, SA – Stability-Based Activities, PC – PACER, PU – Push-ups, CU – Curl-ups, MT – Motivation, CN – Confidence, ER – Emotional Regulation, PR – Physical Regulation, ET – Ethics, CB – Collaboration, RL – Relationships, CL – Culture, CK – Content Knowledge. Marker indicators are colored red; error terms are omitted for clarity; F2 is equal to F3 without freely estimated covariances between indicators.

When there is evidence of a strong factor underlying the results (with high indicator loadings), then the correlation of the derived total sum-score with the *true* factor score will be high. Otherwise, the use of scores for each of identified domains (as in the correlated factors model) might be justified (Reise et al., 2000). To assess the tenability of a total score, along with eventual subscales scores, a bifactor model can be fit, and various model-based indices derived (Reise et al., 2010; Reise, Bonifay, et al., 2013; Rodriguez et al., 2016). These allow the researcher to evaluate the amount of variance accounted by the general factor and group-factors (specific domain), and whether their strength warrant statistical or empirical interpretation. It also allows to test whether the use of a unidimensional model in SEM settings would adequately convey the general trait, or whether a bifactor model should be fit.

Formative measurement

Ontological and conceptual issues

Two different views exist within this perspective according to the conceived ontology of the construct, causality, error, conceptual unity of its indicators, and subsequent estimation: a) causal indicators, b) composite indicators (Bollen & Bauldry, 2011). We will firstly address their differences, and later describe their similitudes.

Causal indicators

Similarly to the reflective approach, a latent variable is still posited to *exist*, however, instead of it accounting for the variation in its indicators, causal indicators form, or influence the latent variable (Brown, 2015; Diamantopoulos, 2008). Included indicators of a construct must share conceptual unity (i.e., pertain to the same concept), and must cover all possible content domains of the construct (Bollen & Diamantopoulos, 2017). Measurement error is conceived at the construct-level through the estimation of a disturbance term, and it is posited to account for possible unincluded indicators or facets (Bollen & Bauldry, 2011; Diamantopoulos & Siguaw, 2006).

Estimation of this type of models can only be achieved in covariance-based methods (i.e. CFA or factor-based SEM) (Sarstedt et al., 2016) by specifying two emitted paths from each causal-formative variable (Bollen, 1989) – either intended outcomes, or direct reflective measures of the same construct (creating a multiple indicators-multiple causes model).

Composite indicators

Conversely to both earlier approaches, constructs defined by composite indicators (i.e., emergent variables, composites or artifacts; Henseler, 2021) have no ascribed existence independent of measurement (constructivism) and thus, are created for mostly analytic purposes (operationalism) (c.f., Borsboom et al., 2003; Edwards, 2011). Included indicators need not share conceptual unity (Bollen & Bauldry, 2011), and are assumed to completely define the construct (Henseler, 2021), since no error term is estimated– neither at item, nor construct-level (Figure 18).

Composite estimation is best done through variance-based methods (i.e., composite-based) (Sarstedt et al., 2016), and one of the its many available estimators – with the most studied estimator being Partial Least Squares (PLS; Schuberth et al., 2018). In this framework, identification of the composite requires only a connected construct (i.e., non-isolation condition; Schuberth et al., 2018).

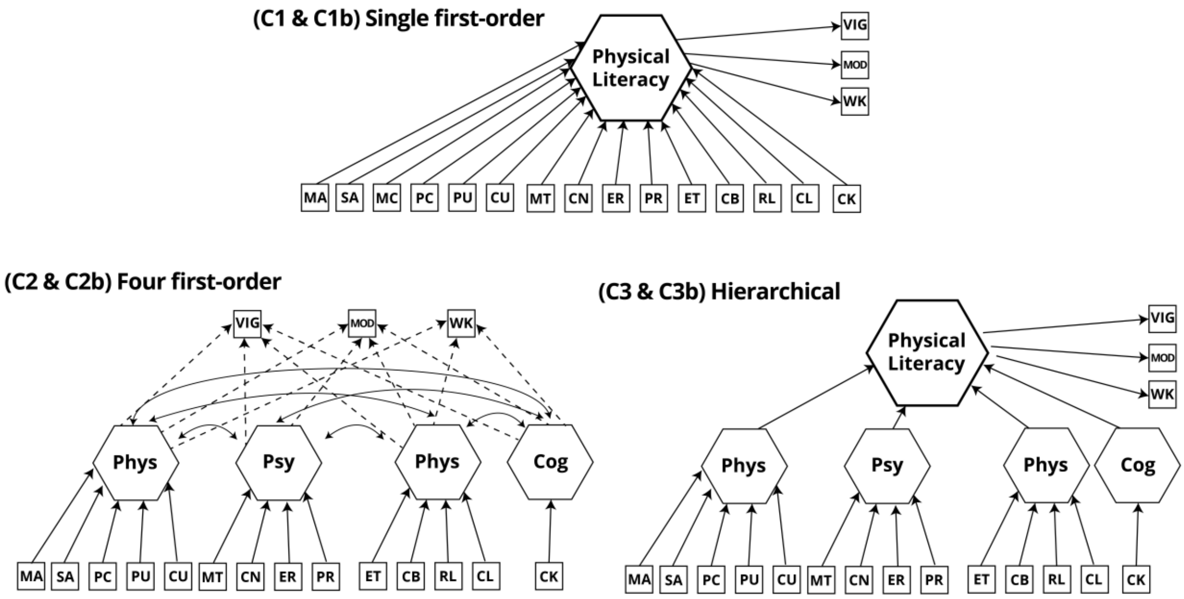


Figure 18. Composite-based models estimated in the study

Legend: PL – Physical Literacy, MA – Manipulative-Based Activities, SA – Stability-Based Activities, PC – PACER, PU – Push-ups, CU – Curl-ups, MT – Motivation, CN – Confidence, ER – Emotional Regulation, PR – Physical Regulation, ET – Ethics, CB – Collaboration, RL – Relationships, CL – Culture, CK – Content Knowledge

Similitudes

Despite the aforementioned differences, both formative approaches share the fact that constructs are estimated as addictive - i.e., a linear combination of weighted indicators (Law et al., 1998). These constructs are multidimensional by definition, formed by indicators that each capture a non-redundant facet of the concept; as such, removal of any indicator will change the meaning of the estimated construct (Bollen & Bauldry, 2011; Diamantopoulos & Siguaw, 2006; Henseler, 2021). These constructs are not limited to a single first-order conceptualization, and can take a similar multitude of structures implied by the reflective measurement paradigm, including a) correlated first-order and b) higher order models (c.f., Schuberth et al., 2020).

Theoretically, no degree of correlation among indicators is required (Bollen & Diamantopoulos, 2017; Henseler, 2021), since a change in one indicator might not be accompanied by change in all indicators (Avila et al., 2015); in practice, high levels of correlations among indicators can cause issues of multicollinearity that difficult interpretation of parameters (Cenfetelli & Bassellier, 2009; Rigdon, 2012).

Scoring implications

The main difference regarding scoring in formative models is that the score loses its conceptualization as a position on a posited trait, and is rather equated to an index – a summary of data reduction (c.f., Borsboom et al., 2003). As such, a simple sum-score might provide a parsimonious estimate, at the cost of distinct information in each indicator (Coltman et al., 2008), especially when correlations among indicators are low (Howell et al., 2007). As in the reflective model, usage of differential weights might be an option to address this issue. An advantage of composite-based methods is the inherent determinacy of construct scores (Esposito Vinzi et al., 2010) – since they represent linear combinations of weighted estimates. As such, weighting indicator scores by their regression weights will be equivalent to the estimated construct scores.

Other implication, albeit disputed (Bollen, 2007), is that of susceptibility to interpretational confounding in weight estimates – i.e., difference in the weights attributed to each indicators depending on the variables used to identify the model (Aguirre-Urreta et al., 2016; J. R. Edwards & Bagozzi, 2000; Guyon & Tensaout, 2016; Howell et al., 2007). This is also argued to compromise theoretical development (Wilcox et al., 2008) and meaningful interpretation of the construct (Bagozzi, 2011), since the same construct might change depending on the nomological network into which it is inserted. In order to resolve this issue, some researchers suggest the use of predetermined weights based on theory (Avila et al., 2015; N. Lee & Cadogan, 2013), or revert to the simpler solution: unit weights (Cadogan & Lee, 2013; Henseler, 2021; Rigdon, 2012).

Methods

Participants

A sample of 521 (58% female) grade 10-12 students (M_{age} = 16, SD= 1 years) from 6 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 (Mota et al., 2022c, 2022a). 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. Student sample was 58% composed of female, with a mean age of 16 years (SD =1). A participant flow is available in Figure 19. A minimum sample size of 275 was initially chosen based on a power analysis conducted in WarpPLS software (Kock, 2020), using the Inverse Square Root Method (Kock & Hadaya, 2018), for a minimum absolute path coefficient of .15 and power of .80. This sample size also conforms to guidelines posed for CFA (Mokkink et al., 2018; Prinsen et al., 2018; Wolf et al., 2013). Data collection took place between January and March 2021. PPLA-Q was self-administered (students) both in a pen and paper and online format, in the presence of the lead investigator, PPLA-O was self-administered (PE teachers) using a spreadsheet.

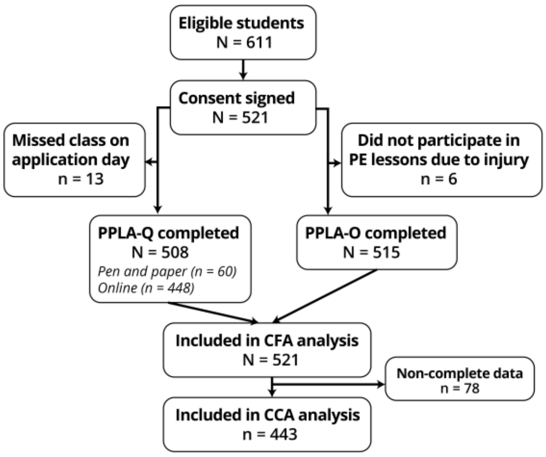


Figure 19. Portuguese Physical Literacy Assessment Validation studies participants flowchart; CFA – Confirmatory Factor Analysis; CCA – Confirmatory Composite Analysis

Measures

PPLA measures

Physical domain

The Physical domain of the PPLA was assessed through the PPLA-O (Mota et al., 2022a), an instrument that integrates teacher-reported data into the same PL framework as other domains. 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), with previous evidence for its construct validity and reliability (Mota et al., 2022a): estimate of empirical reliability of .89 and .73; 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-posterior (EAP) were transformed into a 0-100 score.

Health-related physical fitness module included seven indicators, all assessed through existing FITescola® protocols (Direção-Geral da Educação & Faculdade de Motricidade Humana, 2015), in three major subareas: 1) *Cardiorespiratory endurance* was assessed through the 20-meter Progressive Aerobic Cardiorespiratory Endurance Run (PACER), using the number of laps completed; 2) *Muscular Endurance* was assessed through the Curl-Up, and the 90º Push-Ups protocols, both in number of executions; 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 teacher's initial formation curriculum.

Psychological domain

The Psychological domain included four indicators assessed through the PPLA-Q: *Motivation*, *Confidence*, *Emotional Regulation*, and *Physical Regulation*. All these indicators consisted of the total summed score of responses in each respective scale (composed of seven, nine, seven and eight items, respectively), and then transformed into percentage of maximum points (0-100 score) to normalize different number of items in each scale. All scales have been calibrated through Mokken Scale Analysis (e.g., Sijtsma & van der Ark, 2017) using non-parametric Item Response Theory models (IRT), and have shown evidence supporting adequate score reliability (Molenaar-Sijtsma's ρ of .83 to .94; Molenaar & Sijtsma, 1984) and construct validity in this sample (Mota et al., 2022c): dimensionality (Loevinger's H of .47 to .66; Molenaar, 1990), discriminant validity (disattenuated correlations between subscales of .27 to .73) and convergent validity. To facilitate interpretation, these and all measures in the Social domain below, were transformed in a 0-100 score.

Social domain

The Social domain included four indicators: *Culture*, *Ethics*, *Collaboration* and *Relationships*. These indicators followed the same logic as the those of the Psychological domain presented above, using a total summed score across seven, and six items (for the last three mentioned subscales), respectively. Previous validation using Mokken Scale Analysis (Mota et al., 2022c) resulted in adequate score reliabilities (ρ of .86 to .91) and construct validity (H of .54 to .64; disattenuated correlations of .18 to .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, which was then transformed to percentage metric to facilitate interpretation (Mota et al., 2022b). This calibration gathered evidence on construct validity and score reliability of the test (marginal reliability of .60) to distinguish students with descriptive (*Foundation*) knowledge from those with higher analytical knowledge (*Mastery*). Factor scores derived with expected a-posterior (EAP) were transformed into a 0-100 score.

Self-reported physical activity

The short form of the *International Physical Activity Questionnaire* (IPAQ-SF; Craig et al., 2003), as used in the National Food, Nutrition and Physical Activity Survey (Lopes et al., 2017) was used to obtain weighted estimates of each intensity of physical activity per week (MET/min/week). Data cleaning and coding procedures follow the recommendations of the IPAQ Research Committee (2005). No total summed score was used since this instrument has shown different validity across intensities (Kim et al., 2013; P. H. Lee et al., 2011); and since it is tenable that different intensities might interact differently with the different domains of PL.

Statistical analysis

All statistical analysis used RStudio 1.4.1106 (RStudio Team, 2020), with *R* 4.0.1 (R Core Team, 2020). Missing data and descriptive statistics were computed using the packages *naniar* (Tierney et al., 2021) and psych (Revelle, 2021), and are available in Table 37. To test whether the Missing Completely at Random (MCAR) missing mechanism was plausive, Little's test (1988) was employed, with a resulting statistic of χ^2 (593) =791.65 , p < .001 with 38 missing patterns, resulting in evidence against MCAR. Since missing data most likely originates from students who missed class on the day of measures' application, it is tenable to assume that data is Missing at Random.

Data was also screened for univariate and multivariate normality with Shapiro-Wilk's (Shapiro & Wilk, 1965) and Mardia's test (1974), through the *MVN* package (Korkmaz et al., 2014) – shoulder stretch had to be removed from the later test to achieve convergence, due to being a binary indicator. Results of the univariate tests are presented in Table 37, for each measure; these results, complemented by statistically significant Mardia's statistics (Mardia skewness = 2739.39, p <.001; kurtosis = 13.33, p<.001) render any normality assumption untenable.

Data was screened for multivariate outliers using Minimum Covariance Determinant approach (Leys et al., 2019) through the *Routliers* package (Klein & Delacre, 2021), highlighting 69 multivariate outliers. Sensitivity analysis revealed no differences in model fit or parameters in the main analysis, and so outliers were kept. Bivariate Pearson and polyserial correlations between measures were obtained in *polycor* (Fox, 2019) and are available in Table 38. For factor-based analysis, the *Push-ups* indicator was multiple by a factor of 2 to rescale its variance.

Table 37. 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)	2071 (2084.2)	1440.0	.85	<.001
Moderate[a]	26 (5.0)	1071 (1122.2)	720.0	.80	<.001
Walking[a]	26 (5.0)	767.5 (950.1)	396.0	.76	<.001
PPLA-O Measures					
PACER	22 (4.2)	49.5 (22)	44.0	.93	<.001
Push-ups	26 (5.0)	18.1 (9.6)	18.0	.93	<.001
Curl-ups	23 (4.4)	48.6 (21.7)	45.0	.91	<.001
Shoulder Stretch (frequency of achievement)					
Right	83 (15.9)	95%		.21	<.001
Left	83 (15.9)	89%		.37	<.001
Sit and Reach (cm)					
Right	85 (16.3)	30.7 (8.3)	31.0	.99	.009
Left	84 (16.1)	30.2 (8.2)	31.0	.99	.006
Manipulative-Based activities[b]	6 (1.2)	54.9 (21.8)	55.4	.99	.076

Table 37. 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
Stability-based activities[b]	6 (1.2)	43.4 (16.4)	43.7	.98	<.001
PPLA-Q Measures					
Content Knowledge[b]	13 (2.5)	68.8 (15.2)	70.0	.99	<.001
Motivation[b]		74.9 (14.8)	77.1	.97	<.001
Confidence[b]		68.4 (16.8)	68.9	.98	<.001
Emotional Regulation[b]		69.9 (14.6)	71.4	.98	<.001
Physical Regulation[b]		75.1 (12.2)	75.0	.98	<.001
Culture [b]		58.7 (19.5)	57.1	.98	<.001
Ethics[b]		81.5 (11.9)	83.3	.92	<.001
Collaboration[b]		85.0 (11.4)	86.7	.94	<.001
Relationships[b]		77.7 (13.5)	80.0	.97	<.001

PACER – Progressive Aerobic Cardiovascular Endurance Run
[a]MET/week; [b]Maximum score = 100

Confirmatory factor analysis

Since PPLA is based on the Australian Physical Literacy Framework (APLF; Sport Australia, 2019), a clear rationale for factorial structure has already been laid out (Mota et al., 2021). We employed CFA to test the previously hypothesized model structure against other tenable competing models presented in the literature (Cairney et al., 2019; Gunnell et al., 2018), assuming a reflective measurement model. Six models were estimated (Figure 17): F1) unidimensional, F2 and F3) correlated first-order factors, F4) second-order, F5) canonical (symmetric) bifactor, F6) bifactor S·I-1. All models were estimated in *lavaan* 0.6.9 (Roseel, 2012), using raw data as input. All variables were specified as continuous, and given the violation of multivariate normality, robust maximum likelihood estimation (MLR) with robust "Huber-White" standard errors (Huber, 1967) and a scaled test statistic (equivalent to Yuan-Bentler T2*; Yuan & Bentler, 2000) was used. Given the existence of missing data on many variables, and assumption of MAR, Full Information Maximum likelihood (FIML; Arbuckle, 1996) was used to estimate unbiased parameters (Dong & Peng, 2013).

In all models, the metric of latent factors was fixed by using the first indicator as marker. Error covariances were constrained to zero, unless otherwise specified. In all multiple factors models, the Cognitive factor was specified as a single indicator factor, and its error variance was constrained to _(1- reliability)*variance*(indicator)_ (Gana & Broc, 2019; R. B. Kline, 2016). Estimation of the last three models (F4-F6) used bounded parameters to stabilize the solution (Jonckere & Roseel, 2021). Initial estimation of model F5 resulted in a negative variance (Heywood case); changing the marker indicator resolved this issue. For the sixth model, a 5-point direct indicator of PL ("*I can lead a healthy and active life*") was inserted into the model. A sensitivity analysis compared MLR estimation with WLSMV estimation, with no substantial differences in fit indices or parameters (WLSMV CFI= .96, RMSEA = .05 [.04, .06], SRMR = .033), as such, we present the results for the MLR estimation for comparability with other models.

Model fit and selection

Robust chi-square statistic, along with the Standardized Root Mean Square Residual (SRMR) were used to assess the absolute fit of each model to the data, while the robust versions of the Root Mean Square Error of Approximation (RMSEA) – along with its 90% confidence interval – and the comparative fit index (CFI) were used as indexes of approximate fit of the model. Guided by suggestions from the literature (Hu & Bentler, 1999; Schreiber et al., 2006) cut-offs of SRMR < .08, CFI ≥ .95 and RMSEA ≤ .06 – with .10 not included in its 90% confidence interval – as guidelines for quantifying global fit, rather than as strict rules (F. Chen et al., 2008; Gana & Broc, 2019; Marsh et al., 2004). For model comparisons we used scaled chi-square difference tests (Satorra & Bentler, 2001; with a significance level of .05), as well as information-based indices Akaike Information Criteria (AIC; Akaike, 1998), and Bayesian Information Criteria (BIC; Schwarz, 1978) – with lower values indicating better model fit. Local fit of the models was assessed through examination of the modification's indices and standardized covariance residuals (with a cutoff of |1.96|; Brown, 2015).

Convergent and discriminant validity

In the interest of brevity, the mean of standardized factor loadings was used to summary and assess convergent validity of each factor; these were evaluated as excellent, very good, good, fair or poor when higher than .71, .63, .55, .45 and .32, respectively (Comrey & Lee, 1992). Only solutions that achieved acceptable or borderline global fit were summarized. Correlations among factors were used to assess discriminant validity with a threshold of r<.85 (Brown, 2015).

Variance and reliability

For the final selected model (i.e., asymmetrical bifactor), coefficient omega (McDonald, 1999) was calculated for both the general (ω), and group factors (ω_S) to quantify the total-score reliability, and subscale-score reliability; a value higher than .70 was considered acceptable (Nunnaly & Bernstein, 1994), and a value higher than .80 good (Price, 2017).

Omega hierarchical (McDonald, 1999) was also calculated at general (ω_H) and group level (ω_{HS}). ω_H was used to evaluate tenability of interpretation of a sole total score, with a threshold of .80 (Dueber, 2020; Rodriguez et al., 2016). ω_{HS} for group factors were used in tandem with Explained Common Variance by each group factor (ECV_{SS}) to assess whether use of subscales add unique information, as proposed by Dueber (2020).

Explained Common Variance (ECV; Sijtsma, 2009) and percentage of uncontaminated correlations (PUC; Reise, Scheines, et al., 2013) were used as measures of essential unidimensionality of the general factor to assess whether a measurement model for use in SEM can be tenably specified as unidimensional, without considerable bias. Complementarily, indicator-level ECV (I-ECV; Stucky & Edelen, 2015) indexed the common variance attributable to the general factor in each indicator. All aforementioned indices were calculated using the *BifactorIndicesCalculator* package (Dueber, 2021).

Composite Confirmatory Analysis

Our initial postulated model conceptualized PL and its domains as composites, as such we used *Composite Confirmatory Analysis* (CCA; Schuberth, 2020) to mimic the analysis done through CFA and compare both measurement models.

All composite models were estimated in cSEM 0.4.0.9000 (Rademaker & Schuberth, 2020) using the PLS estimator with 1000 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 18): C1) a single composite model, C2) a correlated composite model, C3) and a second-order model of PL (using the "two-stage approach"; Schuberth et al., 2020; van Riel et al., 2017); 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 modelled composites.

Bootstrapped-based (1000 replications) overall tests of model fit (d_L and d_G) and SRMR were used to assess global fit of the model against a saturated model (good fitting model as having a lower value that its 95% quantile distribution, and tentatively SRMR < .08; Benitez et al., 2020; Henseler, 2021). *Root Mean Square Residual Covariance* (RMS_θ) was used as secondary measure of model fit, with a cut-off of .012 (Henseler, 2021). The magnitude and statistical significance (at p <.05) of indicator was used as measures of local fit. Multicollinearity was assessed with a combined analysis of indicator correlations and the Variance Inflation Factor (VIF) (Cenfetelli & Bassellier, 2009) to identify possible suppressor effects among variables.

Results

Preliminary Analysis

Bivariate correlations between all indicators displayed results compatible with the *a priori* factorial structure: indicators theoretically in same factor (domain) correlated higher with indicators in the same factor that with those of other domains (Table 38). An exception to this, were 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); and seemed to cluster only with the measure of the other limb (e.g., right shoulder with left shoulder). As such, we chose to remove 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 (Table 39), attaining an acceptable fit to the data– a statistically significant χ^2 ruled out an excellent fit. 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 .90 for CFI, and borderline to RMSEA and SRMR of .08 (e.g., Gana & Broc, 2019).

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 modification emerged: a) 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; and b) between the *Emotional Regulation* and *Physical Regulation* indicators , given shared similarities in wording and structure of the items. We specified a post-hoc error covariance between these indicators, resulting in the F3 model, which were kept for testing in further models. F3 was an improvement over F2 according to all indices.

MI analysis of F3 suggested that the model could be further improved by allowing a cross-loading of the *Culture* indicator on the Psychological factor. 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 chose to keep this parameter constrained since the following bifactorial specification would assess whether a general trait could best account for this implied correlation. Fitting the hierarchical solution (F4), provided a worse fitting solution comparing to F3, with large MI (>1000) and an inadmissible estimate of 1.0 (disturbance = 0) for the second order loading of the Psychological factor. Alternatively 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); as such, we fit the asymmetrical model (F6) which resulted in better overall indices and lower MI and most residuals below the |1.96| threshold. No direct comparison between was possible due to the insertion of a global indicator of PL to estimate F6.

Table 38. 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	.38																				
3. Self-Reported PA – Walking	.25	.36																			
4. PACER	.31	.01	-.04																		
5. Push-ups	.29	.03	-.07	.61																	
6. Curl-ups	.24	.05	-.03	.44	.41																
7. Shoulder Stretch. Right	-.13	-.02	-.08	-.03	.08	.24															
8. Shoulder Stretch. Left	-.10	-.05	-.06	-.02	-.04	.00	.75														
9. Sit-and-reach Right	-.09	.05	-.03	-.14	-.07	-.05	.31	.32													
1. Sit-and-reach Left	-.07	.08	.01	-.14	-.05	-.01	.32	.30	.93												
11. Manipulative-Based Activities	.23	.05	.01	.37	.43	.34	-.23	-.25	-.23	-.22											
12. Stability-Based Activities	.18	.06	-.01	.31	.35	.27	-.06	-.11	-.03	-.01	.79										
13. Content Knowledge	-.01	-.08	-.07	.17	.10	.10	.19	-.03	.15	.14	-.06	-.05									
14. Motivation	.45	.13	.03	.40	.35	.30	-.05	-.04	.05	.09	.33	.31	.09								
15. Confidence	.40	.10	.04	.49	.45	.37	-.08	.02	-.01	.03	.37	.33	.02	.73							
16. Emotional Regulation	.06	.06	.05	.19	.21	.11	-.13	-.04	-.03	-.02	.16	.14	.03	.28	.42						
17. Physical Regulation	.24	.10	.14	.28	.19	.25	-.13	-.09	.05	.06	.20	.17	.06	.50	.56	.44					
18. Culture	.31	.16	.06	.26	.22	.23	-.10	.07	.10	.10	.21	.20	.10	.47	.46	.16	.36				
19. Ethics	.03	.03	.13	-.08	.01	.04	-.03	.03	.11	.13	.05	.04	.11	.21	.16	.23	.32	.16			
20. Collaboration	.08	.07	.06	-.13	.01	.02	-.15	-.09	.05	.07	.04	.01	.04	.23	.19	.17	.32	.23	.63		
21. Relationships	.21	.10	.13	.07	.07	.07	-.17	-.05	-.03	-.01	.13	.14	.03	.36	.39	.20	.39	.34	.42	.63	

PA – Physical Activity; PACER – Progressive Aerobic Cardiovascular Endurance Run
Note: all correlations with Shoulder Stretch variables are polyserial correlations

Table 39. Model fit, mean factor loadings and inter-factor correlations for factor-based models

	First-order models			Second-order models		
Fit measure	**Unidimensional (F1)**	**Correlated Factors(F2)**	**Correlated Factors (F3)**[a]	**Hierarchical (F4)**	**Bifactor (F5)**	**Bifactor S·I – 1 (F6)**
MLRχ^2	1217.07 (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	.55	.78	.90	.89	.95	.97
Robust RMSEA [90%CI]	.18 [.17, .19]	.13 [.12, .14]	.09 [.08, .10]	.09 [.08,.10]	.07 [.06, .08]	.05 [.04, .06]
SRMR	.12	.11	.09	.09	.06	.04
AIC	58075.108	57452.753	57110.133	57150.092	56991.07	
BIC	58253.850	57652.773	57318.665	57350.112	57233.65	
χ^2 robust different test		**F1 vs F2**: $\chi2$ = 319.51, Δ df = 5, *p* = <.001	**F2 vs F3**: $\chi2$ = 7338.7, Δ df = 2, *p* = <.001	**F4 vs F3**: $\chi2$ = 151.35, Δ df = 2, *p* = <.001	**F3 vs F5**: $\chi2$ = 712.34, Δ df = 8, *p* = <.001	
Mean factor loadings						
General					.44 (.20)	.46 (.20)
Physical			.61 (.15)		.44 (.16)	.51 (.17)
Psychological			.69 (.21)		.32 (.14)	.30 (.22)
Social			.65 (.22)		.55 (.36)	.48 (.37)
Cognitive			.77		.77	.77
Factor Correlations (SE)						
Physical ~ Psychological			.67 (.03)			.48 (.11)
Physical ~ Social			.02 (.06)			-.33 (.07)
Physical ~ Cognitive			.19 (.06)			.14 (.08)
Psychological ~ Social			.42 (.07)			-.20 (.12)
Psychological ~ Cognitive			.06 (.06)			-.14 (.09)
Social ~ Cognitive			.09 (.07)			.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

[a] Correlated residuals between Emotional Regulation and Physical Regulation, and Manipulative-based Activities and Stability-based Activities

Table 40. 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	Standardized loading (SE)	S^2	Standardized loading (SE)	S^2	Standardized loading (SE)	S^2	Standardize d loading (SE)	S^2	
PL Global	.76 (.03)	.58									1.00
Manipulative-Based Activities	.30 (.05)	.09	.41 (.06)	.17							.35
Stability-Based Activities	.30 (.05)	09	.32 (.06)	.10							.46
PACER	.41 (.05)	.16	.70 (.04)	.49							.25
Push-ups	.35 (.05)	.12	.67 (.05)	.45							.21
Curl-ups	.35 (.05)	.12	.44 (.05)	.19							.38
Motivation	.74 (.04)	.55			.31 (.11)	.10					.85
Confidence	.76 (.04)	.58			.58 (.08)	.34					.63
Emotional Regulation	.33 (.06)	.11			.28 (.08)	.08					.58
Physical Regulation	.68 (.04)	.47			.04 (.09)	.00					.996
Ethics	.34 (.06)	.12					.54 (.04)	.30			.28
Collaboration	.38 (.06)	.14					.91 (.06)	.83			.15
Relationships	.54 (.04)	.29					.47 (.04)	.22			.57
Culture	.57 (.04)	.33					.01 (.05)	.00			.999
Content Knowledge	.11 (.05)	.01							**.77 (.02)**	.59	.02
ECV/ ECV$_{SS}$		.49		.71		.23		.61		.98	
ω/ω$_S$		.89		.76		.82		.80		.60	
ω$_H$/ ω$_{HS}$		.71		.52		.15		.42		.59	
PUC		.79									

PL – Physical Literacy; S^2-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$_S$ – Omega coefficient subscale; ω$_H$ / ω$_{HS}$ – Omega hierarchical coefficient / omega hierarchical subscale

Convergent and discriminant validity

Model F3 mean factor loadings were mostly very good, with correlations between factors ranging from .01 to .66 supporting discriminant validity. The correlation pattern was uneven, with the Psychological factor showing moderate correlations with only the Physical and Social factors – other factors had negligible to low correlations.

Mean factor loadings in the group factors of the F5 and F6 solution 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 (Δ = .02) in mean factor loadings in the general factor, and concomitant reduction in two of the group factors (i.e., Psychological and Social).

In the asymmetrical bifactor model (F6), 3 indicators had excellent loadings on the general factor (Physical Literacy), 1 had very good, 1 had good, and the remaining 7 indicators had borderline (~.32) to poor loadings – out of these, *Content Knowledge* had no statistically significant loading (Table 40). Loadings on the group factors (residual variance not explained by the general factor) ranged from .32 to .70, .04 to .58, .01 to .91, and .77, on the Physical, Psychological, Social and Cognitive domain respectively; except for the indicators in the Psychological domain, all indicators had on average higher loadings on their group factors than on the general factor (Table 39). After accounting for the general factor (Physical Literacy), half of the inter-factor correlations became negative, with the remaining three showing a decrease in their positive correlations - this effect was negligible in the case of the Physical ~ Cognitive correlation.

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 were due to both the general and group factors (ω =.89), and the remaining 11% due to error. An estimated 71% of total score variance was due to individual differences in the general factor (ω_H = .71; this ω_H value is lower than the tentative .80 for meaningful interpretation of the total score in isolation (Dueber, 2020; Rodriguez et al., 2016). Reliabilities for the subscale scores were all adequate or good, except for the Cognitive score (ω_s from .60 to .82). Based on the relationship between ω_s and $\omega_{Hs,}$ three subscales attained the recommended cut-offs for adding statistical value over and beyond that of the total score (recommended ω_{Hs} = .212, .192, .192 and .244, respectively).

Similar results were estimated by the ECV: 46% of the total common variance (inherent to both general and group factors) is explained by the general factor (Table 40). Of all indicators, only three (*Motivation*, *Physical Regulation*, and *Culture*), achieved the tentative .80 - .85 cut-off for I-ECV (Stucky & Edelen, 2015) and can be regarded as essentially being influenced by the general trait alone. Most reliable variance in indicators was explained by their respective group factor, resulting in marked dimensional uniqueness, except in the Psychological factor (ECV_{SS} = .56, .22, .71 and .99 respectively for the Physical, Psychological, Social and Cognitive group factors). Based on ECV_{ss} and ω_S, all group factors attain the recommended value for dimensional uniqueness (i.e., warrant interpretation) (recommended ω_S = .479, .815, .479 and .479; Dueber, 2020). Finally, 77% of all correlations were saturated by the general factor (PUC = .77), bordering on the 80% recommendation (Reise, Scheines, et al., 2013) 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 (Table 41) suggesting excellent fit to the data. Both C2 and C3 models provide an acceptable fit to the data, with estimated borderline SRMR (both cases) and RMS_θ (not available for hierarchical model) 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). It is relevant to note that an alternative to model C3b (not available directly in *cSEM*) would be to attribute equal weights to domain scores and sum them into a total PL score, which would render the model equivalent to C1b.

Variance

Standardized weights in the single composite model (C1; Table 42) ranged from -.18 to 62, with *Motivation* being the only indicator with a statistically significant result – all other indicators do not contribute beyond this indicator. High correlations (<.70) existed between some indicators, with corresponding VIF in the 2.26 to 3.14 range, suggesting existence of suppressor effects.

In the correlated composites model (C2; Table 43), standardized weights increased for most variables as expected, given the reduced number of competing predictors in each composite – ranging from -.23 to .53; -.22 to .70; -.24 to .76; and 1.0 (single-indicator composite), for the Physical, Psychological, Social and Cognitive composites, respectively. Four indicators had non-statistically significant weights and two were borderline ($p \approx .05$). VIF values at indicator-level were generally lower, with all composites showing existence of high correlations among some of its indicators, with unexpected, inverted signs in three indicators – symptoms of suppressor effects still in play. Correlations among composites ranged from negligible (.05) to moderate (.53) (Hinkle et al., 2003), with correlations with the Cognitive composite being all negligible and non-statistically significant (.05 to .11).

For the most part, weights, VIF and indicator correlation kept their magnitude in the second-order composite (Table 44), except for an increase in weights for *Push-ups*, *Curl-ups*, *Physical Regulation* indicators, and 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 = .69), in explaining variance in the second-order composite of Physical Literacy. Analysis of the first-order loadings (bivariate correlations) suggested that the Physical, and Social composites (loadings = .68) were still important in composing this second-order composite, despite explaining less amounts of variance.

Table 41. Model fit and inter-factor correlations for Composite–based models (C1 - C3b); n= 443

	First-order				Second-order	
Fit measure	**Single First-Order Composite (C1)**	**Single First-Order Composite – Unit Weighted (C1b)**	**Four First-Order Composites (C2)**	**Four First-Order Composites (C3) – Unit Weighted (C2b)**	**Hierarchical (C3)**	**Hierarchical – Unit Weighted (C3b)[b]**
RMS$_{\theta}$	<.001	<.001	.04	.03		
SRMR (Critical Value 95%)	.050 (.054)	.059 (.059)	.078 (.068)	.087 (.061)	.071 (.060)	.090 (.062)
d$_L$ (Critical Value 95%)	.391 (.443)	.534 (.533)	.928 (.707)	1.147 (.563)	.767 (.547)	1.228 (.586)
d$_G$ (Critical Value 95%)	.085 (.106)	.118 (.113)	.179 (.146)	.215 (.138)	.169 (.134)	.224 (.140)
Construct Correlations						
Physical ~ Psychological			**.49 (.04)**	**.47 (.03)**		
Physical ~ Social			**.27 (.06)**	**.12 (.04)**		
Physical ~ Cognitive			**.11 (.05)**	.06 (.04)		
Psychological ~ Social			**.53 (.05)**	**.49 (.04)**		
Psychological ~ Cognitive			.07 (.05)	.05 (.05)		
Social ~ Cognitive			.05 (.05)	.09 (.05)		

RMS – Root Mean Square Error Covariance; SRMR – Standardized Root Mean Square Residual; d$_L$ – Squared Euclidean Distance; d$_G$ – Geodesic Distance

[b]Only 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.

Note: statistically significant weights (*p* < .05) are boldened

Table 42. Item parameters and total effects of Single First-order Composite Models (C1 and C1b); n= 443

Composite ← indicators	**β (SE)**	**VIF**	**Indicator Correlation**
PL ←			-.14 - .77
Manipulative-Based Activities	.20 (.14)	2.88	
Stability-Based Activities	-.25 (.13)	2.55	
PACER	.15 (.16)	2.03	
Push-ups	.12 (.15)	1.89	
Curl-ups	.09 (.10)	1.39	
Motivation	.61 (.13)	2.39	
Confidence	.13 (.10)	3.14	
Emotional Regulation	-.17 (.10)	1.41	
Physical Regulation	.03 (.13)	1.84	
Ethics	-.07 (.12)	1.76	
Collaboration	-.05 (.14)	2.36	
Relationships	.15 (.12)	1.89	
Culture	.18 (.11)	1.38	
Content Knowledge	-.14 (.10)	1.10	

Table 42. Item parameters and total effects of Single First-order Composite Models (C1 and C1b); n= 443

Optimal weights (C1)	Total effects β (SE)	f²	R²adj.
IPAQ ← PL			
Vigorous	**.52** (.03)	.36	.27
Moderate	**.16** (.06)	.03	.02
Walking	.07 (.07)	.00	.00
Unit weights (C1b)			
IPAQ ← PL			
Vigorous	.39 (.04)	.18	.15
Moderate	.12 (.05)	.02	.01
Walking	.07 (.05)	.01	.00

β – Standardized Weights; SE – standard error; VIF – Variance Inflation Factor; R²adj. – adjusted R²; PACER – Progressive Aerobic Cardiovascular Endurance Run; PL – Physical Literacy; IPAQ – International Physical Activity Questionnaire
Note: statistically significant (*p*<.05) weights are boldened

Table 43. 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	.29 - .77
Manipulative-Based Activities	**.41 (.20), *p*=.050**	2.82	
Stability-Based Activities	-.23 (.19)	2.50	
PACER	**.53 (.17)**	1.72	
Push-ups	.28 (.19)	1.79	
Curl-ups	.22 (.15)	1.34	
Psychological ←		1.70	.29 - .73
Motivation	**.70 (.13)**	2.22	
Confidence	**.40 (.15)**	2.57	
Emotional Regulation	**-.22 (.11), *p* =.065**	1.35	
Physical Regulation	.06 (.15)	1.64	
Social ←		1.40	.13 - .63
Ethics	-.02 (.22)	1.64	
Collaboration	-.24 (.24)	2.22	
Relationships	**.59 (.17)**	1.69	
Culture	**.76 (.14)**	1.12	
Cognitive ←		1.01	-
Content Knowledge	1.0		

Optimal weights (C2)	Total effects β (SE)	f²	R²adj.
IPAQ ← Physical			
Vigorous	**.19 (.05)**	.04	.25
Moderate	-.04 (.06)	.00	.04
Walking	-.08 (.07)	.01	.02
IPAQ ← Psychological			
Vigorous	**.36 (.05)**	.10	

Table 43. Item parameters and total effects of Four First-order Composite Models (C2 and C2b); n= 443

Moderate	.08 (.06)	.00	
Walking	.04 (.07)	.00	
IPAQ ← Social			
Vigorous	.06 (.05)	.00	
Moderate	**.15 (.06)**	.02	
Walking	.13 (.08)	.01	
IPAQ ← Cognitive			
Vigorous	-.04 (.04)	.00	
Moderate	-.08 (.05).	.01	
Walking	**-.08 (.04)**	.01	
Unit weights (C2b)			
IPAQ ← Physical			
Vigorous	.22 (.05)	.04	
Moderate	-.00 (.06)	.00	
Walking	-.09 (.06)	.00	
IPAQ ← Psychological			
Vigorous	.25 (.05)	.05	
Moderate	.11 (.06)	.01	
Walking	.08 (.07)	.00	.17
IPAQ ← Social			.02
Vigorous	.03 (.06)	.00	.03
Moderate	.08 (.07)	.01	
Walking	.12 (.06)	.01	
IPAQ ← Cognitive			
Vigorous	-.02 (.05)	.00	
Moderate	-.08 (.04)	.01	
Walking	-.09 (.05)	.01	

β – Standardized Weights; SE – standard error; VIF – Variance Inflation Factor; R^2adj. – adjusted R^2; PACER – Progressive Aerobic Cardiovascular Endurance Run; PL – Physical Literacy; IPAQ – International Physical Activity Questionnaire

Note: statistically significant (p<.05) and borderline weights are boldened

Table 44. 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 ←			.29 - .77
Manipulative-Based Activities	.19 (.14)	2.82	
Stability-Based Activities	.09 (.14)	2.50	
PACER	**.54 (.11)**	1.72	
Push-ups	**.21 (.12) *p* =.072**	1.79	
Curl-ups	**.29 (.10)**	1.34	
Psychological ←			
Motivation	**.44 (.08)**	2.22	.29 - .73
Confidence	**.56 (.10)**	2.57	
Emotional Regulation	-.10 (.07)	1.35	
Physical Regulation	**.16 (.08), *p* = .06**	1.64	
Social ←			.13 - .63
Ethics	.08 (.12)	1.64	
Collaboration	-.14 (.14)	2.22	
Relationships	**.55 (.09)**	1.69	
Culture	**.73 (.07)**	1.12	
Cognitive ←			-
Content Knowledge	1.0	-	
Second order			
PL ←			
Physical	**.27 (.13)**	1.42	
Psychological	**.69 (.13)**	1.93	
Social	.22 (.15)	1.48	
Cognitive	-.14 (.10)	1.02	
Optimal-weights (C5)	**Total effects β (SE)**	**f^2**	**R^2adj.**
IPAQ ~ PL			
Vigorous	**.48 (.04)**	.30	.23
Moderate	**.16 (.05)**	.03	.02
Walking	.09 (.06)	.01	.01
Unit-weights (C3b)			
IPAQ ~ PL			
Vigorous	**.41 (.04)**	.20	.17
Moderate	**.14 (.05)**	.02	.02
Walking	.09 (.06)	.01	.01

β – Standardized Weights; SE – standard error; VIF – Variance Inflation Factor; R^2adj. – adjusted R^2; PACER – Progressive Aerobic Cardiovascular Endurance Run; PL – Physical Literacy; IPAQ – International Physical Activity Questionnaire
Note: statistically significant (p<.05) weights and borderline weights are boldened

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^2adj. = .27; Table 42 - 43); variance explained on moderate and walking self-reported PA was all-around negligible. Using the correlated composite (C2) approach revealed different contributions by composite: Psychological and Physical had a higher effect size (f^2) on vigorous PA, while Social had a low effect size on moderate PA; Cohen, 1988). 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 decrease in contributions of the Psychological composite, with others maintaining their relative magnitudes.

Discussion

The aims of this paper were to a) establish evidence on construct validity of the PPLA, integrating measures derived from the PPLA-Q and PPLA-O, by comparing results drawn from factor-based methods, and composite-based methods; and b) assess adequacy of using a PL total-score, and respective subscales. As secondary research aim, we discuss practical and ontological implications of the different methods for PL.

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 were different from those found in other PL measuring batteries. In the most recent construct validation effort of the *Canadian PL Assessment* (CAPL), Gunnell and colleagues (2018) modelled a second-order factor to account for correlations between domains of PL; they, however, did not report fitting a bifactorial model. Similarly, in a validation of a PL measuring model for children and youth (Cairney et al., 2019), a second-order factor model was chosen as 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), 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 (*r* = .67, SE = .03; Table 39), and moderately with the Cognitive factor (*r* = .19, SE = .06), it did not correlate with the Social factor (*r*=.02, SE = .06). Despite using different measures, and operational definitions of the constructs (with a *Daily Behavior* domain, and without the Social domain), 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 low loading (.21) on the second-order Physical Literacy factor (Gunnell et al., 2018); again, 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 models direct effects of the general factor (Physical Literacy) on indicators, with the asymmetrical version allowing for correlations among group factors, resulting in a better fit than that of its symmetrical counterpart - which suggests that the orthogonality constraint was overly restrictive, and that the PL general factor fails to account for all shared variance among domains (group factors).

Our results from the bifactor model (F6) analysis suggest 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, complementary use of unit-weighted subscale scores (per domain) has additional value over and beyond the single total score since they present enough dimensional uniqueness (akin to convergent validity in a correlated factor model). A noteworthy exception is that of the interpretation of the Psychological subscale: since indicators on this domain seem to be saturated by the general factor (i.e., small loadings on the group factor, with corresponding strong factors on the general one), 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 (Babic et al., 2014; Park & Kim, 2008), and adults (Amireault et al., 2013); similarly, these variables might play a mediating roles between other domains – an hypothesis that can be researched in the future with competing models.

Similarly, despite achieving borderline values to be considered as essentiality measuring the single trait of Physical Literacy (PUC >.80, or ECV > .60 and $\omega_{H} >$ .70; Reise, Scheines, et al., 2013) high values of ECV_{SS} and moderate values of ω_{HS} suggest that further research in structural equation modelling contexts should use a bifactor measurement model for the PPLA, instead of alternative structures (i.e., unidimensional, or correlated factors) to capture as much information as possible from the group factors, and avoid bias. This would also allow to test different effects of the general factor and group factors.

Comparison with composite-based methods

On our application of 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 a) the existence of high correlations among indicators; and b) the number of indicators estimated in the same composite.

Although no assumptions regarding covariation of indicators is made in a composite model (e.g., Jarvis et al., 2003), high correlation patterns among indicators – even if VIF values are below the usual 3.3 recommendation (Diamantopoulos & Siguaw, 2006) – will generally result in multicollinearity, and cause suppression effects, co-occurrence of positive and negative weights (i.e., "flipped signs"), and preclude meaningful interpretation of these weights in general (Cenfetelli & Bassellier, 2009). Since a multiple regression is used to estimate the weights of indicators, these are competing for explained variance (i.e., maximum average weight would be .267, assuming 13 uncorrelated indicators; Cenfetelli & Bassellier, 2009), 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 explain variance in their respective composites over and beyond other indicators; while in the later the magnitudes of standardized loadings obtained by these indicators in the general factor of Physical Literacy, along with I-ECV were poor. A similar case occurred with the *Content Knowledge* indicator, albeit in the former case, its poor performance was carried into the first-order weight (since it was defined as a single-indicator composite). Analyzing the first-order weights and the inverse of the ECV_{SS} (i.e., $ECV_{GS\,=}$ 1- ECV_{SS} , variance explained by the general factor in each set of indicators; not shown) suggests a similar pattern: the Psychological indicators contribute more to the general factor/higher-order composite (β = .69, $ECV_{GS\,=}$ .78), the Physical (β = .27, $ECV_{GS\,=}$ .29) and Social (β = .22, $ECV_{GS\,=}$ .39) indicators contribution is similar, and then the Cognitive indicator contributes marginally (β = -.14, $ECV_{GS\,=}$ .02). This is parallel to our earlier discussion on the absorption of the most of the Psychological indicators' variance into the general Physical Literacy factor (asymmetrical bifactor model).

Thus, these models could be further improved by dropping indicators with statistically non-significant weights (Cenfetelli & Bassellier, 2009), indicators with low I-ECV, or with poor loadings on both the general and group factors (Rodriguez et al., 2016). This, however, would 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 done, this analysis should be replicated in a large independent sample, outside of COVID19 restrictions – which might pose unexpected 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 remaining correlations being lower. A noteworthy difference is the increase in correlation among the Physical and Social domain in the composite model, which could be attributed to difference in indicator weighting between models.

Regarding the structural model estimates (inner model) obtained with the composite-based methods we will not offer an extensive discussion, since our focus for these studies was on construct validity, and not on predictive validity. Lower levels of PA during COVID-19 (Stockwell et al., 2021) might have had a substantial effect on estimated paths. Nonetheless, our results establish preliminary evidence in favor of predictive validity of the PPLA to predict vigorous (self-reported) PA, and suggest that different domains of PL might interact differently with different intensities of PA; a finding also reported by Belanger et al. (2018).

Summing up, evidence in favor of a measurement model with a higher-order Physical Literacy construct (either represented by an asymmetrical bifactor or composite hierarchical model) was mostly robust across methods, with comparable results. Regarding 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 (i.e., being more useful than its isolated parts; Henseler, 2021); 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

Despite having achieved analogous results, both methodologies stem from different philosophical and ontological backgrounds regarding the nature of the construct. While from a reflective, realist perspective (i.e., factor-based methods) a bifactor view seems the most empirically and conceptually plausible one, since it is tenable that a transversal broadband meta-learning or disposition (i.e., common factor) influences all different elements (here represented as indicators) in a movement context; with domain-specific processes, inherent to physical, affective, social and cognitive skills' development, originating clusters of highly interdependent variance (i.e., group-factors). This also 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, Biggs & Collis, 1982).

Also, if a higher-order common-factor perspective is tenable, then different domains and elements are theoretically interchangeable (e.g., Jarvis et al., 2003) since they are merely a sample of infinite possible indicators and facets that could be chosen to measure Physical Literacy; and variation at PL-level should be reflected with equivalent magnitude (after accounting for measurement error) in all domains and indicators. While it seems certainly plausible that a different set of indicators could be selected according to the research questions and applications at hand – as we did when we selected from the 30 elements which the original APLF posits – it might dimmish the integrated perspective that researchers have been seeking all along, if no proper care is taken to ensure representation across all domains. The asymmetrical bifactor model offers a compromise solution, as it becomes possible to acknowledge that while indicators are interchangeable within a domain, domains themselves are not interchangeable (Eid, 2020; Eid et al., 2017), and are essential to defining the PL construct. Other plausible interpretations include that of PL as a network of interconnected latent variables – similar to our F3 model, and initial efforts of the CAPL (Longmuir et al., 2015)– that may or may not correlate. This, however, could compromise the place of PL in educational policy discussion, since it would present no added value as a whole variable, and could easily be picked apart based on convenience.

From a diametrical point-of-view, viewing PL as an emergent variable (composite indicators), 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 (Bollen & Bauldry, 2011; Bollen & Diamantopoulos, 2017). This would also recognize that selected indicator variable might share a distal common causes mediated through a complex chain of mediators and moderators; a vision more compatible with the epistemic phenomenology position of the *Whiteheadian* school of thought, wherein each individual 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 (Bollen & Diamantopoulos, 2017), and similarly hinder meaningful progress on educational policies and research, if care is not taken in interpretation and dissemination. A possible 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 – which is unavailable in PLS-PM, given its composite-based nature (Benitez et al., 2020; Sarstedt et al., 2016). According to this, the scope of PL 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, without mandating concomitant variation in all elements (Bollen & Diamantopoulos, 2017; Sarstedt et al., 2016). Further research should seek to reconcile and/or discuss these two paradigms.

Since our results are compatible with both interpretations, we take a more practical stance and tentatively recommend the common factor lens of analysis, as it is the implicit foundation of both Classical Test Theory (CTT) and Item Response Theory (IRT), affording access to a more robust analysis toolkit (e.g., FIML estimation) to explore dimensionality, score adequacy and response patterns; as we have shown in initial validation of the different PPLA measures (Mota et al., 2022c, 2022b, 2022a), and in earlier sections of this article. It might also afford the possible to disentangle the impact of different group factors on intended outcomes of PL, 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 and shed light about the adequacy of the conceptual interpretation we described.

Strengths and limitations

A major strength in this study was the comparison between two different methodologies to draw inferences about the construct validity and reliability of the PPLA, as measured by its 13 indicators. Secondly, to our knowledge it is the first study to demonstrate the application of bifactor models to a PL assessment tool, to assess the adequacy of interpretation or use of scores and sub scores. Thirdly, our study builds upon measures which have gather evidence of construct validity and reliability at item-level using Item Response Theory methodologies, providing more accurate estimates, as well as capability to study item quality and psychometric behavior in detail.

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. And the use of measures for Movement Competence based on exploratory GRM. We did not account for multilevel grouping within data (i.e., schools and classes) which could also hold some bias over the results. Also, despite mimicking the relative composition of grade 10 to 12 students' 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 ideally, 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 (e.g., sex, grade, major), measurement invariance should be assessed for the full model, along with its predictive validity on meaningful outcomes (e.g., objectively measured PA, well-being) which was not one of our foci in this study. We also recognize that PL could cogently be modelled using equivalent or alternative models and encourage further research into it.

Another limitation created using IRT-calibrated measures (i.e., scales and scores) 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 bias 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 that these indicators are relevant to the whole-picture of PL and its inclusion should be considered and scrutinized in future efforts, despite its joint-specific nature (Committee on Fitness Measures and Health Outcomes in Youth et al., 2012).

Finally, we acknowledge that there are likely statistical and methodology nuances that eluded us regarding differences and similarities among reflective and formative (both in causal-formative and composite forms); although some literature has been generated towards finding a *non-versus* perspective between methods, clear guidelines are still warranted if conceptually different measurement perspectives are to have meaningful dissemination.

Conclusion

Using both confirmatory factor analysis and confirmatory composite analysis, we gathered evidence supporting construct validity and reliability of the PPLA as an integrated tool to measure Physical Literacy as a multidimensional construct in 15 to 18 years old Portuguese students. Out of all 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 is responsible for a considerable amount of variance in all indicators – albeit with insufficient strength to be interpreted in isolation - with demarked domain-specific variance.

Based on this, we advise 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 to quickly compare different classes and schools in low-stakes settings, the latter allows for more meaningful interpretation of students PL profiles and needs. We also recommend use of total scores per element/indicator – which have gathered previous content and construct validity evidence; as these will allow a richer and more specific feedback to students and families, arguably better serving teaching practices and self-development towards a more meaningful and impactful PL journey.

For future research efforts, we recommend the use of a bifactor measurement model, which permits disentangling of variance attributed on the general PL trait and its domains. This capability opens the possibility to estimate independent associations of the four domains of PL with external outcomes, along with the general trait, to identify different antecedent or outcome variables, which could improve the dissemination and testing of the concept as key attribute in lifelong PA practice. It also seems to provide a better representation of the multidimensionality inherent to this construct. Nonetheless, instrument development and validation are an iterative process, as such, further comparisons could be made in the future, to better argue for other conceptualizations or specifications of measurement models for PL.

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 instrument 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 all the tireless PE teachers and students which participated in these studies. Similarly, we thank António Rosado, and Dean Dudley for their helpful insights during this project. The lead author would also like to thank his co-authors for their ever-present guidance and support during his PhD journey.

CHAPTER 7 – Conclusions, limitations, and future perspectives

The main purpose of this PhD thesis was to develop and validate a novel criterion-referenced PL assessment system for application in Portuguese PE that makes uses of the APLF model's domains, elements and learning continuum conceptualization to provide a detailed and feasible assessment of each student's PL journey. This may complement pedagogical decisions (at local, regional, and national level) towards a more meaningful and targeted PE environment to promote PL learning of grade 10-12 (15-18 years) adolescents.

Throughout studies detailed in the five papers (synthetized in Table 45) the PPLA was developed and initially validated into two parts using scientific instrument development guidelines and methodologies, and state-of-the-art conceptual frameworks: a) PPLA-Questionnaire, a self-administration questionnaire targeting part of the cognitive domain (*Content Knowledge*), and the psychological and social domains; and b) PPLA-Observation, an instrument that integrates regular assessment data collected by PE teachers regarding the physical and cognitive domains of PL. These studies also described its sequential validation process using multiple sources of validity, to ensure that the PPLA has gathered enough evidence to support its use and interpretation in the intended adolescent student population.

Overall, the PPLA emerges as a highly feasible tool for the PE context that can be completed in around 20 minutes (student self-administration of the PPLA-Q), plus time spent by the PE teacher inserting/copying data (already routinely collected during regular PE classes) into a spreadsheet (PPLA-O), with no further teacher training or specialized logistics required. It also employs a common criterion-referenced assessment framework, based on a non-linear learning continuum – as warranted by the developers of the APLF (Keegan et al., 2019) and multiple authors (L. Edwards, Bryant, Keegan, Morgan, Cooper, et al., 2017; Young et al., 2021). This focus on a comparison with standards deemed indicative of each stage of PL development; instead of a normative-referenced system that would compare and derive (usually through percentiles) each individual's standing related to results of other students (i.e., the results of a norming sample). The former also provides a position compatible with the idea of an individual journey of PL development, oriented through multiple milestones (objectives/goals dealing with mastery of relevant skills), rather than a peer-comparison mindset that could reinforce alienation to PA and movement altogether. Although a more ipsative focus on PL – considering every student as incomparable to any other standard than his/her own progress – as stressed by the *Whiteheadian* school of thought (Whitehead, 2010) could lead to higher internalization and adequacy if done properly, it arguably requires resources (e.g., additional PE teacher training, and time) which would pose a limitation to large scale application in PE, and could compromise the advancement of research in the field by reducing comparability across individuals and contexts. As such, a criterion-oriented focus might provide a tenable way to operationalize such a philosophically rich concept, while adhering to educational assessment's best practices.

Results on the PPLA are summarized through class and individual reports[3]
(Additional File 6 and

[3] It should be noted that these reports were produced before studies were finished, so that participating students and teachers could receive them in a timely manner before the end of the school year. As such, there are some minor differences regarding the terminology and elements used. They were produced in Portuguese and translated to English for the sole purpose of this thesis.

Additional File 7), providing a systematic and formative view on each student's PL journey, and support to differentiation of teaching-learning strategies – specially in domains and elements that are harder for teachers to assess and track over time (e.g., Psychological and Social); as well as enabling older adolescents to autonomously delineate strategies for their own journey through suggested strategies, complemented by the essential guidance of their PE teachers.

PPLA also opens the door to more general implications, both at national and international level, using assessment as a key driver of curriculum and pedagogy (Young et al., 2021). Firstly, widespread use of this tool has the potential to contribute to change in discourse of assessment and pedagogy policy in national PE by providing an auxiliary standardized and validated tool to assess, and thus, dedicate time to the development of domains that, although essential in our PL conception, as well as in the Portuguese PE curriculum (i.e., motivation, confidence, collaboration, interpersonal relationships), lacked readily available and integrated assessment tools in PE. Simultaneously, it also could provide a national advocacy mean to disseminate the PL concept in teacher's practice, reinforcing and updating the vision that has been in the syllabus for many years, while also contributing to the legitimization of PE as an indelible part of the Portuguese school curriculum to develop active citizens. Concomitantly, given its publication in international peer-reviewed forums, it affords an opportunity for worldwide recognition of the quality of the Portuguese PE curriculum, its didactics practices, and its professionals.

Multiple implications for research and theory development are also in sight. PPLA is, to our knowledge, the first published and validated instrument to make use of the APLF's application of a learning continuum to draw inferences about progression within and across each element, through a conciliatory conceptual and methodological take on diverse constructs that have inherently different assessment perspectives and methodologies depending on the discipline (motor development, sport psychology, sport sociology and educational assessment). It is also the first published instrument to operationalize the APLF for adolescents, and one of the few available for this age-range across different conceptualizations and models. Its use in research will allow for furthering of the theoretical discourse, by enabling comparation and hypothesis testing among different frameworks of PL

(CAPL vs APLF, or others) and their usability towards developing meaningful lifelong PA participation, in different contexts.

Table 45. Synthesis of research aims methods, main results, and recommendations from each data chapter

Chapter	Research aims	Methods	Main Results	Recommendations
Chapter 2 (Paper 1)	a) Develop the PPLA- Questionnaire b) Assess its Content Validity c) Assess feasibility d) Assess preliminary construct validity and reliability	a) Content Analysis + Expert input b) Evaluation by experts (CVI and κ) + Cognitive interviews with students c) Pilot testing d) CTT item analysis (difficulty, discrimination, distractor analysis) + PLS-SEM	a) Three modules, version 0.6 post-revisions: Cognitive - 10 Cognitive items (Content knowledge) Psychological – 4 scales (8 subscales) – 46 Likert-type items Social – 5 scales (8 subscales) – 43 Likert-type items b) i. Expert evaluation (N= 11, 2 rounds): most items had an Item-Content Validity Index ≥.78 and Cohen's κ ≥ .76. At module-level, S-CVI/Ave and UA were .87/.60, .98/.93 and .96/.84 for the cognitive, psychological, and social modules ii. Revisions made based on student's response processes and feedback (N = 12) c) Completion time *M*= 27 minutes; low response errors and no evidence of problematic response patterns d) i. Item analysis: item difficulty ranged from .10 (very hard) to .95 (very easy), with an average difficulty of .50. 6 good or very good discriminating items (D > .30) ii. PLS-SEM: adequate reliability in 10 out of 16 subscales (α >.70 and composite reliability > .60)	b) Further study the content validity of the Cognitive Module through expert evaluation d) Robust validation efforts with a larger sample size and eventual usage of IRT methodologies
Chapter 3 (Paper 2)	a) Assess dimensionality of PPLA-Q Psychological and Social modules b) Assess measurement invariance by sex (DIF and DTF) c) Assess reliability (total score, and test-retest 15 days-interval) d) Assess convergent and discriminant validity	a) Mokken Scale Analysis (Non-parametric IRT) b) Scalability stratified by sex (Loevinger's H and H_i) c) Molennar-Sijtsma ρ and ICC d) Bivariate attenuated Spearman correlations	a) 16 subscales can be coherently interpreted as 8 moderate to strong Mokken scales adhering to the MHM and DMM, H ranging from .47 to .66; 4 scales with an interpretable IIO (H^T>.30) – total sum score can be used as an indicator in the latent trait b) 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 – further scrutiny needed before any sex comparisons are made in this scale c) i. Score reliability: all scales had good total score reliability ρ>.80, ranging from .93 to .94 ii. Test-retest reliability (n=73): 3 scales had good to excellent reliability (ICC95%CI ranging from .72 to .95), and 5 scales presented moderate to good reliability (ICC95%CI ranging from .51 to .85). d) *Motivation* and *Confidence* scales showed a high disattenuated correlation (*r* >.85) – threat to discriminant validity; all other scales behaved as theoretically expected, with low to moderate across domain correlations providing support of convergent and discriminant validity	a) Adjustments to items of scales without an interpretable IIO (*Confidence*, *Emotional Regulation*, *Collaboration* and *Relationships*) to increase the overall distinction among each item's difficulty in the continuum of development; refinement and further analysis of items targeting *Relational* Thinking b) Explore causes of measurement invariance in *Physical Regulation* scale through qualitative methods c) Additional test-retest reliability assessment outside of COVID-19 lockdown restrictions – equate using IRT growth models d) Assessment of tenability of a higher-order factor underlying both constructs

Table 45. Synthesis of research aims methods, main results, and recommendations from each data chapter

Chapter	Research aims	Methods	Main Results	Recommendations
Chapter 4 (Paper 3)	a) Assess internal structure/dimensionality b) Assess measurement invariance by sex (DIF and DTF) c) Assess reliability (score and test-retest 15 days-interval) d) Assess whether using distractor information is useful for higher precision e) Assess whether the sum-score has enough precision for applied settings	a) Parametric IRT models b) LRT – based DIF and DTF analysis (sDTF) c) Marginal reliability and conditional reliability, ICC and Svenson's ordinal agreement d) Polytomous IRT models and Nested Logit models e) Bivariate Pearson correlations	a) Mixed 2-parameter nested logit + graded response model provided the best fit; C2 (21) = 23.92, *p* = .21; CFI = .98; RMSEA_{C2}= .017 [0,.043] with no misfitting items b) Evidence of DIF in Item 1 in favor of male students, however, does not translate in statistically significant DTF (sDTF = -0.06; sDTF% = -0.14) – further scrutiny recommended with higher sample size c) i. Score reliability: lower than acceptable .70 threshold (ρ_{xx}= .60); adequate reliability in the -2 to -1 θ range – useful for distinguishing student with transitional knowledge (between foundation and mastery) – further revisions needed to target full spectrum of θ ii. ICC (n=73): poor to moderate test-retest reliability (ICC = .56, [.38, .70]); iii. Svenson's method (n=73): 6 out of 10 items with acceptable agreement (>.70), 4 remaining items with small individual variability – suggested use of 3PL model with guessing parameter d) Modelling distractor information provides increase in available information and thus, reliability, along with possibility to identify correctness of distractors and their popularity e) High correlation (*r* = .91 [.90,.93]) among sum-score and scores derived from calibrated mixed model – sum-score might be useful for applied settings to get a quick overview of student's knowledge; for precision IRT score is recommended	b) Further scrutiny of DIF and DTF is recommend with higher sample size c) i. Refine, or add items to assess the whole θ spectrum, especially in the higher levels ii. and iii. Assess test-retest reliability using IRT growth models or 3-parameter logistic models accounting for guessing
Chapter 5 (Paper 4)	a) Develop the PPLA- Observation instrument b) Assess dimensionality of the MCRT module c) Assess measurement invariance by sex (DIF) d) Assess convergent and discriminant validity e) Assess score reliability;	a) Content Analysis b) MIRT c) LRT-based DIF analysis d) Bivariate Pearson and polyserial correlations e) EAP empirical reliability	a) Two modules: MCRT (22 physical activities) and HRF (5 FITescola ® protocols) – teacher-reported data entered in a spreadsheet b) Two-dimensional GRM showed best fit – 1) Manipulative-based Activities (MA), and 2) Stability-based Activities (SA) (N=515) c) No evidence of DIF d) r = .68 between dimensions; boys with higher scores in both dimensions; *r* = .23 and r = .18 of dimensions with age; r = -.13 between BMI and SA; similar magnitudes of correlations with previous meta-analysis and systematic reviews e) Adequate to good marginal reliabilities: MA = .89, and SA = .73	b) Explore the possibility of developing separate modules for assessment of *Rules*, and *Tactics* elements to allow the disentanglement of their variance from *Movement Competence*-related one in reported proficiency levels e) Evaluate inter and intra-rater reliability of PE teachers in activities' assessment, along with any patterns that may emerge related to teaching experience of other variables

Table 45. Synthesis of research aims methods, main results, and recommendations from each data chapter

Chapter	Research aims	Methods	Main Results	Recommendations
Chapter 6 (Paper 5)	a) Assess construct validity at higher conceptual level b) Assess adequacy of usage of a total PL-score and respective subscales by domain c) Discuss conceptual and practical implications of reflective vs formative measurement for PL	a) Confirmatory Factor Analysis (CFA) and Confirmatory Composite Analysis (CCA) b) Bifactor model indices (ω and ω_{H}, ECV), and unit-weighted composite models c) Literature review + empirical evidence	a) i. CFA (N=521): Asymmetrical Bifactor model (S*1-1) provided best fit to the data: Robust CFI .97, Robust RMSEA .05 [.04,.06], SRMR .04 ii. CCA (n= 443): Best absolute fit for single first-order composite models (dG, dL, and SRMR below or borderline of their 95% critical value, in both the optimal and unit weighted models, in both optimal and unit-weighted models); adequate approximate fit of four correlated composite and hierarchical (second-order) model b) Total score should not be used in isolation ($\omega_{H\,=}$ .71 < .80); subscales can be used and interpreted (Psychological subscale should not be used for high-stakes decisions) ω_{S} ranged from .60 to .82, ECV_{SS} ranged from .23 (psychological sub-score) to .93 c) Tenability of both paradigms to assess PL is discussed: reflective paradigm with more robust toolkit for validity assessment, however formative paradigm is also tenable and could be further researched	a and b) Replication of these findings with a higher sample size outside of restrictions due to the COVID-19 context Study the possibilities of adding the Flexibility indicators to the PPLA system c) Further comparisons and relations among paradigms and conceptual discussion regarding the ontology of PL

Note: points a)-e) are kept consistent across columns to indicate relationship between research aims, methods used, main results and recommendations according to those results
PPLA- Portuguese Physical Literacy Assessment; CVI – Content Validity Index; CTT – Classical Test Theory; IRT – Item Response Theory; PLS – Partial Least Squares; SEM – Structural Equation Modelling; DIF -Differential Item Functioning; DTF – Differential Test Functioning; IIO – Invariant Item Ordering; ICC – Intraclass Correlation Coefficient; MHM – Monotone Homogeneity Model; DMM – Double Monotonicity Model; MCRT – Movement Competence, Rules and Tactics; HRF – Health-related Fitness; PE – Physical Education; MIRT – Multidimensional IRT; LRT – Likelihood Ratio Test; EAP – Expected a posteriori; ECV – Explained Common Variance

This thesis contributions also reinforce two apparently diametrical sides of PL theoretical development. On the operationalization-craving arena, it reinforces the growing body of instruments that can be used to gather empirical support for PL and its impacts; while on the conceptual-driven one, discusses issues that might be paramount to the ontology of PL (i.e., whether PL could/should be regarded as an existing, common cause of the multiple indicators we construe as elements, or whether it is a man-made artifact useful to explain and index a rich, diverse network of variables that might impact meaningful PA participation) and respective implications on its measurement and study. We argue that both these sides should not be mutually exclusive if PL is to be used as an umbrella, multi-disciplinary lens to understand human movement; otherwise, we could be forever shackled to philosophical/conceptual battles, that while rich in metaphors and meaning, will never reach those who need it the most: those who are (not) enjoying movement as part of their lives; or, the opposite: blind progress towards measures that are empty of value towards understanding the whole-picture of meaningful and fruitful interaction with the environment through movement.

Limitations

Nonetheless the strengths and implications reported above, there are some limitations that merit underlining. The most general of these pertains precisely to the flip side of using a rich and interdisciplinary network of constructs: multiple concessions had to be made regarding the essence of each construct to allow for a synthetic perspective transversal to all domains and elements, which could have decreased the accuracy of the overall theoretical framework of each individual construct.

Multiple limitations were caused by the advent of the COVID-19 pandemic, including severe limitations in access to a fully representative sample of students, despite multiple safeguards through stratification by grade and major. These limitations motivated the choice to use schools with PE teacher's preservice protocol with the Faculty of Human Kinetics, and mostly classes led by PE preservice teachers; this could have biased the results mostly in teacher-reported data in the PPLA-O. Other COVID-19 impacts on data quality are also tenable, and as such, replication of these results is in order.

Future Perspectives

Following this thesis' work, multiple threads are open to further work. One of these is motivated by the nature of validity (Cizek, 2020), which requires an on-going gathering of evidence to support PPLA's intended uses; as such, some sources of validity are left unexplored including predictive validity, which could further reinforce it the utility of the construct of PL (as delineated in the APLF) to improvement participation in lifelong PA; and concurrent validity that, could, through comparison with other PL assessment tools, strengthen the collective understanding of PL assessment. Similarly, this initial validation effort, highlighted the need for diverse revisions and improvements which could further improve the measurement quality of the PPLA instruments (summarized in Table 45), especially in the *Tactics* and *Rules* elements, along with eventual augmentation of the *Content Knowledge* test with other PL relevant themes. Also, evaluation of the effectiveness and practical usefulness of the PPLA cut-scores and reports is warranted to guarantee that results resonate with adolescents, teachers, and other relevant stakeholders (e.g., school's directing board, or the PE teacher's group).

Future venues of work also include the adaptation of the PPLA to other age-ranges inside, and outside school-age and settings. This includes adapting the PPLA to assess children and preadolescents, enabling tracking and comparison of the PL development mechanisms, trends, and meanings throughout school-age, working also to strengthen the advocacy of the concept as outlined above. It could also include adaptation of the PPLA to assess and study PL's development in adults – a yet much unexplored research venue – reinforcing its lifelong nature and its added value to society and individual's flourishing. Similarly, multiple expansions to the PPLA could be entailed by developing specific integrated modules to assess PL development in other environments (e.g., water, snow), which would, again, reinforce the transversal nature of this concept across movement and PA settings.

Finally, further work is suggested in a) development of automatic scoring platforms (e.g., web or smartphone applications) to facilitate student's and teacher's access to results without the research team; b) use of Computerized Adaptative Testing (i.e., a promising Item Response Theory method that tailors the test to the individual to allow for more precise, quicker results); c) integration of PPLA within national/international efforts to monitor quality PE practices and longitudinal impact of educational policies; d) expansion of the ontological and methodological discussions presented in chapter 6 of this thesis and its implications for PL.

References

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

Aguirre-Urreta, M. I., Rönkkö, M., & Marakas, G. M. (2016). Omission of Causal Indicators: Consequences and Implications for Measurement. *Measurement: Interdisciplinary Research and Perspectives*, *14*(3), 75–97. https://doi.org/10.1080/15366367.2016.1205935

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

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

Amireault, S., Godin, G., & Vézina-Im, L.-A. (2013). Determinants of physical activity maintenance: A systematic review and meta-analyses. *Health Psychology Review*, 7(1), 55–91. https://doi.org/10.1080/17437199.2012.701060

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8

Arbuckle, J. (1996). Full Information Estimation in the Presence of Incomplete Data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced Structural Equation Modeling*. Psychology Press.

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

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

Armstrong, T. S., Cohen, M. Z., Eriksen, L., & Cleeland, C. (2005). Content Validity of Self-Report Measurement Instruments: An Illustration From the Development of the Brain Tumor Module of the M.D. Anderson Symptom Inventory. *Oncology Nursing Forum*, *32*(3), 669–676. https://doi.org/10.1188/05.ONF.669-676

Artero, E. G., España-Romero, V., Castro-Piñero, J., Ortega, F. B., Suni, J., Castillo-Garzon, M. J., & Ruiz, J. R. (2011). Reliability of field-based fitness tests in youth. *International Journal of Sports Medicine*, *32*(3), 159–169. https://doi.org/10.1055/s-0030-1268488

Artino, A. R., La Rochelle, J. S., Dezee, K. J., & Gehlbach, H. (2014). Developing questionnaires for educational research: AMEE Guide No. 87. *Medical Teacher*, *36*(6), 463–474. https://doi.org/10.3109/0142159X.2014.889814

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.

Avila, M. L., Stinson, J., Kiss, A., Brandão, L. R., Uleryk, E., & Feldman, B. M. (2015). A critical review of scoring options for clinical measurement tools. *BMC Research Notes*, *8*, 612. https://doi.org/10.1186/s13104-015-1561-6

Babic, M. J., Morgan, P. J., Plotnikoff, R. C., Lonsdale, C., White, R. L., & Lubans, D. R. (2014). Physical Activity and Physical Self-Concept in Youth: Systematic Review and Meta-Analysis. *Sports Medicine*, *44*(11), 1589–1601. https://doi.org/10.1007/s40279-014-0229-z

Bagozzi, R. P. (2011). Measurement and Meaning in Information Systems and Organizational Research: Methodological and Philosophical Foundations. *MIS Quarterly*, *35*(2), 261–292. https://doi.org/10.2307/23044044

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R* (1st ed. 2017). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-54205-8

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory* (Vol. xiii). Prentice-Hall, Inc.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. W.H. Freeman and Company.

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

Barnett, L. M., Dudley, D. A., Telford, R. D., Lubans, D. R., Bryant, A. S., Roberts, W. M., Morgan, P. J., Schranz, N. K., Weissensteiner, J. R., Vella, S. A., Salmon, J., Ziviani, J., Okely, A. D., Wainwright, N., Evans, J. R., & Keegan, R. J. (2019). Guidelines for the Selection of Physical Literacy Measures in Physical Education in Australia. *Journal of Teaching in Physical Education*, *38*(2), 119–125. https://doi.org/10.1123/jtpe.2018-0219

Barnett, L. M., Lai, S. K., Veldman, S. L. C., Hardy, L. L., Cliff, D. P., Morgan, P. J., Zask, A., Lubans, D. R., Shultz, S. P., Ridgers, N. D., Rush, E., Brown, H. L., & Okely, A. D. (2016). Correlates of Gross Motor Competence in Children and Adolescents: A Systematic Review and Meta-Analysis. *Sports Medicine (Auckland, N.Z.)*, *46*(11), 1663–1688. https://doi.org/10.1007/s40279-016-0495-z

Barnett, L. M., Mazzoli, E., Hawkins, M., Lander, N., Lubans, D. R., Caldwell, S., Comis, P., Keegan, R. J., Cairney, J., Dudley, D., Stewart, R. L., Long, G., Schranz, N., Brown, T. D., & Salmon, J. (2020). Development of a self-report scale to assess children's perceived physical literacy. *Physical Education and Sport Pedagogy*, *0*(0), 1–26. https://doi.org/10.1080/17408989.2020.1849596

Barnett, L. M., van Beurden, E., Morgan, P. J., Brooks, L. O., & Beard, J. R. (2009). Childhood Motor Skill Proficiency as a Predictor of Adolescent Physical Activity. *Journal of Adolescent Health*, 44(3), 252–259. https://doi.org/10.1016/j.jadohealth.2008.07.004

Baumeister, R. F., & Vohs, K. D. (2007). Self-Regulation, Ego Depletion, and Motivation: Motivation and Ego Depletion. *Social and Personality Psychology Compass*, *1*(1), 115–128. https://doi.org/10.1111/j.1751-9004.2007.00001.x

Belanger, K., Barnes, J. D., Longmuir, P. E., Anderson, K. D., Bruner, B., Copeland, J. L., Gregg, M. J., Hall, N., Kolen, A. M., Lane, K. N., Law, B., MacDonald, D. J., Martin, L. J., Saunders, T. J., Sheehan, D., Stone, M., Woodruff, S. J., & Tremblay, M. S. (2018). The relationship between physical literacy scores and adherence to Canadian physical activity and sedentary behaviour guidelines. *BMC Public Health*, *18*(S2), 1042. https://doi.org/10.1186/s12889-018-5897-4

Benitez, J., Henseler, J., Castillo, A., & Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. *Information & Management*, 57(2), 103168. https://doi.org/10.1016/j.im.2019.05.003

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

Bernaards, C. A., & Sijtsma, K. (1999). Factor Analysis of Multidimensional Polytomous Item Response Data Suffering From Ignorable Item Nonresponse. *Multivariate Behavioral Research*, *34*(3), 277–313. https://doi.org/10.1207/S15327906MBR3403_1

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

Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of Observed Learning Outcomes)*. Academic Press.

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

Bisi, M. C., Pacini Panebianco, G., Polman, R., & Stagni, R. (2017). Objective assessment of movement competence in children using wearable sensors: An instrumented version of the TGMD-2 locomotor subtest. *Gait & Posture*, *56*, 42–48. https://doi.org/10.1016/j.gaitpost.2017.04.025

Blanchard, J., Van Wyk, N., Ertel, E., Alpous, A., & Longmuir, P. E. (2020). Canadian Assessment of Physical Literacy in grades 7-9 (12-16 years): Preliminary validity and descriptive results. *Journal of Sports Sciences*, *38*(2), 177–186. https://doi.org/10.1080/02640414.2019.1689076

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*. https://doi.org/10.3389/fpubh.2018.00149

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

Bock, R. D., & Gibbons, R. D. (2021). *Item response theory* (First edition). Wiley.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, *53*(1), 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Bollen, K. A. (2007). Interpretational confounding is due to misspecification, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods*, *12*(2), 219–228; discussion 238-245. https://doi.org/10.1037/1082-989X.12.2.219

Bollen, K. A., & Bauldry, S. (2011). Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581–596. https://doi.org/10.1037/met0000056

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

Boreham, C., & Riddoch, C. (2001). The physical activity, fitness and health of children. *Journal of Sports Sciences*, *19*(12), 915–929. https://doi.org/10.1080/026404101317108426

Borsboom, D., Mellenbergh, G. J., & Heerden, J. V. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Britton, U., Issartel, J., Symonds, J., & Belton, S. (2020). What Keeps Them Physically Active? Predicting Physical Activity, Motor Competence, Health-Related Fitness, and Perceived Competence in Irish Adolescents after the Transition from Primary to Second-Level School. *International Journal of Environmental Research and Public Health*, *17*(8), E2874. https://doi.org/10.3390/ijerph17082874

Bronikowska, M., Korcz, A., Pluta, B., Krzysztoszek, J., Ludwiczak, M., Łopatka, M., Wawrzyniak, S., Kowalska, J. E., & Bronikowski, M. (2019). Fair Play in Physical Education and Beyond. *Sustainability*, *11*(24), 7064. https://doi.org/10.3390/su11247064

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brooks, A. W., Schroeder, J., Risen, J. L., Gino, F., Galinsky, A. D., Norton, M. I., & Schweitzer, M. E. (2016). Don’t stop believing: Rituals improve performance by decreasing anxiety. *Organizational Behavior and Human Decision Processes*, *137*, 71–85.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.

Bunker, D., & Thorpe, R. (1982). A model for the teaching of games in the secondary school. *Bulletin of Physical Education*, *10*, 9–16.

Burrell, G., & Morgan, G. (2019). *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life* (Reprint). Routledge.

Burton, A. W., & Rodgerson, R. W. (2001). New Perspectives on the Assessment of Movement Skills and Motor Abilities. *Adapted Physical Activity Quarterly*, *18*(4), 347–365. https://doi.org/10.1123/apaq.18.4.347

Bushman, B. A., & American College of Sports Medicine (Eds.). (2017). *ACSM's complete guide to fitness & health* (Second edition). Human Kinetics.

Butler, J., & Griffin, L. L. (Eds.). (2010). *More teaching games for understanding: Moving globally*. Human Kinetics.

Cadogan, J. W., & Lee, N. (2013). Improper use of endogenous formative variables. *Journal of Business Research*, *66*(2), 233–241. https://doi.org/10.1016/j.jbusres.2012.08.006

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

Cairney, J., Clark, H., Dudley, D., & Kriellaars, D. (2019). Physical Literacy in Children and Youth—A Construct Validation Study. *Journal of Teaching in Physical Education*, *38*(2), 84–90. https://doi.org/10.1123/jtpe.2018-0270

Cairney, J., Veldhuizen, S., Graham, J. D., Rodriguez, C., Bedard, C., Bremer, E., & Kriellaars, D. (2018). A Construct Validation Study of PLAYfun. *Medicine & Science in Sports & Exercise*, *50*(4), 855–862. https://doi.org/10.1249/MSS.0000000000001494

Caspersen, C. J., Powell, K. E., & Christenson, G. M. (1985). Physical activity, exercise, and physical fitness: Definitions and distinctions for health-related research. *Public Health Reports*, *100*(2), 126.

Cattuzzo, M. T., Dos Santos Henrique, R., Ré, A. H. N., de Oliveira, I. S., Melo, B. M., de Sousa Moura, M., de Araújo, R. C., & Stodden, D. (2016). Motor competence and health related physical fitness in youth: A systematic review. *Journal of Science and Medicine in Sport*, *19*(2), 123–129. https://doi.org/10.1016/j.jsams.2014.12.004

Cenfetelli & Bassellier. (2009). Interpretation of Formative Measurement in Information Systems Research. *MIS Quarterly*, *33*(4), 689. https://doi.org/10.2307/20650323

Chaddock-Heyman, L., Hillman, C. H., Cohen, N. J., & Kramer, A. F. (2014). III. The importance of physical activity and aerobic fitness for cognitive control and memory in children. *Monographs of the Society for Research in Child Development*, *79*(4), 25–50. https://doi.org/10.1111/mono.12129

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

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

Chan, E. K. H. (2014). Standards and Guidelines for Validation Practices: Development and Evaluation of Measurement Instruments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 9–24). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_2

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

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods & Research*, *36*(4), 462–494. https://doi.org/10.1177/0049124108314720

Chen, F. F., West, S., & Sousa, K. (2006). A Comparison of Bifactor and Second-Order Models of Quality of Life. *Multivariate Behavioral Research*, *41*(2), 189–225. https://doi.org/10.1207/s15327906mbr4102_5

Chen, S.-T., Tang, Y., Chen, P.-J., & Liu, Y. (2020). The Development of Chinese Assessment and Evaluation of Physical Literacy (CAEPL): A Study Using Delphi Method. *International Journal of Environmental Research and Public Health*, *17*(8), 2720. https://doi.org/10.3390/ijerph17082720

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

Cho, E. (2016). Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organizational Research Methods*, *19*(4), 651–682. https://doi.org/10.1177/1094428116656239

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

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use* (1st ed.). Routledge.

Clark, L., & Watson, D. (1995). Constructing Validity: Basic Issues in Objective Scale Development. *Psychological Assessment*, 7(3), 309–319.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research*, *61*(12), 1250–1262. https://doi.org/10.1016/j.jbusres.2008.01.013

Committee on Fitness Measures and Health Outcomes in Youth, Food and Nutrition Board, & Institute of Medicine. (2012). *Fitness Measures and Health Outcomes in Youth* (R. Pate, M. Oria, & L. Pillsbury, Eds.). National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK241315/

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Psychology Press.

Comtois, D. (2021). *Summarytools* (R package 1.0.0) [Computer software]. https://CRAN.R-project.org/package=summarytools

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, *12*(1), 19–24. https://doi.org/10.1016/S1322-7696(08)60478-3

Corbin, C. B. (2016a). Implications of physical literacy for research and practice: A commentary. *Research Quarterly for Exercise and Sport*, *87*, 14–27. https://doi.org/10.1080/02701367.2016.1124722

Corbin, C. B. (2016b). 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

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

Cox, A., Duncheon, N., & McDavid, L. (2009). Peers and Teachers as Sources of Relatedness Perceptions, Motivation, and Affective Responses in Physical Education. *Research Quarterly for Exercise and Sport*, *80*(4), 765–773. https://doi.org/10.1080/02701367.2009.10599618

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

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (Fifth edition). SAGE.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Crum, B. (1993). Conventional Thought and Practice in Physical Education: Problems of Teaching and Implications for Change. *Quest*, *45*(3), 339–356. https://doi.org/10.1080/00336297.1993.10484092

Dana, J., & Dawes, R. M. (2004). The Superiority of Simple Alternatives to Regression for Social Science Predictions. *Journal of Educational and Behavioral Statistics*, *29*(3), 317–331.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4

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

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Deci, E. L., & Ryan, R. M. (Eds.). (2002). *Handbook of Self-Determination Research* (1st edition). University of Rochester Press.

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie Canadienne*, *49*(3), 182–185. https://doi.org/10.1037/a0012801

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

DeMars, C. E. (2013). A Tutorial on Interpreting Bifactor Model Scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

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

DeVellis, R. (2017). *Scale Development: Theory and Applications* (4th ed.). SAGE Publications Ltd. https://us.sagepub.com/en-us/nam/scale-development/book246123

Diamantopoulos, A. (2008). Formative indicators: Introduction to the special issue. *Journal of Business Research*, *61*(12), 1201–1202. https://doi.org/10.1016/j.jbusres.2008.01.008

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management*, *17*(4), 263–282. https://doi.org/10.1111/j.1467-8551.2006.00500.x

Díaz-Cueto, M., Hernández-Álvarez, J. L., & Castejón, F. J. (2010). Teaching Games for Understanding to In-Service Physical Education Teachers: Rewards and Barriers Regarding the Changing Model of Teaching Sport. *Journal of Teaching in Physical Education*, *29*(4), 378–398. https://doi.org/10.1123/jtpe.29.4.378

Dima, A. L. (2018). Scale validation in applied health research: Tutorial for a 6-step R-based psychometrics protocol. *Health Psychology and Behavioral Medicine*, *6*(1), 136–161. https://doi.org/10.1080/21642850.2018.1472602

Direção-Geral da Educação & Faculdade de Motricidade Humana. (2015). *FITescola*. https://fitescola.dge.mec.pt/home.aspx

DiStefano, C., Zhu, M., & Mîndrilã, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research, and Evaluation*, *14*(20). https://doi.org/10.7275/DA8T-4G52

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2. https://doi.org/10.1186/2193-1801-2-222

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

Dudley, D. (2015). A Conceptual Model of Observed Physical Literacy. *The Physical Educator*, 72(5), 236–260. https://doi.org/10.18666/TPE-2015-V72-I5-6020

Dudley, D., Cairney, J., Wainwright, N., Kriellaars, D., & Mitchell, D. (2017). Critical Considerations for Physical Literacy Policy in Public Health, Recreation, Sport, and Education Agencies. *Quest*, *69*(4), 436–452. https://doi.org/10.1080/00336297.2016.1268967

Dudley, D., Keegan, R., & Barnett, L. (2017). *Physical Literacy: Informing a Definition and Standard for Australia*. https://www.researchgate.net/publication/321310128_Physical_Literacy_Informing_a_Definition_and_Standard_for_Australia

Dueber, D. (2020). *A Bifactor Approach to Dimensionality Assessment* [University of Kentucky Libraries]. https://doi.org/10.13023/ETD.2020.078

Dueber, D. (2021). *BifactorIndicesCalculator* (R Package version 0.2.2) [Computer software]. https://CRAN.R-project.org/package=BifactorIndicesCalculator

Durden-Myers, E. J., Green, N. R., & Whitehead, M. E. (2018). Implications for Promoting Physical Literacy. *Journal of Teaching in Physical Education*, *37*(3), 262–271. https://doi.org/10.1123/jtpe.2018-0131

Earley, P. C., & Ang, S. (2003). *Cultural Intelligence: Individual Interactions Across Cultures*. Stanford University Press.

Eastman, S. T., & Riggs, K. E. (1994). Televised Sports and Ritual: Fan Experiences. *Sociology of Sport Journal*, *11*(3), 249–274. https://doi.org/10.1123/SSJ.11.3.249

Ebel, R., & Frisbie, D. (1991). *Essentials of Educational Measurement*. Prentice-Hall, Inc. https://en.pt1lib.org/book/907387/7d3493

Edwards, J. R. (2011). The Fallacy of Formative Measurement. *Organizational Research Methods*, *14*(2), 370–388. https://doi.org/10.1177/1094428110378369

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Edwards, L., Bryant, A., Keegan, R., Morgan, K., Cooper, S., & Jones, A. (2017). 'Measuring' Physical Literacy and Related Constructs: A Systematic Review of Empirical Findings. *Sports Medicine*, *48*(3), 659–682. https://doi.org/10.1007/s40279-017-0817-9

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

Eid, M. (2020). Multi-Faceted Constructs in Abnormal Psychology: Implications of the Bifactor S - 1 Model for Individual Clinical Assessment. *Journal of Abnormal Child Psychology*, *48*(7), 895–900. https://doi.org/10.1007/s10802-020-00624-9

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 341–349.

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

Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (Eds.). (2010). *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-32827-8

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

Evermann, J., & Rönkkö, M. (2021). Recent Developments in PLS. *Communications of the Association for Information Systems*, *44*, 123–133.

Faught, B. E., Cairney, J., Hay, J., Veldhuizen, S., Missiuna, C., & Spironello, C. A. (2008). Screening for motor coordination challenges in children using teacher ratings of physical ability and activity. *Human Movement Science*, *27*(2), 177–189. https://doi.org/10.1016/j.humov.2008.02.001

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

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

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Fox, J. (2019). *polycor: Polychoric and Polyserial Correlations* (R package version 0.7-10) [Computer software]. https://CRAN.R-project.org/package=polycor

Francis, C. E., Longmuir, P. E., Boyer, C., Andersen, L. B., Barnes, J. D., Boiarskaia, E., Cairney, J., Faigenbaum, A. D., Faulkner, G., Hands, B. P., Hay, J. A., Janssen, I., Katzmarzyk, P. T., Kemper, H. C. G., Knudson, D., Lloyd, M., McKenzie, T. L., Olds, T. S., Sacheck, J. M., … Tremblay, M. S. (2016). The Canadian Assessment of Physical Literacy: Development of a Model of Children's Capacity for a Healthy, Active Lifestyle Through a Delphi Process. *Journal of Physical Activity and Health*, *13*(2), 214–222. https://doi.org/10.1123/jpah.2014-0597

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

Furr, R. M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. SAGE Publications Ltd. https://doi.org/10.4135/9781446287866

Gagné, M., Forest, J., Gilbert, M.-H., Aubé, C., Morin, E., & Malorni, A. (2010). The Motivation at Work Scale: Validation Evidence in Two Languages. *Educational and Psychological Measurement*, *70*(4), 628–646. https://doi.org/10.1177/0013164409355698

Gallahue, D. L. (1996). *Developmental Physical Education for Today's Children* (3rd ed.). Brown & Benchmark.

Gallahue, D. L., Goodway, J., & Ozmun, J. C. (2020). *Understanding motor development: Infants, children, adolescents, adults* (Eighth edition). Jones & Bartlett Learning.

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

Gamerman, D., Gonçalves, F. B., & Soares, T. M. (2019). Differential Item Functioning. In *Handbook of Item Response Theory Volume 3*. Chapman and Hall/CRC.

Gana, K., & Broc, G. (2019). *Structural equation modeling with lavaan*. ISTE Ltd.

Gandrieau, J., Schnitzler, C., Derigny, T., & Potdevin, F. (2021). Évaluation de la littératie physique: Création d'un outil de mesure pour les jeunes adultes. In *SEPAPS 2020*. https://doi.org/10.25518/sepaps20.485

Gibbs, J. C. (2014). *Moral development and reality: Beyond the theories of Kohlberg, Hoffman, and Haidt* (Third edition). Oxford University Press.

Giblin, S., Collins, D., & Button, C. (2014). Physical Literacy: Importance, Assessment and Future Directions. *Sports Medicine*, *44*(9), 1177–1184. https://doi.org/10.1007/s40279-014-0205-7

Goleman, D. (2005). *Emotional Intelligence: Why It Can Matter More Than IQ* (10th Anniversary edition). Bantam.

Goodway, J., Ozmun, J. C., & Gallahue, D. (2020). *Understanding motor development: Infants, children, adolescents, adults* (Eighth edition). Jones & Bartlett Learning.

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

Gréhaigne, J.-F., & Godbout, P. (2013). *Collective Variables for Analysing Performance in Team Sports*. Routledge Handbooks Online. https://doi.org/10.4324/9780203806913.ch9

Gréhaigne, J.-F., & Godbout, P. (2014). Dynamic Systems Theory and Team Sport Coaching. *Quest*, *66*(1), 96–116. https://doi.org/10.1080/00336297.2013.814577

Gréhaigne, J.-F., Godbout, P., & Bouthier, D. (1997). Performance Assessment in Team Sports. *Journal of Teaching in Physical Education*, *16*(4), 500–516. https://doi.org/10.1123/jtpe.16.4.500

Gréhaigne, J.-F., Richard, J.-F., & Griffin, L. L. (2005). *Teaching and learning team sports and games*. RoutledgeFalmer.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450.

Griffiths, A., Toovey, R., Morgan, P. E., & Spittle, A. J. (2018). Psychometric properties of gross motor assessment tools for children: A systematic review. *BMJ Open*, *8*(10), e021734. https://doi.org/10.1136/bmjopen-2018-021734

Gunnell, K. E., Longmuir, P. E., Barnes, J. D., Belanger, K., & Tremblay, M. S. (2018). Refining the Canadian Assessment of Physical Literacy based on theory and factor analyses. *BMC Public Health*, *18 (Suppl 2)*, 131–145. https://doi.org/10.1186/s12889-018-5899-2

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

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

Guyon, H., & Tensaout, M. (2016). Are Formative Indicators Superfluous? An Extension of Aguirre-Urreta, Rönkkö, and Marakas Analysis. *Measurement: Interdisciplinary Research and Perspectives*, *14*, 101–104. https://doi.org/10.1080/15366367.2016.1224966

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

Hair, J. F., Hult, G., Ringle, C., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)* (Second edition). Sage.

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, *31*(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203

Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (Eighth edition). Cengage.

Hair Jr., J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. (2018). *Advanced Issues in Partial Least Squares Structural Equation Modelling*. SAGE Publications.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed). Lawrence Erlbaum Associates.

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

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

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *Curriculum Journal*, *16*(2), 207–223. https://doi.org/10.1080/09585170500136093

Harlen, W. (2007). *Assessment of Learning*. SAGE Publications Ltd.

Harlen, W. (2009). Improving assessment of learning and for learning. *Education 3-13*, *37*(3), 247–257. https://doi.org/10.1080/03004270802442334

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

Hassandra, M., Goudas, M., & Hatzigeorgiadis, A. (2002). Development of a questionnaire assessing fair play in elementary school physical education. *Athlitki Psychol*, *13*, 105–126.

Hassandra, M., Goudas, M., & Hatzigeorgiadis, A. (2003). Attitudes towards fair play in physical education: The role of intrinsic motivation and gender. *Proceedings, XIth European Congress of Sport Psychology*. https://doi.org/10.1037/e547922012-140

Hassandra, M., Goudas, M., Hatzigeorgiadis, A., & Theodorakis, Y. (2007). A fair play intervention program in school Olympic education. *European Journal of Psychology of Education*, 22(2), 99. https://doi.org/10.1007/BF03173516

Hay, J., & Donnelly, P. (1996). Sorting out the boys from the girls: Teacher and student perceptions of student physical ability. *Avante*, 2, 26–52.

Hellison, D. (2011). *Teaching Personal and Social Responsibility Through Physical Activity* (3rd edition). Human Kinetics.

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

Henseler, J. (2018). Partial least squares path modeling: Quo vadis? *Quality & Quantity*, 52(1), 1–8. https://doi.org/10.1007/s11135-018-0689-6

Henseler, J. (2021). *Composite-based structural equation modeling: Analyzing latent and emergent variables*. The Guilford Press.

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., Ketchen, D. J., Hair, J. F., Hult, G. T. M., & Calantone, R. J. (2014). Common Beliefs and Reality About PLS: Comments on Rönkkö and Evermann (2013). *Organizational Research Methods*, *17*(2), 182–209. https://doi.org/10.1177/1094428114526928

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health*, *31*(2), 180–191. https://doi.org/10.1002/nur.20247

Hervé, M. (2021). *RVAideMemoire: Testing and Plotting Procedures for Biostatistics* (R package version 0.9-80) [Computer software].

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (Vol. 663). Houghton Mifflin College Division.

Holfelder, B., & Schott, N. (2014). Relationship of fundamental movement skills and physical activity in children and adolescents: A systematic review. *Psychology of Sport and Exercise*, *15*(4), 382–391. https://doi.org/10.1016/j.psychsport.2014.03.005

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, 2(1), 41–54. https://doi.org/10.1007/BF02287965

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

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

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205–218. https://doi.org/10.1037/1082-989X.12.2.205

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, *5.1*, 221–234.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, *103*(2), 219–230. https://doi.org/10.1007/s11205-011-9843-4

Hughes, A., Galbraith, D., & White, D. (2011). Perceived Competence: A Common Core for Self-Efficacy and Self-Concept? *Journal of Personality Assessment*, *93*(3), 278–289. https://doi.org/10.1080/00223891.2011.559390

Hulteen, R. M., Barnett, L. M., True, L., Lander, N. J., del Pozo Cruz, B., & Lonsdale, C. (2020). Validity and reliability evidence for motor competence assessments in children and adolescents: A systematic review. *Journal of Sports Sciences*, *38*(15), 1717–1798. https://doi.org/10.1080/02640414.2020.1756674

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00045

International Physical Literacy Association (IPLA). (2017). *IPLA definition*. https://www.physical-literacy.org.uk/

IPAQ Research Commitee. (2005). *Guidelines for Interpreting the IPAQ*.

Janssen, A., Leahy, A. A., Diallo, T. M. O., Smith, J. J., Kennedy, S. G., Eather, N., Mavilidi, M. F., Wagemakers, A., Babic, M. J., & Lubans, D. R. (2020). Cardiorespiratory fitness, muscular fitness and mental health in older adolescents: A multi-level cross-sectional analysis. *Preventive Medicine*, *132*, 105985. https://doi.org/10.1016/j.ypmed.2020.105985

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. *Journal of Consumer Research*, *30*(2), 199–218. https://doi.org/10.1086/376806

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00109

Johanson, G. A., & Brooks, G. P. (2010). Initial Scale Development: Sample Size for Pilot Studies. *Educational and Psychological Measurement*, *70*(3), 394–400. https://doi.org/10.1177/0013164409355692

Jonckere, J. D., & Roseel, Y. (2021). *Using bounded estimation to avoid nonconvergence in small sample structural equation modeling*. https://osf.io/f7z6j/

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

Keegan, R., Barnett, L., & Dudley, D. (2017). *Literature sampling to inform development of a physical literacy definition and standard for Australia*.

Keegan, R., Barnett, L. M., Dudley, D. A., Telford, R. D., Lubans, D. R., Bryant, A. S., Roberts, W. M., Morgan, P. J., Schranz, N. K., Weissensteiner, J. R., Vella, S. A., Salmon, J., Ziviani, J., Okely, A. D., Wainwright, N., & Evans, J. R. (2019). Defining Physical Literacy for Application in Australia: A Modified Delphi Method. *Journal of Teaching in Physical Education*, *38*(2), 105–118. https://doi.org/10.1123/jtpe.2018-0264

Kemper, H. C. G., & Koppes, L. L. J. (2006). Linking Physical Activity and Aerobic Fitness: Are We Active Because We Are Fit, or Are We Fit Because We Are Active? *Pediatric Exercise Science*, *18*(2), 173–181. https://doi.org/10.1123/pes.18.2.173

Kim, Y., Park, I., & Kang, M. (2013). Convergent validity of the international physical activity questionnaire (IPAQ): Meta-analysis. *Public Health Nutrition*, *16*(3), 440–452. https://doi.org/10.1017/S1368980012002996

Klein, O., & Delacre, M. (2021). *Routliers: Robust Outliers Detection* (R package version 0.0.0.3) [Computer software].

Kline, P. (2000). *Handbook of Psychological Testing* (2nd ed.). Routledge.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth edition). The Guilford Press.

Kock, N. (2020). *WarpPLS* (7.0) [Computer software]. ScriptWarp Systems.

Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods: Sample size in PLS-based SEM. *Information Systems Journal*, *28*(1), 227–261. https://doi.org/10.1111/isj.12131

Kohlberg, L. (1964). Development of Moral Character and Moral Ideology. In L. W. Hoffman & M. L. Hoffman (Eds.), *Review of Child Development Research: Volume 1*. Russell Sage Foundation.

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

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, *6*(2), 151–162.

Krathwohl, Bloom, & Masia (Eds.). (1964). *Taxonomy of education objectives: The classification of education goals: Handbook 2—Affective domain*. David McKay.

Kyriazos, T. A., & Stalikas, A. (2018). Applied Psychometrics: The Steps of Scale Development and Standardization Process. *Psychology*, *09*(11), 2531–2560. https://doi.org/10.4236/psych.2018.911145

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

Law, K. S., Wong, C.-S., & Mobley, W. H. (1998). Toward a Taxonomy of Multidimensional Constructs. *The Academy of Management Review*, *23*(4), 741. https://doi.org/10.2307/259060

Lazarus, R. S., & Folkman, S. (1984). *Stress, Appraisal, and Coping* (1st edition). Springer Publishing Company.

Lee, N., & Cadogan, J. W. (2013). Problems with formative and higher-order reflective variables. *Journal of Business Research*, *66*(2), 242–247. https://doi.org/10.1016/j.jbusres.2012.08.004

Lee, P. H., Macfarlane, D. J., Lam, T. H., & Stewart, S. M. (2011). Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): A systematic review. *The International Journal of Behavioral Nutrition and Physical Activity*, *8*, 115. https://doi.org/10.1186/1479-5868-8-115

Leptokaridou, E. T., Vlachopoulos, S. P., & Papaioannou, A. G. (2015). Associations of Autonomy, Competence, and Relatedness with Enjoyment and Effort in Elementary School Physical Education: The Mediating Role of Self-Determined Motivation. *Hellenic Journal of Psychology*, *12*, 105–128.

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Psychology*, *32*(1), 5. https://doi.org/10.5334/irsp.289

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

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

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

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

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

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

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd edition). Wiley.

Liu, Y., & Chen, S. (2021). Physical literacy in children and adolescents: Definitions, assessments, and interventions. *European Physical Education Review*, 27(1), 96–112. https://doi.org/10.1177/1356336X20925502

Logan, S. W., Ross, S. M., Chee, K., Stodden, D. F., & Robinson, L. E. (2018). Fundamental motor skills: A systematic review of terminology. *Journal of Sports Sciences*, *36*(7), 781–796. https://doi.org/10.1080/02640414.2017.1340660

Longmuir, P. E., Boyer, C., Lloyd, M., Yang, Y., Boiarskaia, E., Zhu, W., & Tremblay, M. S. (2015). The Canadian Assessment of Physical Literacy: Methods for children in grades 4 to 6 (8 to 12 years). *BMC Public Health*, *15*(1), 767. https://doi.org/10.1186/s12889-015-2106-6

Longmuir, P. E., Gunnell, K. E., Barnes, J. D., Belanger, K., Leduc, G., Woodruff, S. J., & Tremblay, M. S. (2018). Canadian Assessment of Physical Literacy Second Edition: A streamlined assessment of the capacity for physical activity among children 8 to 12 years of age. *BMC Public Health*, *18 (Suppl 2)*, 169–180. https://doi.org/10.1186/s12889-018-5902-y

Longmuir, P. E., & Tremblay, M. S. (2016). Top 10 Research Questions Related to Physical Literacy. *Research Quarterly for Exercise and Sport*, *87*(1), 28–35. https://doi.org/10.1080/02701367.2016.1124671

Longmuir, P. E., Woodruff, S. J., Boyer, C., Lloyd, M., & Tremblay, M. S. (2018). Physical Literacy Knowledge Questionnaire: Feasibility, validity, and reliability for Canadian children aged 8 to 12 years. *BMC Public Health*, *18 (Suppl 2)*, 19–29. https://doi.org/10.1186/s12889-018-5890-y

Lopes, C., Torres, D., Oliveira, A., Severo, M., Alarcão, V., Guiomar, S., Mota, J., Teixeira, P., Ramos, E., Rodrigues, S., Vilela, S., Oliveira, L., Nicola, P., Soares, S., Andersen, L. F., & Consórcio IAN-AF. (2017). *Inquérito Alimentar Nacional e de Atividade Física, IAN- AF 2015-2016: Relatório Metodológico*. Universidade do Porto. www.ian-af.up.pt

Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, *17*(2), 181–194. https://doi.org/10.1007/BF02288781

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Information Age Publ.

Lubans, D. R., & Cliff, D. P. (2011). Muscular fitness, body composition and physical self-perception in adolescents. *Journal of Science and Medicine in Sport*, *14*(3), 216–221. https://doi.org/10.1016/j.jsams.2010.10.003

Lubans, D. R., Morgan, P., Callister, R., Plotnikoff, R. C., Eather, N., Riley, N., & Smith, C. J. (2011). Test-retest reliability of a battery of field-based health-related fitness measures for adolescents. *Journal of Sports Sciences*, *29*(7), 685–693. https://doi.org/10.1080/02640414.2010.551215

Lubans, D. R., Morgan, P. J., Cliff, D. P., Barnett, L. M., & Okely, A. D. (2010). Fundamental Movement Skills in Children and Adolescents: Review of Associated Health Benefits. *Sports Medicine*, *40*(12), 1019–1035. https://doi.org/10.2165/11536850-000000000-00000

Luz, C., Rodrigues, L. P., Almeida, G., & Cordovil, R. (2016). Development and validation of a model of motor competence in children and adolescents. *Journal of Science and Medicine in Sport*, *19*(7), 568–572. https://doi.org/10.1016/j.jsams.2015.07.005

Luz, C., Rodrigues, L. P., Meester, A. D., & Cordovil, R. (2017). The relationship between motor competence and health-related fitness in children and adolescents. *PloS One*, *12*(6), e0179993. https://doi.org/10.1371/journal.pone.0179993

Lynn, M. R. (1986). Determination and Quantification Of Content Validity. *Nursing Research*, *35*(6), 382–386.

Mardia, K. V. (1974). Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, *36*(2), 115–128.

Markland, D., & Tobin, V. (2004). A modification to the Behavioural Regulation in Exercise Questionnaire to include an assessment of amotivation. *Journal of Sport & Exercise Psychology*, *26*(2), 191–196. https://doi.org/10.1123/jsep.26.2.191

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. Routledge. https://doi.org/10.4324/9780203501207

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

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). 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. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Martínez-Vizcaíno, V., & Sánchez-López, M. (2008). Relationship Between Physical Activity and Physical Fitness in Children and Adolescents. *Revista Española de Cardiología*, *61*(2), 108–111. https://doi.org/10.1016/S1885-5857(08)60084-5

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

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

Martins, J., Marques, A., Loureiro, N., Carreiro da Costa, F., Diniz, J., & Gaspar de Matos, M. (2019). Trends and Age-Related Changes of Physical Activity Among Portuguese Adolescent Girls From 2002–2014: Highlights From the Health Behavior in School-Aged Children Study. *Journal of Physical Activity and Health*, *16*(4), 281–287. https://doi.org/10.1123/jpah.2018-0092

Martins, J., Marques, A., Sarmento, H., & Carreiro da Costa, F. (2015). Adolescents' perspectives on the barriers and facilitators of physical activity: A systematic review of qualitative studies. *Health Education Research*, *30*(5), 742–755. https://doi.org/10.1093/her/cyv042

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

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

Mayorga-Vega, D., Aguilar-Soto, P., & Viciana, J. (2015). Criterion-Related Validity of the 20-M Shuttle Run Test for Estimating Cardiorespiratory Fitness: A Meta-Analysis. *Journal of Sports Science & Medicine*, *14*(3), 536–547.

Mayorga-Vega, D., Merino-Marban, R., & Viciana, J. (2014). Criterion-Related Validity of Sit-and-Reach Tests for Estimating Hamstring and Lumbar Extensibility: A Meta-Analysis. *Journal of Sports Science & Medicine*, *13*(1), 1–14.

Mazurkiewicz, M. (2011). Some observations about ritual in sport. *Studies in Physical Culture and Tourism*, *18*(4), 12.

McBride, R. E., & Xiang, P. (2004). Thoughtful Decision Makingin Physical Education: A Modest Proposal. *Quest*, *56*(3), 337–354. https://doi.org/10.1080/00336297.2004.10491830

McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

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

Memmert, D., & Harvey, S. (2008). The Game Performance Assessment Instrument (GPAI): Some Concerns and Solutions for Further Development. *Journal of Teaching in Physical Education*, 27, 220–240. https://doi.org/10.1123/jtpe.27.2.220

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

Ministério da Educação. (2001b). *Programa Nacional Educação Física (Reajustamento): Ensino Básico 3ºCiclo*. DEB.

Ministério da Educação. (2005). *Programa Nacional Educação Física (reajustamento): Ensino Básico 2ºCiclo*. DEB.

Ministério da Educação. (2018a). *Aprendizagens Essenciais: 12º ano*. https://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/12_educacao_fisica.pdf

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

Ministério da Educação [Ministry of Education]. (2019). *Infoescolas—Estatísticas do Ensino Básico e Secundário*. http://infoescolas.mec.pt/

Mohammadzadeh, M., Sheikh, M., Houminiyan Sharif Abadi, D., Bagherzadeh, F., & Kazemnejad, A. (2021). Design and psychometrics evaluation of Adolescent Physical Literacy Questionnaire (APLQ). *Sport Sciences for Health*, 1–9. https://doi.org/10.1007/s11332-021-00818-8

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. DE GRUYTER MOUTON. https://doi.org/10.1515/9783110813203

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

Molenaar, I. W. (1990). A Weighted Loevinger H-coefficient Extending Mokken Scaling to Multicategory Items. *Kwantitatieve Methoden*, *12*(37), 97–117.

Molenaar, I. W. (1997). *Nonparametric Models for Polytomous Responses* (W. van der Linden & R. Hambleton, Eds.; pp. 369–380). Springer-Verlag.

Molenaar, I. W., & Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschrift Voor Onderwijsresearch*, 9(5), 257–268.

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.

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

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

Mota, J., Martins, J., & Onofre, M. (2022a). *Portuguese Physical Literacy Assessment Observation (PPLA-O) for adolescents (15-18 years) from grades 10-12: Development and initial validation through Item Response Theory*. Research Square. https://doi.org/10.21203/rs.3.rs-1488826/v1

Mota, J., Martins, J., & Onofre, M. (2022b). *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*. Research Square. https://doi.org/10.21203/rs.3.rs-1458688/v2

Mota, J., Martins, J., & Onofre, M. (2022c). *Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) for adolescents (15-18 years) from grades 10-12: Validity and reliability evidence of the Psychological and Social modules using Mokken Scale Analysis*. Research Square. https://doi.org/10.21203/rs.3.rs-1458709/v4

Murdoch, E., & Whitehead, M. (2010). Physical literacy, fostering the attributes and curriculum planning. In *Physical Literacy: Through the Lifecourse* (pp. 175–188). Routledge.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed). Pearson/Prentice Hall.

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

Nering, M. L. (Ed.). (2010). *Handbook of polytomous item response theory models*. Routledge.

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

Nielson, W. R., Jensen, M. P., Karsdorp, P. A., & Vlaeyen, J. W. S. (2013). Activity Pacing in Chronic Pain: Concepts, Evidence, and Future Directions. *The Clinical Journal of Pain*, *29*(5), 461–468. https://doi.org/10.1097/AJP.0b013e3182608561

Ntoumanis, N., & Myers, N. (Eds.). (2016). *An introduction to intermediate and advanced statistical analyses for sport and exercise scientists*. John Wiley & Sons Inc.

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

Onofre, M. (2017). A Qualidade da Educação Física como Essência da Promoção de uma Cidadania Ativa e Saudável. *Retos: Nuevas Tendencias En Educación Física, Deporte y Recreación*, *31*, 328–333.

Onofre, M., Costa, J., Martins, J., & Quitério, Ana. (2020). Physical Education and School Sport in Portugal. In R. Naul & C. Scheuer (Eds.), *Research on Physical Education and School Sport in Europe*. Meyer & Meyer.

Organisation for Economic Co-operation and Development. (2018). *The Future of Education and Skills: Education 2030*. Organisation for Economic Co-operation and Development. https://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf

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

Orlando, M., & Thissen, D. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied*

Psychological Measurement, 27(4), 289–298. https://doi.org/10.1177/0146621603027004004

Ortega, F. B., Ruiz, J. R., Castillo, M. J., & Sjöström, M. (2008). Physical fitness in childhood and adolescence: A powerful marker of health. *International Journal of Obesity*, *32*(1), 1–11. https://doi.org/10.1038/sj.ijo.0803774

Oslin, J. L., Mitchell, S. A., & Griffin, L. L. (1998). The Game Performance Assessment Instrument (GPAI): Development and Preliminary Validation. *Journal of Teaching in Physical Education*, *17*(2), 231–243. https://doi.org/10.1123/jtpe.17.2.231

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

Park, H., & Kim, N. (2008). Predicting Factors of Physical Activity in Adolescents: A Systematic Review. *Asian Nursing Research*, 2(2), 113–128. https://doi.org/10.1016/S1976-1317(08)60035-3

Patterson, P., Bennington, J., & La-Rosa, T. D. (2001). Psychometric Properties of Child- and Teacher-Reported Curl-Up Scores in Children Ages 10–12 Years. *Research Quarterly for Exercise and Sport*, 72(2), 117–124. https://doi.org/10.1080/02701367.2001.10608941

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

Physical Literacy for Life Consortium. (2021). *Physical Literacy for Life Self-Assessment Tools*. https://physical-literacy.isca.org/tools/

Pierce, S. (2021). *piercer: Functions for Research and Statistical Computing* (R package version 0.9.1) [Computer software]. https://github.com/sjpierce/piercer

Plowman, S. A., & Meredith, M. D. (Eds.). (2013). *FITNESSGRAM /ACTIVITYGRAM Reference Guide (4th Edition)*. http://www.cooperinst.org/vault/2440/web/files/662.pdf

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

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459–467. https://doi.org/10.1002/nur.20199

Pons, J., Viladrich, C., Ramis, Y., & Polman, R. (2018). The Mediating Role of Coping between Competitive Anxiety and Sport Commitment in Adolescent Athletes. *The Spanish Journal of Psychology*, *21*. https://doi.org/10.1017/sjp.2018.8

Pot, N., Whitehead, M. E., & Durden-Myers, E. J. (2018). Physical Literacy From Philosophy to Practice. *Journal of Teaching in Physical Education*, *37*(3), 246–251. https://doi.org/10.1123/jtpe.2018-0133

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

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

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, 27(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

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

Rademaker, M., & Schuberth, F. (2020). *CSEM: Composite-Based Structural Equation Modeling* (Package version 0.4.0) [Computer software]. https://m-e-rademaker.github.io/cSEM/.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research.

Raykov, T., & Marcoulides, G. A. (2016). On the Relationship Between Classical Test Theory and Item Response Theory. *Educational and Psychological Measurement*, *76*(2), 325–338. https://doi.org/10.1177/0013164415576958

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer New York. https://doi.org/10.1007/978-0-387-89976-3

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and Modeling Psychological Measures in the Presence of Multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations: Exploring the Extent to which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., & Revicki, D. A. (Eds.). (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge, Taylor & Francis Group.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educational and Psychological Measurement*, *73*(1), 5–26. https://doi.org/10.1177/0013164412449831

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

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*(3), 287–297. https://doi.org/10.1037/1040-3590.12.3.287

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

Rigdon, E. E. (2012). Rethinking Partial Least Squares Path Modeling: In Praise of Simple Methods. *Long Range Planning*, *45*(5–6), 341–358. https://doi.org/10.1016/j.lrp.2012.09.010

Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal*, *34*(6), 598–605. https://doi.org/10.1016/j.emj.2016.05.006

Rigdon, E. E., Sarstedt, M., & Ringle, C. M. (2017). On Comparing Results from CB-SEM and PLS-SEM: Five Perspectives and Five Recommendations. *Marketing ZFP*, *39*(3), 4–16. https://doi.org/10.15358/0344-1369-2017-3-4

Rindskopf, D., & Rose, T. (1988). Some Theory and Applications of Confirmatory Second-Order Factor Analysis. *Multivariate Behavioral Research*, *23*(1), 51–67. https://doi.org/10.1207/s15327906mbr2301_3

Ringle, C. M., Wende, S., & Becker, J.-M. (2015). *SmartPLS 3*. SmartPLS. http://www.smartpls.com

Robinson, D. B., Randall, L., & Barrett, J. (2018). Physical Literacy (Mis)understandings: What do Leading Physical Education Teachers Know About Physical Literacy? *Journal of Teaching in Physical Education*, *37*(3), 288–298. https://doi.org/10.1123/jtpe.2018-0135

Robinson, L. E., Stodden, D. F., Barnett, L. M., Lopes, V. P., Logan, S. W., Rodrigues, L. P., & D'Hondt, E. (2015). Motor Competence and its Effect on Positive Developmental Trajectories of Health. *Sports Medicine (Auckland, N.Z.)*, *45*(9), 1273–1284. https://doi.org/10.1007/s40279-015-0351-6

Rodrigues, I. B., Adachi, J. D., Beattie, K. A., & MacDermid, J. C. (2017). Development and validation of a new tool to measure the facilitators, barriers and preferences to exercise in people with osteoporosis. *BMC Musculoskeletal Disorders*, *18*(1), 540. https://doi.org/10.1186/s12891-017-1914-5

Rodrigues, L. P., Luz, C., Cordovil, R., Bezerra, P., Silva, B., Camões, M., & Lima, R. (2019). Normative values of the motor competence assessment (MCA) from 3 to 23 years of age. *Journal of Science and Medicine in Sport*, *22*(9), 1038–1043. https://doi.org/10.1016/j.jsams.2019.05.009

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Roetert, E. P., & MacDonald, L. C. (2015). Unpacking the physical literacy concept for K-12 physical education: What should we expect the learner to master? *Journal of Sport and Health Science*, *4*(2), 108–112. https://doi.org/10.1016/j.jshs.2015.03.002

Roseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

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

Ryan, R. M., & Deci, E. L. (2002). Overview of self-determination theory: An organismic-dialectical perspective. In *Handbook of self-determination research* (pp. 3–33). University of Rochester Press.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97. https://doi.org/10.1007/BF03372160

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

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research*, *69*(10), 3998–4010. https://doi.org/10.1016/j.jbusres.2016.06.007

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Schoemaker, M. M., Niemeijer, A. S., Flapper, B. C. T., & Smits-Engelsman, B. C. M. (2012). Validity and reliability of the Movement Assessment Battery for Children-2 Checklist for children with and without motor impairments. *Developmental Medicine and Child Neurology*, *54*(4), 368–375. https://doi.org/10.1111/j.1469-8749.2012.04226.x

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Schuberth, F. (2020). Confirmatory composite analysis using partial least squares: Setting the record straight. *Review of Managerial Science*. https://doi.org/10.1007/s11846-020-00405-0

Schuberth, F., Henseler, J., & Dijkstra, T. K. (2018). Confirmatory Composite Analysis. *Frontiers in Psychology*, *9*, 2541. https://doi.org/10.3389/fpsyg.2018.02541

Schuberth, F., Rademaker, M. E., & Henseler, J. (2020). Estimating and assessing second-order constructs using PLS-PM: The case of composites of composites. *Industrial Management & Data Systems*, *120*(12), 2211–2241. https://doi.org/10.1108/IMDS-12-2019-0642

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

Scudder, M. R., Lambourne, K., Drollette, E. S., Herrmann, S. D., Washburn, R. A., Donnelly, J. E., & Hillman, C. H. (2014). Aerobic capacity and cognitive control in elementary school-age children. *Medicine and Science in Sports and Exercise*, *46*(5), 1025–1035. https://doi.org/10.1249/MSS.0000000000000199

Scully, D. (2017). *Constructing Multiple-Choice Items to Measure Higher-Order Thinking*. *22*(4), 14.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611.

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

Siedentop, D. (1998). What is Sport Education and How Does it Work? *Journal of Physical Education, Recreation & Dance*, *69*(4), 18–20. https://doi.org/10.1080/07303084.1998.10605528

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K., & Ark, L. A. van der. (2021). *Measurement models for psychological attributes*. CRC Press.

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

Sijtsma, K., & Molenaar, I. (2002). *Introduction to Nonparametric Item Response Theory*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984676

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

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

Simon, R. L., Torres, C. R., & Hager, P. F. (2015). *Fair Play: The Ethics of Sport* (4th Edition). Westview Press.

Singh, J. (2004). Tackling measurement problems with Item Response Theory. *Journal of Business Research*, *57*(2), 184–208. https://doi.org/10.1016/S0148-2963(01)00302-2

Slade, D. G. (2010). *Transforming play: Teaching tactics and game sense*. Human Kinetics.

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

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

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

Sonderen, E. van, Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLoS ONE*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

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

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

Standal, O. (2016). *Phenomenology and Pedagogy in Physical Education*. Routledge.

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

Stockwell, S., Trott, M., Tully, M., Shin, J., Barnett, Y., Butler, L., McDermott, D., Schuch, F., & Smith, L. (2021). Changes in physical activity and sedentary behaviours from before to during the COVID-19 pandemic lockdown: A systematic review. *BMJ Open Sport & Exercise Medicine*, 7(1), e000960. https://doi.org/10.1136/bmjsem-2020-000960

Stodden, D. F., Goodway, J. D., Langendorfer, S. J., Roberton, M. A., Rudisill, M. E., Garcia, C., & Garcia, L. E. (2008). A Developmental Perspective on the Role of Motor Skill Competence in Physical Activity: An Emergent Relationship. *Quest*, *60*(2), 290–306. https://doi.org/10.1080/00336297.2008.10483582

Stodden, D. F., Langendorfer, S., & Roberton, M. A. (2009). The Association Between Motor Skill Competence and Physical Fitness in Young Adults. *Research Quarterly for Exercise and Sport*, *80*(2), 223–229. https://doi.org/10.1080/02701367.2009.10599556

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

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

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

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

Stucky, B. D., & Edelen, M. O. (2015). Using Hierarchical IRT Models to Create Unidimensional Measures from Multidimensional Data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modelling: Applications to Typical Performance Assessment*. Routledge.

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

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

Sum, R. K. W., Cheng, C.-F., Wallhead, T., Kuo, C.-C., Wang, F.-J., & Choi, S.-M. (2018). Perceived physical literacy instrument for adolescents: A further validation of PPLI. *Journal of Exercise Science & Fitness*, *16*(1), 26–31. https://doi.org/10.1016/j.jesf.2018.03.002

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

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

Taylor, I. M., Ntoumanis, N., Standage, M., & Spray, C. M. (2010). Motivational predictors of physical education students' effort, exercise intentions, and leisure-time physical activity: A multilevel linear growth analysis. *Journal of Sport and Exercise Psychology*, *32*(1), 99–120.

Teixeira, P. J., Carraça, E. V., Markland, D., Silva, M. N., & Ryan, R. M. (2012). *Exercise, physical activity, and self-determination theory: A systematic review*. 30.

Telama, R. (2009). Tracking of physical activity from childhood to adulthood: A review. *Obesity Facts*, 2(3), 187–195. https://doi.org/10.1159/000222244

Telama, R., Yang, X., Leskinen, E., Kankaanpää, A., Hirvensalo, M., Tammelin, T., Viikari, J. S. A., & Raitakari, O. T. (2014). Tracking of physical activity from early childhood through youth into adulthood. *Medicine and Science in Sports and Exercise*, *46*(5), 955–962. https://doi.org/10.1249/MSS.0000000000000181

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.

The Cooper Institute. (2017). *FitnessGram Administration Manual: The Journey to MyHealthyZone* (Fifth edition). Human Kinetics.

Tidén, A., Lundqvist, C., & Nyberg, M. (2015). Development and Initial Validation of the NyTid Test: A Movement Assessment Tool for Compulsory School Pupils. *Measurement in Physical Education and Exercise Science*, *19*(1), 34–43. https://doi.org/10.1080/1091367X.2014.975228

Tierney, N., Cook, D., McBain, M., & Fay, C. (2021). *naniar: Data Structures, Summaries, and Visualisations for Missing Data* (R package version 0.6.1) [Computer software]. https://CRAN.R-project.org/package=naniar

Tinning, R. (2015). 'I don't read fiction': Academic discourse and the relationship between health and physical education. *Sport, Education and Society*, *20*(6), 710–721. https://doi.org/10.1080/13573322.2013.798638

Toland, M. D. (2014). Practical Guide to Conducting an Item Response Theory Analysis. *The Journal of Early Adolescence*, *34*(1), 120–151. https://doi.org/10.1177/0272431613511332

Towns, M. H. (2014). Guide To Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *Journal of Chemical Education*, *91*(9), 1426–1431. https://doi.org/10.1021/ed500076x

Tremblay, M., Costas-Bradstreet, C., Barnes, J. D., Bartlett, B., Dampier, D., Lalonde, C., Leidl, R., Longmuir, P., McKee, M., Patton, R., Way, R., & Yessis, J. (2018). Canada's Physical Literacy Consensus Statement: Process and outcome. *BMC Public Health*, *18*(2), 1034. https://doi.org/10.1186/s12889-018-5903-x

Turiel, E. (2015). Morality and Prosocial Judgments and Behavior. In D. A. Schroeder & W. G. Graziano (Eds.), *The Oxford Handbook of Prosocial Behavior*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399813.013.022

Ubago-Jiménez, J. L., González-Valero, G., Puertas-Molero, P., & García-Martínez, I. (2019). Development of Emotional Intelligence through Physical Activity and Sport Practice. A Systematic Review. *Behavioral Sciences*, *9*(4). https://doi.org/10.3390/bs9040044

Uddin, R., Salmon, J., Islam, S. M. S., & Khan, A. (2020). Physical education class participation is associated with physical activity among adolescents in 65 countries. *Scientific Reports*, *10*(1), 22128. https://doi.org/10.1038/s41598-020-79100-9

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

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

Vallerand, R. J., Rousseau, F. L., Grouzet, F. M. E., Dumais, A., Grenier, S., & Blanchard, C. M. (2006). Passion in Sport: A Look at Determinants and Affective Experiences. *Journal of Sport and Exercise Psychology*, *28*(4), 454–478. https://doi.org/10.1123/jsep.28.4.454

van Riel, A. C. R., Henseler, J., Kemény, I., & Sasovova, Z. (2017). Estimating hierarchical constructs using consistent partial least squares: The case of second-order composites of common factors. *Industrial Management & Data Systems*, *117*(3), 459–477. https://doi.org/10.1108/IMDS-07-2016-0286

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

Vanhelst, J., Béghin, L., Fardy, P. S., Ulmer, Z., & Czaplicki, G. (2016). Reliability of health-related physical fitness tests in adolescents: The MOVE Program. *Clinical Physiology and Functional Imaging*, *36*(2), 106–111. https://doi.org/10.1111/cpf.12202

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

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

Wallhead, T. L., Garn, A. C., & Vidoni, C. (2013). Sport Education and social goals in physical education: Relationships with enjoyment, relatedness, and leisure-time physical activity. *Physical Education and Sport Pedagogy*, *18*(4), 427–441. https://doi.org/10.1080/17408989.2012.690377

Waltz, C., Strickland, O., & Lenz, E. (2010). *Measurement in Nursing and Health Research* (4th ed.). Springer. https://en.book4you.org/book/969788/f90163

Ward, J. T., Nobles, M. R., & Fox, K. A. (2015). Disentangling Self-Control from Its Elements: A Bifactor Analysis. *Journal of Quantitative Criminology*, *31*(4), 595–627. https://doi.org/10.1007/s10940-014-9241-6

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

Werner, P., Thorpe, R., & Bunker, D. (1996). Teaching Games for Understanding: Evolution of a Model. *Journal of Physical Education, Recreation & Dance*, *67*(1), 28–33. https://doi.org/10.1080/07303084.1996.10607176

Whitehead, M. (2001). The Concept of Physical Literacy. *European Journal of Physical Education*, *6*(2), 127–138. https://doi.org/10.1080/1740898010060205

Whitehead, M. (2007). Physical Literacy: Philosophical Considerations in Relation to Developing a Sense of Self, Universality and Propositional Knowledge. *Sport, Ethics and Philosophy*, *1*(3), 281–298. https://doi.org/10.1080/17511320701676916

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

Whitehead, M. (2013a). Definition of physical literacy and clarification of related issues. *ICSSPE Journal of Sport Science and Physical Education*, *65*, 29–34.

Whitehead, M. (2013b). The history and development of physical literacy. *ICSSPE Journal of Sport Science and Physical Education*, *65*, 22–28.

Wiggins, G. (1990). The Case for Authentic Assessment. *Practical Assessment, Research, and Evaluation*, 2(2). https://doi.org/10.7275/FFB1-MM19

Wilcox, J. B., Howell, R. D., & Breivik, E. (2008). Questions about formative measurement. *Journal of Business Research*, *61*(12), 1219–1228. https://doi.org/10.1016/j.jbusres.2008.01.010

Willis, G. B., & Artino, A. R. (2013). What Do Our Respondents Think We're Asking? Using Cognitive Interviewing to Improve Medical Education Surveys. *Journal of Graduate Medical Education*, *5*(3), 353–356. https://doi.org/10.4300/JGME-D-13-00154.1

Willse, J. (2018). *Classical Test Theory Functions (CTT)* (2.3.3) [R]. https://cran.r-project.org/web/packages/CTT/CTT.pdf

Wilson, P. M., Rogers, W. T., Rodgers, W. M., & Wild, T. C. (2006). The Psychological Need Satisfaction in Exercise Scale. *Journal of Sport and Exercise Psychology*, *28*(3), 231–251. https://doi.org/10.1123/jsep.28.3.231

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

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

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety.

Educational and Psychological Measurement, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wong, C.-S., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. *The Leadership Quarterly*, *13*(3), 243–274. https://doi.org/10.1016/S1048-9843(02)00099-1

Woods, C., Moyna, N., & Quinlan, A. (2010). *The children's sport participation and physical activity study (CSPPA study)*. http://doras.dcu.ie/21335/

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

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

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

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two Quantitative Approaches for Estimating Content Validity. *Western Journal of Nursing Research*, *25*(5), 508–518. https://doi.org/10.1177/0193945903252998

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

Ydo, Y. (2021). Physical literacy on the global agenda. *PROSPECTS*, *50*(1), 1–3. https://doi.org/10.1007/s11125-020-09524-8

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

Young, L., O'Connor, J., & Alfrey, L. (2019). Physical literacy: A concept analysis. *Sport, Education and Society*, 25(8), 946–959. https://doi.org/10.1080/13573322.2019.1677586

Young, L., O'Connor, J., Alfrey, L., & Penney, D. (2021). Assessing physical literacy in health and physical education. *Curriculum Studies in Health and Physical Education*, *12*(2), 156–179. https://doi.org/10.1080/25742981.2020.1810582

Yuan, K.-H., & Bentler, P. M. (2000). Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128. https://doi.org/10.1007/BF02294531

Zamanzadeh, V. (2015). *Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication*. 14.

Zeidner, M., Matthews, G., & Roberts, R. D. (Eds.). (2012). *What we know about emotional intelligence: How it affects learning, work, relationships, and our mental health* (1. MIT Press paperback ed). MIT Press.

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

Additional Files

Additional File 1

Supplementary Table S1. Information about experts that participated in the content validation of the PPLA-Q

Expert	PPLA-Q Domain	Expertise	#Citations [1]	h-index[1]	Round participation
1	Cognitive	Test development, educational assessment	201	NA	1st
2	Cognitive	Health promotion and public Health	3468	28	1st
3	Cognitive	Curriculum development, PE didactics	80	NA	1st
4	Cognitive	Curriculum development, PE didactics, developer of the PPES	NA	NA	1st
5	Psychological	Sport psychology, decision-making in sport	16699	72	1st
6	Psychological	Scale development and validation, sport psychology	1052	17	1st*
7	Psychological	Sport psychology, behavioral change	3025	25	1st
8	Social	Sport sociology, school ethnography	59	5	1st & 2nd
9	Social and Psychological	Sport psychology and pedagogy, scale validation	5287	35	1st
10	All	Healthy and active lifestyles, PE Didactics	1162	18	1st & 2nd
11	All	PE didactics, Teacher education	3014	27	1st
12	All	PE didactics, Teacher education	1167	16	2nd

Note. All experts were professors at Graduate-level Education. NA – Not available; PE – Physical Education; PPES – Portuguese Physical Education syllabus.
*Qualitative evaluation only.
[1]Citation data obtained from each expert's Google Scholar profile in January 2021.

Additional File 2

Supplementary Table S2. Item CVI, kappa coefficient and evaluation of each item (PPLA-Q version 0.2)

Item	Relevance					Clarity				
	Number of Experts	Experts Giving Rating of 3 or 4	I-CVI	κ[1]	Evaluation[2]	Number of Experts	Experts Giving Rating of 3 or 4	Proportion of Agreement	κ[1]	Evaluation[3]
Cognitive Module										
C1	6	6	1	1	Excellent	6	6	1	1	Clear
C2	6	2	.33	.13	Eliminate	6	3	.50	.27	Review
C3	5	5	1	1	Excellent	6	6	1	1	Clear
C4	5	5	1	1	Excellent	6	5	.83	.82	Clear
C5	5	5	1	1	Excellent	6	6	1	1	Clear
C6	5	4	.80	.76	Excellent	6	4	.67	.56	Review
C7	5	5	1	1	Excellent	6	5	.83	.82	Clear
C8	5	4	.80	.76	Excellent	6	2	.33	.13	Review
C9	5	5	1	1	Excellent	6	5	.83	.82	Clear
C10	5	4	.80	.76	Excellent	6	6	1	1	Clear
S-CVI/Ave			.87							
S-CVI/UA			.60							
Psychological Module										
P1	5	5	1	1	Excellent	5	5	1	1	Clear
P2	5	5	1	1	Excellent	5	5	1	1	Clear
P3	5	5	1	1	Excellent	5	5	1	1	Clear
P4	5	5	1	1	Excellent	5	5	1	1	Clear
P5	5	5	1	1	Excellent	5	5	1	1	Clear
P6	5	5	1	1	Excellent	5	5	1	1	Clear
P7	5	5	1	1	Excellent	5	5	1	1	Clear
P8	5	5	1	1	Excellent	5	4	.80	.76	Clear
P9	5	5	1	1	Excellent	5	4	.80	.76	Clear
P10	5	5	1	1	Excellent	5	5	1	1	Clear
P11	5	5	1	1	Excellent	5	5	1	1	Clear
P12	5	5	1	1	Excellent	5	5	1	1	Clear
P13	5	5	1	1	Excellent	5	5	1	1	Clear
P14	5	5	1	1	Excellent	5	5	1	1	Clear
P15	5	5	1	1	Excellent	5	5	1	1	Clear
P16	5	4	.80	.76	Excellent	5	4	.80	.76	Clear
P17	5	5	1	1	Excellent	5	5	1	1	Clear
P18	5	5	1	1	Excellent	5	5	1	1	Clear
P19	5	5	1	1	Excellent	5	5	1	1	Clear
P20	5	5	1	1	Excellent	5	5	1	1	Clear
P21	5	5	1	1	Excellent	5	5	1	1	Clear
P22	5	5	1	1	Excellent	5	5	1	1	Clear
P23	5	5	1	1	Excellent	5	5	1	1	Clear
P24	5	4	.80	.76	Excellent	5	5	1	1	Clear
P25	5	5	1	1	Excellent	5	4	.80	.76	Clear
P26	5	5	1	1	Excellent	5	5	1	1	Clear
P27	5	5	1	1	Excellent	5	5	1	1	Clear
P28	5	5	1	1	Excellent	5	5	1	1	Clear
P29	5	5	1	1	Excellent	5	5	1	1	Clear
P30	5	5	1	1	Excellent	5	5	1	1	Clear
P31	5	5	1	1	Excellent	5	5	1	1	Clear
P32	5	5	1	1	Excellent	5	5	1	1	Clear
P33	5	5	1	1	Excellent	5	5	1	1	Clear
P34	5	5	1	1	Excellent	5	5	1	1	Clear
P35	4	4	1	1	Excellent	4	4	1	1	Clear
P36	5	4	.80	.76	Excellent	5	5	1	1	Clear
P37	4	4	1	1	Excellent	4	3	.75	.67	Review
P38	4	4	1	1	Excellent	3	3	1	1	Clear
P39	4	4	1	1	Excellent	3	3	1	1	Clear
P40	4	4	1	1	Excellent	4	2	.50	.20	Review
S-CVI/Ave			.98							
S-CVI/UA			.93							
Social Module										
S1	4	3	.75	.67	Good	3	1	.33	-.07	Review
S2	4	3	.75	.67	Good	4	4	1	1	Clear
S3	4	2	.50	.20	Eliminate	4	3	.75	.67	Review
S4	4	4	1	1	Excellent	4	3	.75	.67	Review
S5	4	4	1	1	Excellent	4	4	1	1	Clear

Supplementary Table S2. Item CVI, kappa coefficient and evaluation of each item (PPLA-Q version 0.2)

Item	Relevance					Clarity				
	Number of Experts	Experts Giving Rating of 3 or 4	I-CVI	κ[1]	Evaluation[2]	Number of Experts	Experts Giving Rating of 3 or 4	Proportion of Agreement	κ[1]	Evaluation[3]
S6	4	3	.75	.67	Good	3	3	1	1	Clear
S7	4	4	1	1	Excellent	3	3	1	1	Clear
S8	4	3	.75	.67	Good	3	3	1	1	Clear
S9	4	4	1	1	Excellent	3	3	1	1	Clear
S10	4	4	1	1	Excellent	4	4	1	1	Clear
S11	4	4	1	1	Excellent	4	4	1	1	Clear
S12	4	4	1	1	Excellent	4	4	1	1	Clear
S13	4	4	1	1	Excellent	4	4	1	1	Clear
S14	4	4	1	1	Excellent	3	3	1	1	Clear
S15	4	4	1	1	Excellent	4	4	1	1	Clear
S16	4	4	1	1	Excellent	4	4	1	1	Clear
S17	4	4	1	1	Excellent	4	4	1	1	Clear
S18	4	4	1	1	Excellent	4	4	1	1	Clear
S19	4	4	1	1	Excellent	4	4	1	1	Clear
S20	4	3	.75	.67	Good	4	4	1	1	Clear
S21	4	3	.75	.67	Good	4	3	.75	.67	Review
S22	4	3	.75	.67	Good	4	3	.75	.67	Review
S23	4	3	.75	.67	Good	4	4	1	1	Clear
S24	3	3	1	1	Excellent	4	4	1	1	Clear
S25	3	3	1	1	Excellent	4	4	1	1	Clear
S26	3	3	1	1	Excellent	4	4	1	1	Clear
S27	3	3	1	1	Excellent	4	4	1	1	Clear
S28	4	4	1	1	Excellent	4	4	1	1	Clear
S29	4	3	.75	.67	Good	4	3	.75	.67	Review
S30	4	2	.50	.20	Eliminate	4	4	1	1	Clear
S31	4	3	.75	.67	Good	4	3	.75	.67	Review
S32	4	4	1	1	Excellent	4	4	1	1	Clear
S33	4	4	1	1	Excellent	4	4	1	1	Clear
S34	4	2	.50	.20	Eliminate	4	3	.75	.67	Review
S35	4	4	1	1	Excellent	4	4	1	1	Clear
S36	4	4	1	1	Excellent	4	4	1	1	Clear
S37	4	4	1	1	Excellent	3	2	.67	.47	Review
S38	4	4	1	1	Excellent	3	3	1	1	Clear
S39	4	4	1	1	Excellent	3	3	1	1	Clear
S40	4	4	1	1	Excellent	3	3	1	1	Clear
S-CVI/Ave			.90							
S-CVI/UA			.68							

I-CVI- Item Content Validity Index; Ave- Average; UA- Universal Agreement; κ – kappa coefficient; S-CVI- Scale Content Validity Index.

[1] Multirater modified kappa designating agreement on relevance: κ= (I-CVI - pc)/(1 -pc), with pc (probability of a chance occurrence) computed using the formula for a binomial random variable, with one specific outcome described in Polit et al. (2007)

[2] Evaluation criteria for kappa, using guidelines described in Cicchetti and Sparrow (1981) and Fleiss (1981): Fair kappa of .40 to .59; Good kappa .60 to .74; and Excellent kappa > .74.

[3] Modified criteria for kappa: Needs Revision < .74; Clear > .74

Supplementary Table S3. Social Module's Item CVI, kappa coefficient and evaluation (PPLA-Q version 0.3)

	Relevance					Clarity				
Item	Number of Experts	Number Giving Rating of 3 or 4	I-CVI	κ[1]	Evaluation[2]	Number of Experts	Experts Giving Rating of 3 or 4	Proportion of Agreement	κ[1]	Evaluation[3]
S1*	3	3	1	1	Excellent	3	3	1	1	Clear
S2*	3	3	1	1	Excellent	3	3	1	1	Clear
S3*	3	3	1	1	Excellent	3	2	.67	.47	Review
S4*	3	3	1	1	Excellent	3	2	.67	.47	Review
S5*	3	3	1	1	Excellent	3	2	.67	.47	Review
S6*	3	3	1	1	Excellent	3	3	1	1	Clear
S7*	3	3	1	1	Excellent	3	2	.67	.47	Review
S8*	3	3	1	1	Excellent	3	3	1	1	Clear
S9*	3	3	1	1	Excellent	3	1	.33	-.07	Review
S10	4	4	1	1	Excellent	4	4	1	1	Clear
S11	4	4	1	1	Excellent	4	4	1	1	Clear
S12	4	4	1	1	Excellent	4	4	1	1	Clear
S13	4	4	1	1	Excellent	4	4	1	1	Clear
S14	4	4	1	1	Excellent	3	3	1	1	Clear
S15	4	4	1	1	Excellent	4	4	1	1	Clear
S16	4	4	1	1	Excellent	4	4	1	1	Clear
S17	4	4	1	1	Excellent	4	4	1	1	Clear
S18	4	4	1	1	Excellent	4	4	1	1	Clear
S19	4	4	1	1	Excellent	4	4	1	1	Clear
S20	4	3	.75	.67	Good	4	4	1	1	Clear
S21	4	3	.75	.67	Good	4	3	.75	.67	Review
S22	4	3	.75	.67	Good	4	3	.75	.67	Review
S23	4	3	.75	.67	Good	4	4	1	1	Clear
S24	3	3	1	1	Excellent	4	4	1	1	Clear
S25	3	3	1	1	Excellent	4	4	1	1	Clear
S26	3	3	1	1	Excellent	4	4	1	1	Clear
S27	3	3	1	1	Excellent	4	4	1	1	Clear
S28	4	4	1	1	Excellent	4	4	1	1	Clear
S29	4	3	.75	.67	Good	4	3	.75	.67	Review
S30	4	3	.75	.67	Good	4	3	.75	.67	Review
S31	4	4	1	1	Excellent	4	4	1	1	Clear
S32	4	4	1	1	Excellent	4	4	1	1	Clear
S33	4	4	1	1	Excellent	4	4	1	1	Clear
S34	4	4	1	1	Excellent	4	4	1	1	Clear
S35*	3	3	1	1	Excellent	3	3	1	1	Clear
S36	4	4	1	1	Excellent	3	3	1	1	Clear
S37	4	4	1	1	Excellent	3	3	1	1	Clear
S38	4	4	1	1	Excellent	3	3	1	1	Clear
S-CVI/Ave			.96							
S-CVI/UA			.84							

I-CVI- Item Content Validity Index; Ave- Average; UA- Universal Agreement; κ – kappa coefficient; S-CVI- Scale Content Validity Index.
*Items included in 2nd round of expert validation.

[1] Multirater modified kappa designating agreement on relevance: κ=(I-CVI - pc)/(1 -pc), with pc (probability of a chance occurrence) computed using the formula for a binomial random variable, with one specific outcome described in Polit et al. (2007)

[2] Evaluation criteria for kappa, using guidelines described in Cicchetti and Sparrow (1981) and Fleiss (1981): Fair kappa of .40 to .59; Good kappa .60 to .74; and Excellent kappa > .74

[3] Modified criteria for kappa: Needs Revision < .74; Clear > .74

Additional File 3

English version

For the next question group, you should remember that PA = Physical Activity(ies), and includes every situation that require movement, like your Physical Education classes, sport-based activities (team or individual), rhythmic activities (e.g., dance), exercise (e.g., strength training, jogging), and any activities that you use as a mean of transportation or in your spare time.

In each statement below, select how much it describes you, circling one of the options from 0 to 4 0 = ***Not at all***; 1= ***Slightly***; 2 =***Moderately***; 3 =***Quite a lot***; 4 = ***Totally)***.

How much do these statements describe you?	Not at all	Slightly	Moderately	Quite a lot	Totally
P1. I am motivated to practice PA	0	1	2	3	4
P2. I practice PA because others tell me I should	0	1	2	3	4
P3. I feel guilty when I do not practice PA	0	1	2	3	4
P4. I feel bad about myself when I do not practice PA	0	1	2	3	4
P5. I feel pressured by others to practice PA	0	1	2	3	4
P6. I practice PA because I feel others would be unhappy if I did not	0	1	2	3	4
P7. I practice PA because it is fun	0	1	2	3	4
P8. I feel good when I practice PA	0	1	2	3	4
P9. I consider PA a part of me	0	1	2	3	4
P10. I value the benefits of PA	0	1	2	3	4
P11. I see PA as a fundamental part of who I am	0	1	2	3	4
P12. I enjoy practicing PA	0	1	2	3	4
P13. I feel confident to practice PA	0	1	2	3	4
In Physical Activity contexts:					
P14. I am confident in my abilities	0	1	2	3	4
P15. I can participate with success	0	1	2	3	4
P16. I consider myself competent	0	1	2	3	4
P17. I trust my skills	0	1	2	3	4
P18. I feel good about the way I am able to participate	0	1	2	3	4
P19. I can participate in PA that I consider challenging	0	1	2	3	4
P20. I know how to become more confident in myself	0	1	2	3	4
P21. I feel competent even when I am criticized	0	1	2	3	4

P22. I believe in myself even when I lose	0	1	2	3	4
P23. I can manage my emotions	0	1	2	3	4
P24. I can recognize other's emotions	0	1	2	3	4
P25. I can recognize my emotions	0	1	2	3	4
In Physical Activity contexts:					
P26. I am sensitive to the feelings of others	0	1	2	3	4
P27. I understand what others feel	0	1	2	3	4
P28. I can identify what I feel	0	1	2	3	4
P29. I can anticipate what I will feel	0	1	2	3	4
P30. I can deal with difficulties rationally	0	1	2	3	4
P31. I can manage my emotions when necessary	0	1	2	3	4
P32. I have a good control of my emotions	0	1	2	3	4
P33. I can manage my effort	0	1	2	3	4
P34. I know when I am tired	0	1	2	3	4
P35. I can recognize changes in my breathing	0	1	2	3	4
P36. I can recognize changes in my heart rate	0	1	2	3	4
P37. I recognize my physical limits	0	1	2	3	4
P38. I can recognize the effect that different intensities have in me	0	1	2	3	4
P39. I use strategies to manage my effort	0	1	2	3	4
P40. I can anticipate when I will be fatigued	0	1	2	3	4
P41. I can control my fatigue	0	1	2	3	4
P42. I take action to improve my physical skills	0	1	2	3	4
In the different contexts of my life:					
P43. I feel more motivated to reach my goals because I practice PA	0	1	2	3	4
P44. I feel more confident in my skills because I practice PA	0	1	2	3	4
P45. I am better at controlling my emotions because I practice PA	0	1	2	3	4
P46. I am better at controlling my fatigue because I practice PA	0	1	2	3	4

How much do these statements describe you?	Not al all	Slightly	Moderately	Quite a lot	Totally
S1. I believe that the cultural aspects of PA are important (e.g., its rituals, terminology, clothing, values)	0	1	2	3	4
S2. I participate in PA rituals (e.g., greetings, hymns/chants, cheers, applauses)	0	1	2	3	4
S3. I use specific PA terminology (e.g., names of technics and tactics, names of equipment, idioms)	0	1	2	3	4
S4. I use specific clothing of the PA I am practicing	0	1	2	3	4
S5. I watch PA events (e.g., competitions, spectacles, shows)	0	1	2	3	4
S6. I like to keep up with PA events (e.g., competitions, spectacles, shows)]	0	1	2	3	4
S7. I am interested in the cultural aspects of PA (e.g., its rituals, terminology, clothing, values)]	0	1	2	3	4
S8. I encourage others to watch PA events (e.g., competitions, spectacles, shows)	0	1	2	3	4
S9. I encourage others to participate in each PA's culture (e.g., rituals, terminology, clothing)]	0	1	2	3	4
S10. I can lead a healthy and active life	0	1	2	3	4
In Physical Activity contexts:					
S11. I work well with others	0	1	2	3	4
S12. I try to behave correctly and justly	0	1	2	3	4
S13. I respect my adversaries	0	1	2	3	4
S14. I follow the rules	0	1	2	3	4
S15.I cheat if it brings me benefits	0	1	2	3	4
S16. I respect the decisions of authorities (e.g., referee, umpire, coach/teacher)]	0	1	2	3	4
S17. I behave according to fair-play / sport ethics 'principles	0	1	2	3	4
S18. I understand the importance of fair play/ sport ethics' principles	0	1	2	3	4
S19. I take action to make others behave according to fair play/sport ethic	0	1	2	3	4
S20. I follow the rules, even if unsupervised	0	1	2	3	4
S21. I behave according to fair play/sport ethics' principles on my initiative	0	1	2	3	4
S22. I take action for others to follow the rules	0	1	2	3	4
S23. I collaborate with others	0	1	2	3	4

S24. I am sympathetic with others	0	1	2	3	4
S25. I control my behavior towards others	0	1	2	3	4
S26. I respect others	0	1	2	3	4
S27. I cooperate with others	0	1	2	3	4
S28. I encourage others	0	1	2	3	4
S29. I care about others' success	0	1	2	3	4
S30. I help others achieve success	0	1	2	3	4
S31. I am helpful with others	0	1	2	3	4
S32. I have a positive relationship with others	0	1	2	3	4
S33. I interact with others	0	1	2	3	4
S34. I share a common goal with others	0	1	2	3	4
S35. I feel close to others	0	1	2	3	4
S36. I feel a sense of camaraderie with others	0	1	2	3	4
S37. I take action to improve my relationship with others	0	1	2	3	4
S38. I know how to improve my relationship with others	0	1	2	3	4
S39. I care about my relationship with others	0	1	2	3	4
In the different contexts of my life:					
S40. I am more involved in other cultural activities (e.g., theater, music) because I practice PA	0	1	2	3	4
S41. I am more honest and just because I practice PA	0	1	2	3	4
S42. I collaborate more with others because I practice PA	0	1	2	3	4
S43. I have better relationships with others because I practice PA	0	1	2	3	4

This question group will be about your opinion about topics of physical activity and healthy lifestyles.

Read each question attentively and circle your option(s).

C1. Select the option which represents a healthy and balanced meal.

(A) Chicken soup, cheeseburger, apple, soda.
(B) Vegetable soup, grilled salmon with sweet potatoes and baked vegetables, pear, water.
(C) Garlic bread, carbonara pasta with salad, pear, water.
(D) Bean soup, fried shrimp patties with tomato rice, apple, soda.

C2. In a normal day, Maria total calorie intake is 2000 kcal, and her total caloric expenditure is 2200 kcal.

Select the option which describes Maria's energetical balance for that day and its effects on her bodyweight.

(A) She will have a negative energetic balance and will lose weight.
(B) She will have a positive energetic balance and will gain weight.
(C) She will have a positive energetic balance and will lose weight.
(D) She will have a negative energetic balance and will lose weight.

C3. The "capacity to capture and use oxygen to produce energy" and the "capacity to move one or more joints through their complete range of motion" describe, respectively:

(A) muscular strength, and flexibility.
(B) aerobic fitness, and flexibility.
(C) aerobic fitness, and muscular strength.
(D) speed, and muscular strength.

C4. João has low strength and flexibility and wants to improve those capacities.

Select the option which describes adequate training methods for João.

	Training	Frequency	Volume per muscle group/joint
(A)	**Strength**	**2- 3 non consecutive sessions per week**	**1 to 3 sets of 12 repetitions**
	Flexibility	**5-7 sessions per week**	**30 seconds per exercise**
(B)	Strength	1 session per week	1 to 3 sets of 12 repetitions
	Flexibility	5-7 sessions per week	30 seconds per exercise
(C)	Strength	2- 3 consecutive sessions per week	1 to 3 sets of 12 repetitions
	Flexibility	5-7 sessions per week	30 seconds per exercise
(D)	Strength	2- 3 non consecutive sessions per week	1 to 3 sets of 5 repetitions
	Flexibility	5-7 sessions per week	30 seconds per exercise

C5. Select **all** the options which describe basic actions to avoid injuries during physical activity.

(A) Perform a warmup in the beginning of the activity, and a cool down at the end of the activity.
(B) In the event of an injury, apply something hot immediately.
(C) Use adequate protective equipment.
(D) Have a meal 10 minutes before an activity.
(E) Follow the safety rules of the activity.
(F) Drink water regularly during the activity.

C6. Fill in the text by selecting in the following table the adequate option for each space.

A competitive athlete is thinking about using anabolic steroids. This practice is 1. since it 2. fair-play, 3. sport performance, and 4. health.

1.	2.	3.	4.
(A) safe **(B) forbidden** **(C)** permitted	**(A)** increases **(B) decreases** **(C)** has no effect on	**(A) increases** **(B)** decreases **(C)** has no effect on	**(A) increases the risk of negative consequences for** **(B)** decreases the risk of negative consequences for **(C)** has no effect on

C7. The World Health Organization has defined recommendations for Physical Activity.

Select the option which describes the general guidelines for children and adolescents, and for adults.

(A) Children and adolescents: 60 minutes per day with moderate to vigorous intensity; adults: 150 minutes per week with moderate to vigorous intensity.
(B) Children and adolescents: 150 minutes per week with moderate to vigorous intensity; **adults**: 150 minutes per week with moderate to vigorous intensity
(C) Children and adolescents: 30 minutes per day with moderate to vigorous intensity; **adults**: 150 minutes per week with moderate to vigorous intensity.
(D) Children and adolescents: 60 minutes per day with moderate to vigorous intensity; **adults**: 90 minutes per week with moderate to vigorous intensity.

C8. Select the option that correctly links each type of physical training with its impact on health.

Types of Fitness Training
1.Strength training
2.Flexibility training
3. Aerobic fitness training

Impacts on health
A. Better blood pressure and cardiovascular health
B. Increased resting heart rate
C. Increased mobility and ability to perform day-to-day activities
D. Increased coordination difficulties
E. Better bone, tendon, and ligament health, and reduced injury risk

(A) 1-D, 2-C, 3-A
(B) 1-A, 2-B, 3-C
(C) 1-E, 2-B, 3-D
(D) 1-E, 2-C, 3-A

C9. The calculation formula for the Body Mass Index (BMI) is

(A) $\frac{\boldsymbol{Weight\ (kg)}}{\boldsymbol{Height\ (m)^2}}$

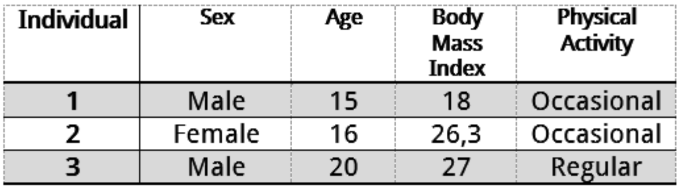
(B) $\frac{Height\ (cm)^2}{Weight\ (kg)}$

(C) $\frac{Weight\ (kg)}{Height\ (cm)^2}$

(D) $\frac{Height\ (m)}{Weight\ (kg)^2}$

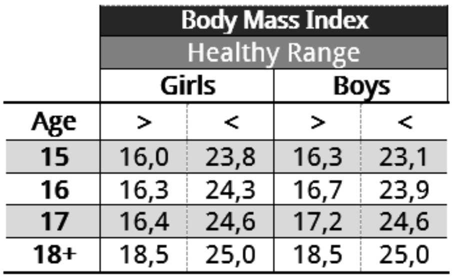
C10. Based on the analysis of the two tables provided, select all true statements.

Table 1



Individual	Sex	Age	Body Mass Index	Physical Activity
1	Male	15	18	Occasional
2	Female	16	26,3	Occasional
3	Male	20	27	Regular

Table 2. Reference values for BMI (FITschool)



	Body Mass Index			
	Healthy Range			
	Girls		Boys	
Age	>	<	>	<
15	16,0	23,8	16,3	23,1
16	16,3	24,3	16,7	23,9
17	16,4	24,6	17,2	24,6
18+	18,5	25,0	18,5	25,0

(A) Individual 1 should practice physical activity regularly.

(B) Individual 1 is below the healthy range of BMI.

(C) Individual 2 should reduce her weight.

(D) Individual 3 should maintain his weight.

(E) Individual 2 has higher risk for disease than individual 3.

Check that you answered all questions.

Thank you for your invaluable participation!

Versão Portuguesa

Para responderes ao próximo grupo, deves relembrar que **AF = Atividade(s) Física(s)**, e que inclui todas as situações que impliquem movimento, como por exemplo, as aulas de Educação Física, as atividades desportivas (coletivas ou individuais), as atividades rítmicas (ex. dança), o exercício físico (ex. treino de força, *jogging*), e as atividades físicas que faças para te deslocar ou nos teus tempos livres.

Para cada afirmação abaixo deves selecionar o quanto essa te descreve circundando uma das opções de 0 a 4 (**0= Nada, 1= Pouco, 2= Moderadamente, 3= Bastante, 4= Totalmente**).

Como é que estas frases te descrevem?	Nada	Pouco	Moderadamente	Bastante	Totalmente
P1. Sinto-me motivado(a) para praticar AF	0	1	2	3	4
P2. Faço AF porque outras pessoas dizem que devo fazer	0	1	2	3	4
P3. Sinto-me culpado(a) quando não faço AF	0	1	2	3	4
P4. Sinto-me mal comigo mesmo(a) quando não faço AF	0	1	2	3	4
P5. Sinto-me pressionado(a) pelos outros para fazer AF	0	1	2	3	4
P6. Faço AF porque outras pessoas vão ficar insatisfeitas comigo se não fizer	0	1	2	3	4
P7. Faço AF porque é divertido	0	1	2	3	4
P8. Sinto-me bem quando faço AF	0	1	2	3	4
P9. Considero que a AF faz parte de mim	0	1	2	3	4
P10. Valorizo os benefícios de praticar AF	0	1	2	3	4
P11. Vejo a AF como parte fundamental de quem sou	0	1	2	3	4
P12. Gosto de praticar AF	0	1	2	3	4
P13. Sinto-me confiante para praticar AF	0	1	2	3	4
Em situações de Atividade Física:					
P14. Sinto-me confiante nas minhas habilidades	0	1	2	3	4
P15. Consigo participar com sucesso	0	1	2	3	4
P16. Considero-me bom/boa praticante	0	1	2	3	4
P17. Confio nas minhas capacidades	0	1	2	3	4
P18. Sinto-me bem com a forma como consigo participar	0	1	2	3	4
P19. Consigo participar em AF que considero desafiantes	0	1	2	3	4
P20. Sei como fazer para me tornar mais confiante	0	1	2	3	4
P21. Sinto-me capaz mesmo quando sou criticado(a)	0	1	2	3	4
P22. Acredito na minha capacidade mesmo quando falho	0	1	2	3	4
P23. Consigo gerir as minhas emoções	0	1	2	3	4

P24. Reconheço as emoções dos outros	0	1	2	3	4
P25. Consigo reconhecer as minhas emoções	0	1	2	3	4
Em situações de Atividade Física:					
P26. Sou um(a) bom(a) observador(a) das emoções dos outros	0	1	2	3	4
P27. Percebo o que os outros sentem	0	1	2	3	4
P28. Consigo identificar o que sinto	0	1	2	3	4
P29. Consigo antecipar as minhas emoções	0	1	2	3	4
P30. Consigo lidar com as dificuldades racionalmente	0	1	2	3	4
P31. Consigo gerir as minhas emoções quando necessário	0	1	2	3	4
P32. Tenho um bom controlo das minhas emoções	0	1	2	3	4
P33. Consigo gerir o meu esforço	0	1	2	3	4
P34. Sei quando me sinto cansado(a)	0	1	2	3	4
P35. Consigo aperceber-me de mudanças na minha respiração	0	1	2	3	4
P36. Consigo reconhecer mudanças no meu ritmo cardíaco	0	1	2	3	4
P37. Tenho noção dos meus limites físicos	0	1	2	3	4
P38. Reconheço o efeito que diferentes intensidades de esforço têm em mim	0	1	2	3	4
P39. Utilizo estratégias para gerir o meu esforço	0	1	2	3	4
P40. Consigo antecipar quando irei ficar cansado(a)	0	1	2	3	4
P41. Consigo controlar o meu cansaço	0	1	2	3	4
P42. Faço por melhorar as minhas capacidades físicas	0	1	2	3	4
Nas diversas situações da minha vida:					
P43. Sinto-me mais motivado(a) a atingir os meus objetivos porque pratico AF	0	1	2	3	4
P 44. Sinto-me mais confiante nas minhas capacidades porque pratico AF	0	1	2	3	4
P45. Controlo melhor as minhas emoções porque pratico AF	0	1	2	3	4
P46. Controlo melhor o meu cansaço porque pratico AF	0	1	2	3	4

Como é que estas frases te descrevem?	Nada	Pouco	Moderadamente	Bastante	Totalmente
S1. Acredito que os aspetos culturais da(s) AF são importantes (ex. os seus rituais, termos específicos, roupa, valores)	0	1	2	3	4
S2. Participo em rituais das AF (ex. saudações, hinos/cânticos, gritos, aplausos)	0	1	2	3	4
S3. Utilizo termos específicos das AF (ex. nomes de técnicas ou táticas, nomes de equipamento, expressões)	0	1	2	3	4
S4. Utilizo vestuário específico da AF que estou a praticar	0	1	2	3	4
S5. Assisto a eventos de AF (ex. competições, demonstrações, programas)	0	1	2	3	4
S6. Gosto de acompanhar eventos de AF (ex. competições, demonstrações, programas)	0	1	2	3	4
S7. Interesso-me pelos aspetos culturais das AF (ex. os seus rituais, termos específicos, roupa, valores)	0	1	2	3	4
S8. Incentivo os outros a assistirem a eventos de AF (ex. competições, demonstrações, programas)	0	1	2	3	4
S9. Encorajo os outros a participar na cultura de cada AF (ex. rituais, termos específicos, roupa)	0	1	2	3	4
S10. Consigo ter uma vida ativa e saudável	0	1	2	3	4
Em situações de Atividade Física:					
S11. Trabalho bem com os outros	0	1	2	3	4
S12. Tento ter comportamentos corretos e justos	0	1	2	3	4
S13. Respeito os meus adversários	0	1	2	3	4
S14. Cumpro as regras	0	1	2	3	4
S15. Faço batota, se isso me trouxer benefícios	0	1	2	3	4
S16. Respeito as decisões de autoridades (ex. árbitro, juiz, treinador/professor)	0	1	2	3	4
S17. Cumpro os princípios do *fair-play*/ética desportiva	0	1	2	3	4
S18. Compreendo a importância dos princípios do *fair-play*/ética desportiva	0	1	2	3	4
S19. Faço para que os outros respeitem o *fair-play*/ética desportiva	0	1	2	3	4
S20. Cumpro as regras, mesmo sem supervisão	0	1	2	3	4
S21. Cumpro os princípios do *fair-play*/ética desportiva por iniciativa própria	0	1	2	3	4
S22. Faço para que os outros cumpram as regras	0	1	2	3	4

S23. Colaboro com os outros	0	1	2	3	4
S24. Sou compreensivo(a) com os outros	0	1	2	3	4
S25. Controlo o meu comportamento para com os outros	0	1	2	3	4
S26. Respeito os outros	0	1	2	3	4
S27. Coopero com os outros	0	1	2	3	4
S28. Encorajo os outros	0	1	2	3	4
S29. Preocupo-me com o sucesso dos outros	0	1	2	3	4
S30. Ajudo os outros a terem sucesso	0	1	2	3	4
S31. Sou prestável com os outros	0	1	2	3	4
S32. Tenho um relacionamento positivo com os outros	0	1	2	3	4
S33. Interajo com os outros	0	1	2	3	4
S34. Partilho um objetivo comum com os outros	0	1	2	3	4
S35. Sinto-me próximo(a) dos outros	0	1	2	3	4
S36. Sinto que há camaradagem entre mim e os outros	0	1	2	3	4
S37. Faço por melhorar o meu relacionamento com os outros	0	1	2	3	4
Em situações de Atividade Física:					
S38. Sei como fazer para melhorar o meu relacionamento com os outros	0	1	2	3	4
S39. Preocupo-me com a qualidade do relacionamento que tenho com os outros	0	1	2	3	4
Nas diversas situações da minha vida:					
S40. Envolvo-me mais noutras atividades culturais (ex. teatro, música) porque pratico AF	0	1	2	3	4
S41. Sou mais correto(a) e justo(a) porque pratico AF	0	1	2	3	4
S42. Colaboro mais com os outros porque pratico AF	0	1	2	3	4
S43. Tenho melhores relações com os outros porque pratico AF	0	1	2	3	4

Este grupo de questões procura conhecer a tua opinião em relação a temas relacionados com a prática de Atividade Física e estilos de vida saudáveis. **Lê cada uma com atenção e coloca um círculo em volta da(s) alínea(s) escolhida(s).**

C1. Seleciona a opção que representa uma refeição saudável e equilibrada.

(A) Canja de galinha, cheeseburger, maçã, refrigerante.
(B) Sopa de legumes, salmão grelhado com batata-doce e vegetais assados, pera, água.
(C) Pão de alho, massa à carbonara com salada, pera, água.
(D) Sopa de feijão, rissóis de camarão fritos com arroz de tomate, maçã, refrigerante.

C2. Num dia normal, a Maria ingere no total 2000 kcal e tem um dispêndio calórico total de 2200 kcal.

Seleciona a opção que descreve o balanço energético da Maria nesse dia e as consequências no seu peso.

(A) Ela terá um balanço energético negativo e ganhará peso.
(B) Ela terá um balanço energético positivo e ganhará peso.
(C) Ela terá um balanço energético positivo e perderá peso.
(D) Ela terá um balanço energético negativo e perderá peso.

C3. A "capacidade de captar e utilizar o oxigénio para produzir energia" e a "capacidade de mover uma ou mais articulações através da sua total amplitude" descrevem, respetivamente

(A) a força muscular e a flexibilidade.
(B) a aptidão aeróbia e a flexibilidade.
(C) a aptidão aeróbia e a força muscular.
(D) a velocidade e a força muscular.

C4. O João tem pouca força e flexibilidade e quer melhorar estas capacidades.

Seleciona a opção que descreve métodos de treino adequados para o João.

	Treino	Frequência	Volume por grupo muscular/articulação
(A)	**Força**	**2- 3 vezes não consecutivas por semana**	**1 a 3 séries de 12 repetições**
	Flexibilidade	**5-7 vezes por semana**	**30 segundos por alongamento**
(B)	Força	1 vez por semana	1 a 3 séries de 12 repetições
	Flexibilidade	5-7 vezes por semana	30 segundos por alongamento
(C)	Força	2- 3 vezes consecutivas por semana	1 a 3 séries de 12 repetições
	Flexibilidade	1 vez por semana	30 segundos por alongamento
(D)	Força	2- 3 vezes não consecutivas por semana	1 a 3 séries de 5 repetições
	Flexibilidade	5-7 vezes por semana	30 segundos por alongamento

C5. Seleciona **todas** as opções que descrevem procedimentos básicos para prevenir lesões e danos físicos durante a prática de atividade física.

(A) Aquecer no início da atividade e retornar à calma no final.
(B) Aplicar quente imediatamente, em caso de lesão.
(C) Utilizar material de proteção adequado.
(D) Comer uma refeição 10 minutos antes da atividade.
(E) Cumprir as regras de segurança da modalidade.
(F) Hidratar regularmente enquanto se estiver em atividade.

6. Completa o texto, selecionando na tabela a opção adequada a **cada espaço**.

Uma atleta de competição está a ponderar utilizar esteroides anabolizantes. Esta prática é __1.__ dado que __2.__ verdade desportiva, __3.__ desempenho desportivo e __4.__ saúde.

1.	2.	3.	4.
(A) segura **(B) proibida** **(C)** permitida	**(A)** aumenta a **(B) diminui a** **(C)** não tem efeito na	**(A) aumenta o** **(B)** diminui o **(C)** não tem efeito no	**(A) aumenta a probabilidade de problemas de** **(B)** diminui a probabilidade de problemas de **(C)** não tem efeito na

C7. A Organização Mundial de Saúde definiu recomendações para a prática de Atividade Física.

Seleciona a opção que descreve as recomendações gerais para crianças e adolescentes, e para adultos.

(A) Crianças e adolescentes: 60 minutos por dia com intensidade moderada a vigorosa; adultos: 150 minutos por semana com intensidade moderada a vigorosa.
(B) Crianças e adolescentes: 150 minutos por semana com intensidade moderada a vigorosa; **adultos**: 150 minutos por semana com intensidade moderada a vigorosa
(C) Crianças e adolescentes: 30 minutos por dia com intensidade moderada a vigorosa; **adultos**: 150 minutos por semana com intensidade moderada a vigorosa.
(D) Crianças e adolescentes: 60 minutos por dia com intensidade moderada a vigorosa; **adultos**: 90 minutos por semana com intensidade moderada a vigorosa.

C8. Seleciona a opção que associa corretamente o tipo de treino físico aos seus impactos na saúde.

Tipos de Treino Físico	Impactos na Saúde
1.Treino de Força	A. Melhoria da tensão arterial e da saúde cardiovascular.
	B. Aumento da frequência cardíaca de repouso
2.Treino de Flexibilidade	C. Maior mobilidade e capacidade para realizar atividades do dia-a-dia
	D. Maiores dificuldades de coordenação
3.Treino de Aptidão Aeróbia	E. Melhoria nos ossos, tendões e ligamentos e menor risco de lesões

(A) 1-D, 2-C, 3-A
(B) 1-A, 2-B, 3-C
(C) 1-E, 2-B, 3-D
(D) 1-E, 2-C, 3-A

C9. A fórmula de cálculo do Índice de Massa Corporal (IMC) é

(A) $\frac{\boldsymbol{Peso\ (kg)}}{\boldsymbol{Altura\ (m)^2}}$

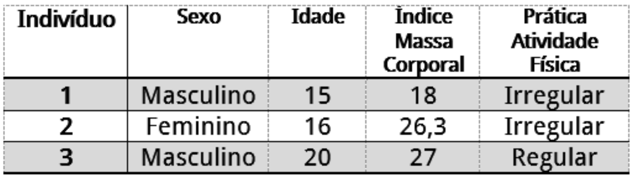
(B) $\frac{Altura\ (cm)^2}{Peso\ (kg)}$

(C) $\frac{Peso\ (kg)}{Altura\ (cm)^2}$

(D) $\frac{Altura\ (m)}{Peso\ (kg)^2}$

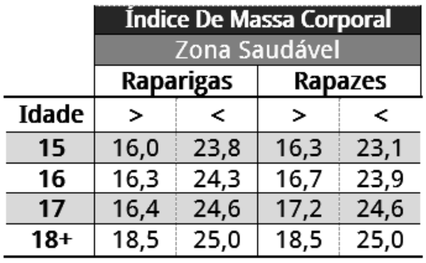
C10. Seleciona, com base na análise das duas tabelas, **todas** as afirmações verdadeiras.

Tabela 1.



Indivíduo	Sexo	Idade	Índice Massa Corporal	Prática Atividade Física
1	Masculino	15	18	Irregular
2	Feminino	16	26,3	Irregular
3	Masculino	20	27	Regular

Tabela 2. Valores de referência de IMC (FITescola)



	Índice De Massa Corporal			
	Zona Saudável			
	Raparigas		**Rapazes**	
Idade	>	<	>	<
15	16,0	23,8	16,3	23,1
16	16,3	24,3	16,7	23,9
17	16,4	24,6	17,2	24,6
18+	18,5	25,0	18,5	25,0

(A) O indivíduo 1 deveria praticar atividade física regularmente.
(B) O indivíduo 1 encontra-se abaixo da zona saudável de IMC.
(C) O indivíduo 2 deveria reduzir o seu peso.
(D) O indivíduo 3 deveria manter o seu peso.
(E) O indivíduo 2 tem maior risco de doenças que o indivíduo 3.

Confirma por favor que respondeste a todas as questões.
Muito obrigado pela tua imprescindível participação!

Additional File 4

Old label (v0.6)	New label (v1.0)
P1	MOT4
P2[R]	MOT3[R]
P5[R]	MOT1[R]
P7	MOT5
P8	MOT2
P11	MOT7
P43	MOT6
P13	CON1
P14	CON6
P16	CON4
P18	CON2
P19	CON5
P20	CON9
P21	CON7
P22	CON8
P44	CON3
P23	EMO5
P25	EMO1
P28	EMO2
P29	EMO7
P30	EMO3
P31	EMO4
P32	EMO6
P33	PHY6
P35	PHY1
P36	PHY2
P37	PHY5
P38	PHY3
P39	PHY7
P41	PHY8
P42	PHY4
S2	CUL6
S3	CUL3
S5	CUL2
S6	CUL1
S7	CUL4
S8	CUL6
S9	CUL7
S12	ETH6
S13	ETH1
S14	ETH2
S17	ETH3
S19	ETH5
S21	ETH4
S23	COL3
S24	COL4
S26	COL1
S27	COL2
S28	COL5
S30	COL6
S32	REL1
S33	REL2
S34	REL5
S36	REL4
S37	REL3
S38	REL6

[R] Item should be reverse scored during analysis

English version

For the next question group, you should remember that PA = Physical Activity(ies), and includes every situation that require movement, like your Physical Education classes, sport-based activities (team or individual), rhythmic activities (e.g., dance), exercise (e.g., strength training, jogging), and any activities that you use as a mean of transportation or in your spare time.

In each statement below, select how much it describes you, circling one of the options from 0 to 4 0 = ***Not at all***; 1= ***Slightly***; 2 =***Moderately***; 3 =***Quite a lot***; 4 = ***Totally***).

How much do these statements describe you?	**Not at all**	**Slightly**	**Moderately**	**Quite a lot**	**Totally**
MOT4. I am motivated to practice PA	0	1	2	3	4
MOT3[R]. I practice PA because others tell me I should	0	1	2	3	4
MOT1[R]. I feel pressured by others to practice PA	0	1	2	3	4
MOT5. I practice PA because it is fun	0	1	2	3	4
MOT2. I feel good when I practice PA	0	1	2	3	4
MOT7. I see PA as a fundamental part of who I am	0	1	2	3	4
PSY. I enjoy practicing PA	0	1	2	3	4
CON1. I feel confident to practice PA	0	1	2	3	4
In Physical Activity contexts:					
CON6. I am confident in my abilities	0	1	2	3	4
CON4. I consider myself competent	0	1	2	3	4
CON2. I feel good about the way I can participate	0	1	2	3	4
CON5. I can participate in PA that I consider challenging	0	1	2	3	4
CON9. I know how to become more confident in myself	0	1	2	3	4
CON7. I feel competent even when I am criticized	0	1	2	3	4
CON8. I believe in myself even when I lose	0	1	2	3	4
EMO5. I can manage my emotions	0	1	2	3	4
EMO1. I can recognize my emotions	0	1	2	3	4
EMO2. I can identify what I feel	0	1	2	3	4
EMO7. I can anticipate what I will feel	0	1	2	3	4
EMO3. I can deal with difficulties rationally	0	1	2	3	4
EMO4. I can manage my emotions when necessary	0	1	2	3	4
EMO6. I have a good control of my emotions	0	1	2	3	4
PHY6. I can manage my effort	0	1	2	3	4
PHY1. I can recognize changes in my breathing	0	1	2	3	4

In Physical Activity contexts:					
PHY2. I can recognize changes in my heart rate	0	1	2	3	4
PHY5. I recognize my physical limits	0	1	2	3	4
PHY3. I can recognize the effect that different intensities have in me	0	1	2	3	4
PHY7. I use strategies to manage my effort	0	1	2	3	4
PHY8. I can control my fatigue	0	1	2	3	4
PHY4. I take action to improve my physical skills	0	1	2	3	4
In the different contexts of my life:					
MOT6. I feel more motivated to reach my goals because I practice PA	0	1	2	3	4
CON3. I feel more confident in my skills because I practice PA	0	1	2	3	4

How much do these statements describe you?	Not at all	Slightly	Moderately	Quite a lot	Totally
CUL6. I participate in PA rituals (e.g., greetings, hymns/chants, cheers, applauses)	0	1	2	3	4
CUL3. I use specific PA terminology (e.g., names of technics and tactics, names of equipment, idioms)	0	1	2	3	4
CUL2. I watch PA events (e.g., competitions, spectacles, shows)	0	1	2	3	4
CUL1. I like to keep up with PA events (e.g., competitions, spectacles, shows)]	0	1	2	3	4
CUL4. I am interested in the cultural aspects of PA (e.g., its rituals, terminology, clothing, values)]	0	1	2	3	4
CUL6. I encourage others to watch PA events (e.g., competitions, spectacles, shows)	0	1	2	3	4
CUL7. I encourage others to participate in each PA's culture (e.g., rituals, terminology, clothing)]	0	1	2	3	4
PL. I can lead a healthy and active life	0	1	2	3	4
In Physical Activity contexts:					
SOC. I work well with others	0	1	2	3	4
ETH6. I try to behave correctly and justly	0	1	2	3	4
ETH1. I respect my adversaries	0	1	2	3	4
ETH2. I follow the rules	0	1	2	3	4
ETH3. I behave according to fair-play / sport ethics 'principles	0	1	2	3	4
ETH5. I take action to make others behave according to fair play/sport ethics	0	1	2	3	4
ETH4. I behave according to fair play/sport ethics' principles on my initiative	0	1	2	3	4
COL3. I collaborate with others	0	1	2	3	4

In Physical Activity contexts:					
COL4. I am sympathetic with others	0	1	2	3	4
COL1. I respect others	0	1	2	3	4
COL2. I cooperate with others	0	1	2	3	4
COL5. I encourage others	0	1	2	3	4
COL6. I help others achieve success	0	1	2	3	4
REL1. I have a positive relationship with others	0	1	2	3	4
REL2. I interact with others	0	1	2	3	4
REL5. I share a common goal with others	0	1	2	3	4
REL4. I feel a sense of camaraderie with others	0	1	2	3	4
REL3. I take action to improve my relationship with others	0	1	2	3	4
REL6. I know how to improve my relationship with others	0	1	2	3	4

Check that you answered all questions.

Thank you for your invaluable participation!

Versão Portuguesa

Para responderes ao próximo grupo, deves relembrar que **AF = Atividade(s) Física(s)**, e que inclui todas as situações que impliquem movimento, como por exemplo, as aulas de Educação Física, as atividades desportivas (coletivas ou individuais), as atividades rítmicas (ex. dança), o exercício físico (ex. treino de força, *jogging*), e as atividades físicas que faças para te deslocar ou nos teus tempos livres.

Para cada afirmação abaixo deves selecionar o quanto essa te descreve circundando uma das opções de 0 a 4 (**0= Nada, 1= Pouco, 2= Moderadamente, 3= Bastante, 4= Totalmente**).

Como é que estas frases te descrevem?	Nada	Pouco	Moderadamente	Bastante	Totalmente
MOT4. Sinto-me motivado(a) para praticar AF	0	1	2	3	4
MOT3[I]. Faço AF porque outras pessoas dizem que devo fazer	0	1	2	3	4
MOT1[I]. Sinto-me pressionado(a) pelos outros para fazer AF	0	1	2	3	4
MOT5. Faço AF porque é divertido	0	1	2	3	4
MOT2. Sinto-me bem quando faço AF	0	1	2	3	4
MOT7. Vejo a AF como parte fundamental de quem sou	0	1	2	3	4
PSY. Gosto de praticar AF	0	1	2	3	4
CON1. Sinto-me confiante para praticar AF	0	1	2	3	4
Em situações de Atividade Física:					
CON6. Sinto-me confiante nas minhas habilidades	0	1	2	3	4
CON4. Considero-me bom/boa praticante	0	1	2	3	4
CON2. Sinto-me bem com a forma como consigo participar	0	1	2	3	4
CON5. Consigo participar em AF que considero desafiantes	0	1	2	3	4
CON9. Sei como fazer para me tornar mais confiante	0	1	2	3	4
CON7. Sinto-me capaz mesmo quando sou criticado(a)	0	1	2	3	4
CON8. Acredito na minha capacidade mesmo quando falho	0	1	2	3	4
EMO5. Consigo gerir as minhas emoções	0	1	2	3	4
EMO1. Consigo reconhecer as minhas emoções	0	1	2	3	4
EMO2. Consigo identificar o que sinto	0	1	2	3	4
EMO7. Consigo antecipar as minhas emoções	0	1	2	3	4
EMO3. Consigo lidar com as dificuldades racionalmente	0	1	2	3	4
EMO4. Consigo gerir as minhas emoções quando necessário	0	1	2	3	4
EMO6. Tenho um bom controlo das minhas emoções	0	1	2	3	4
PHY6. Consigo gerir o meu esforço	0	1	2	3	4
PHY1. Consigo aperceber-me de mudanças na minha respiração	0	1	2	3	4

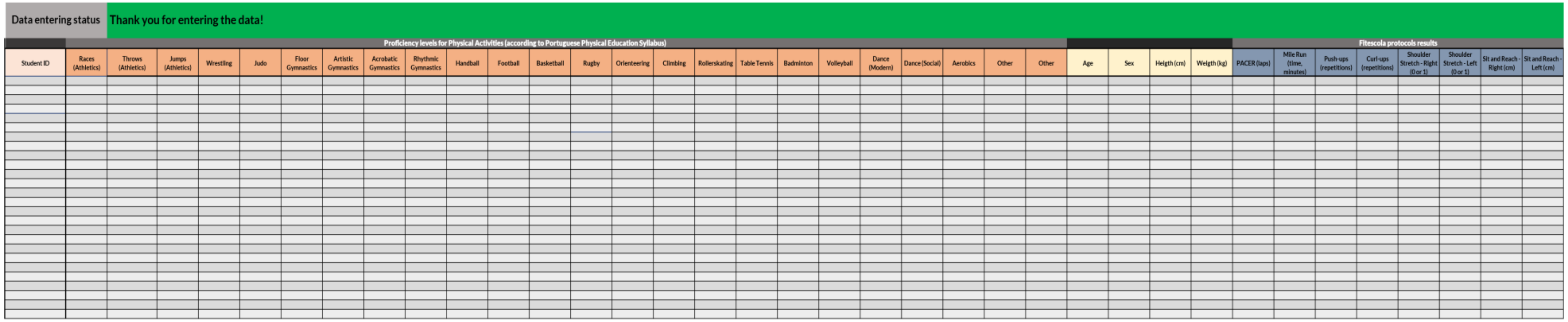
Em situações de Atividade Física:					
PHY2. Consigo reconhecer mudanças no meu ritmo cardíaco	0	1	2	3	4
PHY5. Tenho noção dos meus limites físicos	0	1	2	3	4
PHY3. Reconheço o efeito que diferentes intensidades de esforço têm em mim	0	1	2	3	4
PHY7. Utilizo estratégias para gerir o meu esforço	0	1	2	3	4
PHY8. Consigo controlar o meu cansaço	0	1	2	3	4
PHY4. Faço por melhorar as minhas capacidades físicas	0	1	2	3	4
Nas diversas situações da minha vida:					
MOT6. Sinto-me mais motivado(a) a atingir os meus objetivos porque pratico AF	0	1	2	3	4
CON3. Sinto-me mais confiante nas minhas capacidades porque pratico AF	0	1	2	3	4

Como é que estas frases te descrevem?	Nada	Pouco	Moderadamente	Bastante	Totalmente
CUL6. Participo em rituais das AF (ex. saudações, hinos/cânticos, gritos, aplausos)	0	1	2	3	4
CUL3. Utilizo termos específicos das AF (ex. nomes de técnicas ou táticas, nomes de equipamento, expressões)	0	1	2	3	4
CUL2. Assisto a eventos de AF (ex. competições, demonstrações, programas)	0	1	2	3	4
CUL1. Gosto de acompanhar eventos de AF (ex. competições, demonstrações, programas)	0	1	2	3	4
CUL4. Interesso-me pelos aspetos culturais das AF (ex. os seus rituais, termos específicos, roupa, valores)	0	1	2	3	4
CUL6. Incentivo os outros a assistirem a eventos de AF (ex. competições, demonstrações, programas)	0	1	2	3	4
CUL7. Encorajo os outros a participar na cultura de cada AF (ex. rituais, termos específicos, roupa)	0	1	2	3	4
PL. Consigo ter uma vida ativa e saudável	0	1	2	3	4
Em situações de Atividade Física:					
SOC. Trabalho bem com os outros	0	1	2	3	4
ETH6. Tento ter comportamentos corretos e justos	0	1	2	3	4
ETH1. Respeito os meus adversários	0	1	2	3	4
ETH2. Cumpro as regras	0	1	2	3	4
ETH3. Cumpro os princípios do *fair-play*/ética desportiva	0	1	2	3	4
ETH5. Faço para que os outros respeitem o *fair-play*/ética desportiva	0	1	2	3	4

Em situações de Atividade Física:					
ETH4. Cumpro os princípios do *fair-play*/ética desportiva por iniciativa própria	0	1	2	3	4
COL3. Colaboro com os outros	0	1	2	3	4
COL4. Sou compreensivo(a) com os outros	0	1	2	3	4
COL1. Respeito os outros	0	1	2	3	4
COL2. Coopero com os outros	0	1	2	3	4
COL5. Encorajo os outros	0	1	2	3	4
COL6. Ajudo os outros a terem sucesso	0	1	2	3	4
REL1. Tenho um relacionamento positivo com os outros	0	1	2	3	4
REL2. Interajo com os outros	0	1	2	3	4
REL5. Partilho um objetivo comum com os outros	0	1	2	3	4
REL4. Sinto que há camaradagem entre mim e os outros	0	1	2	3	4
REL3. Faço por melhorar o meu relacionamento com os outros	0	1	2	3	4
REL6. Sei como fazer para melhorar o meu relacionamento com os outros	0	1	2	3	4

Confirma por favor que respondeste a todas as questões.
Muito obrigado pela tua imprescindível participação!

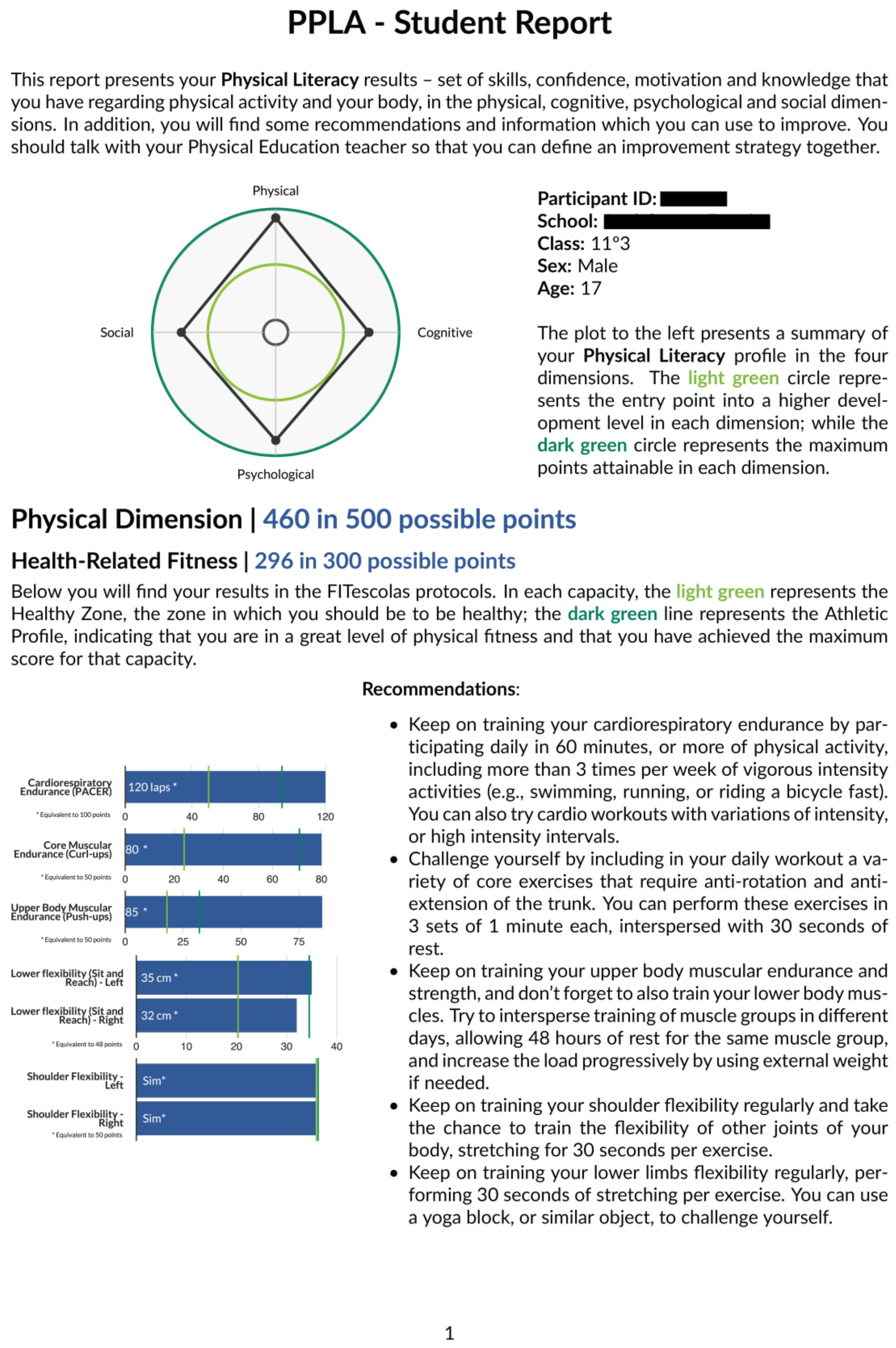
Additional File 5



Data entering status | Thank you for entering the data!

	Proficiency levels for Physical Activities (according to Portuguese Physical Education Syllabus)																												Fitescola protocols results							
Student ID	Races (Athletics)	Throws (Athletics)	Jumps (Athletics)	Wrestling	Judo	Floor Gymnastics	Artistic Gymnastics	Acrobatic Gymnastics	Rhythmic Gymnastics	Handball	Football	Basketball	Rugby	Orienteering	Climbing	Rollerskating	Table Tennis	Badminton	Volleyball	Dance (Modern)	Dance (Social)	Aerobics	Other	Other	Age	Sex	Heigth (cm)	Weigth (kg)	PACER (laps)	Mile Run (time, minutes)	Push-ups (repetitions)	Curl-ups (repetitions)	Shoulder Stretch - Right (0 or 1)	Shoulder Stretch - Left (0 or 1)	Sit and Reach - Right (cm)	Sit and Reach - Left (cm)

Additional File 6



PPLA - Student Report

This report presents your **Physical Literacy** results – set of skills, confidence, motivation and knowledge that you have regarding physical activity and your body, in the physical, cognitive, psychological and social dimensions. In addition, you will find some recommendations and information which you can use to improve. You should talk with your Physical Education teacher so that you can define an improvement strategy together.

Participant ID: [redacted]
School: [redacted]
Class: 11°3
Sex: Male
Age: 17

The plot to the left presents a summary of your **Physical Literacy** profile in the four dimensions. The light green circle represents the entry point into a higher development level in each dimension; while the dark green circle represents the maximum points attainable in each dimension.

Physical Dimension | 460 in 500 possible points

Health-Related Fitness | 296 in 300 possible points

Below you will find your results in the FITescolas protocols. In each capacity, the light green represents the Healthy Zone, the zone in which you should be to be healthy; the dark green line represents the Athletic Profile, indicating that you are in a great level of physical fitness and that you have achieved the maximum score for that capacity.

Cardiorespiratory Endurance (PACER)
120 laps *
* Equivalent to 100 points
0
40
80
120
Core Muscular Endurance (Curl-ups)
80 *
* Equivalent to 50 points
0
20
40
60
80
Upper Body Muscular Endurance (Push-ups)
85 *
* Equivalent to 50 points
0
25
50
75
Lower flexibility (Sit and Reach) - Left
35 cm *
Lower flexibility (Sit and Reach) - Right
32 cm *
* Equivalent to 48 points
0
10
20
30
40
Shoulder Flexibility - Left
Sim*
Shoulder Flexibility - Right
Sim*
* Equivalent to 50 points

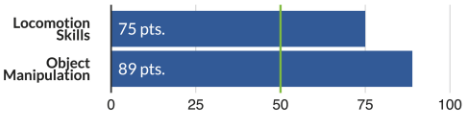
Recommendations:

- Keep on training your cardiorespiratory endurance by participating daily in 60 minutes, or more of physical activity, including more than 3 times per week of vigorous intensity activities (e.g., swimming, running, or riding a bicycle fast). You can also try cardio workouts with variations of intensity, or high intensity intervals.
- Challenge yourself by including in your daily workout a variety of core exercises that require anti-rotation and anti-extension of the trunk. You can perform these exercises in 3 sets of 1 minute each, interspersed with 30 seconds of rest.
- Keep on training your upper body muscular endurance and strength, and don't forget to also train your lower body muscles. Try to intersperse training of muscle groups in different days, allowing 48 hours of rest for the same muscle group, and increase the load progressively by using external weight if needed.
- Keep on training your shoulder flexibility regularly and take the chance to train the flexibility of other joints of your body, stretching for 30 seconds per exercise.
- Keep on training your lower limbs flexibility regularly, performing 30 seconds of stretching per exercise. You can use a yoga block, or similar object, to challenge yourself.

1

Movement Skills | 164 in 200 possible points

Below you will find your Movement Skills results divided into **Object Manipulation** - your skill in controlling object with your hands or feet (e.g., throw, kick, hit with a racquet) – and **Locomotion Skills** - related with whole-body control skills (e.g., jumping, rolling, running).
The light green line marks the transition to a higher level of development in each element, with 100 points as the maximum score.

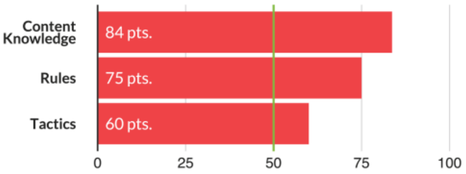


Recommendations:

- Perfect your locomotion skills through complex exercises or situations that involve these skills (e.g., Gymnastics, Dance, Athletics); try to transfer these skills between physical activities you have mastered and others in which you are improving.
- Perfect your object manipulation skills through complex exercises or situations that involve these skills (e.g., Team Sports, Racquets); try to transfer these skills between physical activities you have mastered and others in which you are improving.

Cognitive Dimension | 219 in 300 possible pointss

This dimension includes elements like **Content Knowledge** about themes related with healthy and active lifestyles, as well as your **Tactics** and **Rules** application in physical activities.
The light green line marks the transition to a higher level of development in each element, with 100 points as the maximum score.

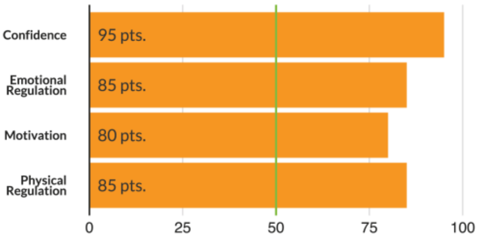


Recommendations:

- You should keep on learning about the following themes:
  - Safety and Risk
  - Benefits of Physical Activity
  - Body Composition
- Support other during physical activities and help them learn and respect existing rules. You can also try the role of referee or judge.
- Perfect your tactical skills by participating in diverse physical activities that include this component. Work on adapting your behaviors in function of your opponent, be that individually or in team.

Psychological Dimension | 345 in 400 possible points

This dimension includes elements like your **Motivation** and **Confidence** to practice physical activity, as well as the way you manage your emotions (**Emotional Regulation**) and effort (**Physical Regulation**) during this practice.
The light green line marks the transition to a higher level of development in each element, with 100 points as the maximum score.

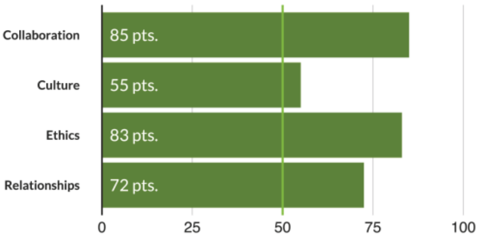


Recommendations:

- Participate in physical activities that challenge you; continuously evaluate your skills and set improvement goals.
- Participate in physical activities that you enjoy and try to acknowledge the benefits that physical activity offers you.
- Keep on developing and applying strategies that allow you to regulate your emotions during physical activity (e.g., deep breaths, visualization).
- Keep on developing and applying strategies that allow you to regulate your effort and persist in intense physical activities.

Social Dimension | 296 in 400 possible points

This dimension includes elements related with your interactions with others during physical activity. **Culture** represents the way you participate in the culture of sport/movement (e.g., through its specific language/terms, rituals, and values); **Ethics** represents your adherence and respect to fair play/sport ethics principles; **Collaboration** refers to the way you respect and collaborate with others; and **Relationships** represents your sense of connection and relationship with others.
The light green line marks the transition to a higher level of development in each element, with 100 points as the maximum score.



Recommendations:

- Take the chance to encourage and support the success of others during physical activities.
- Explore the values present in movement/sport culture and experiment with ways to share its meaning with others.
- Keep on exploring and applying fair play/sport ethics and encourage others to do so in physical activities.
- Invest in relationships that you have with others during physical activity, always trying to keep your interactions positive and encouraging.

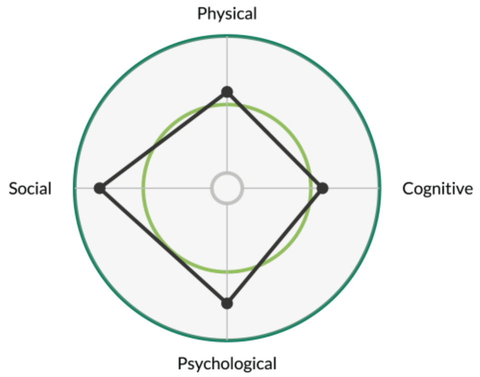
Additional File 7

PPLA - Class Report

This report presents the results of your class's **Physical Literacy** — set of skills, competences, confidence, motivation and knowledge regarding physical activity and one's body, in the physical, cognitive, psychological and social dimensions. You might notice that each of these dimensions is divided into elements. In those elements, results will be presented in two main development levels: **foundation** and **mastery**. The foundation level refers to the development of essential skills in each element and is related with initial learning experiences; the mastery level refers to a deeper level of the respective skills, requiring complex learning tasks. The transition between these levels will be marked by a light green line in all plots – except in Health-Related Fitness, since standards are dependent on age and sex.

In the **Psychological** and **Social** dimensions, you will notice bars of two different colors for each element, since each element was assessed through two different scales; the lighter of the two corresponds to the foundation level score, while the darker corresponds to the mastery score – allowing to better adapt the learning strategy required, since a student might mostly be in a mastery level (already over the light green) line), but still have gaps in the foundation skills.

In addition to the result, you will also receive some suggestions of teaching strategies that you may use to improve your student's development in each element – both in the foundation and mastery levels.



School: [redacted]
Class: 11ºC

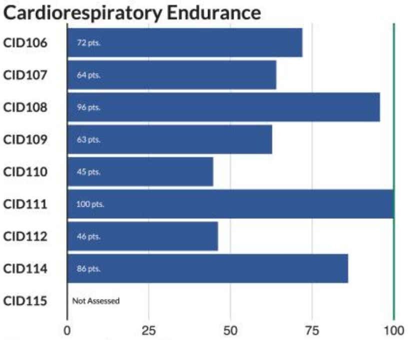
The plot to the left presents a summary of the **Physical Literacy** profile of your class in the four dimensions. The light green circle represents the entry point into the mastery level in each dimension; while the dark green circle represents the maximum points attainable in each dimension.

Physical Dimension | Class average: 296 in 500 possible points

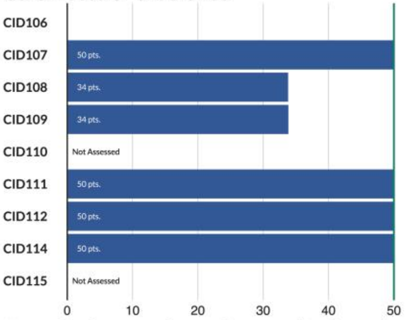
Health-Related Fitness | Class average: 205 in 300 possible points

Below you will find the FITescola protocols results. The **dark green** line represents the maximum score (correspondent to the Athletic Profile). In each individual report, students have access to a light green line that represents the Healthy Zone (mark the transitions between foundation and mastery level, in the fitness case) – since standards for these results are based on age and sex, it is not possible to display them in these class plots.

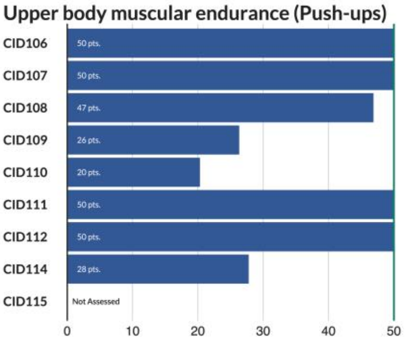
Cardiorespiratory Endurance



Core muscular endurance



Upper body muscular endurance (Push-ups)



Suggested strategies:

Cardiorespiratory Endurance

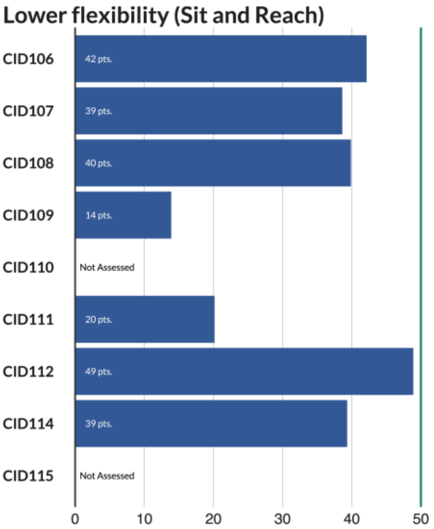
- **Foudantion**: Provide regular opportunities for students to train their cardiorespiratory endurance, promoting that they participate daily in 60 minutes of physical activity, and include at least 3 times per week vigorous intensity activities (e.g., swimming, running, or riding a bicycle fast).
- **Mastery**: Provide regular opportunities for students to train their cardiorespiratory endurance, promoting that they participate daily in 60 or more minutes of physical activity, and include vigorous intensity activities (e.g., swimming, running, or riding a bicycle fast) more than 3 times per week. Explore the use of high intensity interval training during classes. Teach students basic periodization and training methods for cardiorespiratory endurance training.

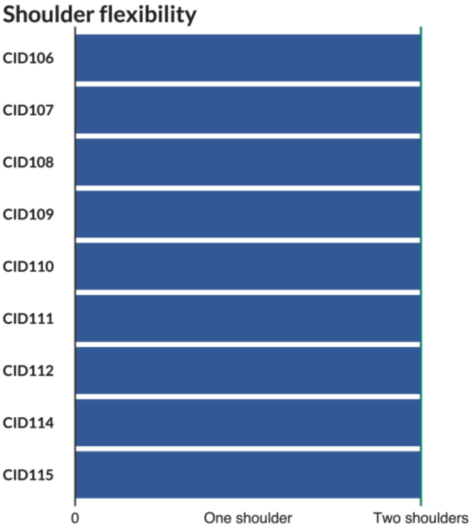
Core muscle endurance

- **Foundation**: Provide opportunities for students to train their core muscle endurance in every class and learn how to train this capacity during the week (e.g., daily workouts with 3 sets of 30 seconds exercises, interspersed with 30 seconds of rest).
- **Mastery**: Provide opportunities for students to train their core muscle endurance with a variety of exercises that involve anti-rotation and anti-extension of the trunk (e.g., using 3 sets of 1 minute, interspersed with 30 seconds or rest).

Upper body muscle endurance

- **Foundation**: Provide opportunities for students to train their upper body strength and endurance during class and learn how to train this capacity during the week (e.g., 3 times per week, 3 sets of 8 to 15 repetitions per muscle group, with 30 seconds of rest between sets).
- **Mastery**: Provide opportunities for students to train their upper and lower muscle strength and endurance; for more advanced students, usage of additional weight might be beneficial. Teach students basic periodization and training methods for strength training.


Lower flexibility (Sit and Reach)

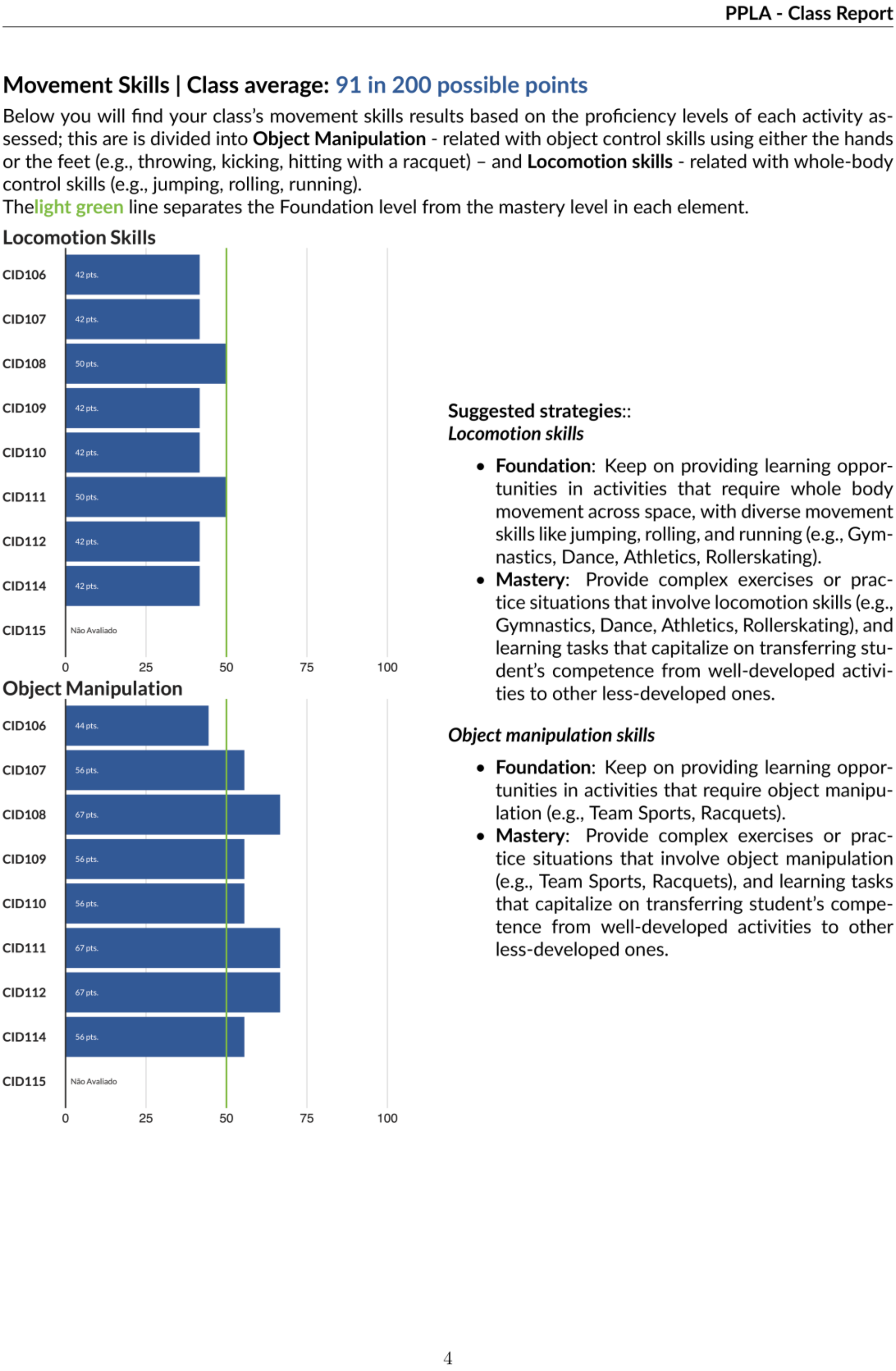

Shoulder flexibility

Lower flexibility

- **Foundation**: Provide opportunities for students to train lower limb flexibility in every class, using, for example, a towel or elastic band to help stretching during 30 seconds per leg. Prescribe daily training of this capacity to students.
- **Mastery**: Provide regular and diverse opportunities for training lower limb flexibility, using, for example a yoga block or similar object to increase the challenge of stretching exercises. Teach students basic periodization and training methods for flexibility training.

Shoulder flexibility

- **Foundation**: Provide opportunities for students to train shoulder flexibility in every class, using, for example, a towel or elastic band to help stretching during 30 seconds per leg. Prescribe daily training of this capacity to students.
- **Mastery**: Provide regular and diverse opportunities for training upper limb flexibility, maintaining the stretch for 30 seconds per exercise, and using techniques to increase the challenge (e.g., peer help, use of walls or wall bars).



Movement Skills | Class average: 91 in 200 possible points

Below you will find your class's movement skills results based on the proficiency levels of each activity assessed; this are is divided into **Object Manipulation** - related with object control skills using either the hands or the feet (e.g., throwing, kicking, hitting with a racquet) – and **Locomotion skills** - related with whole-body control skills (e.g., jumping, rolling, running).

Thelight green line separates the Foundation level from the mastery level in each element.

Locomotion Skills

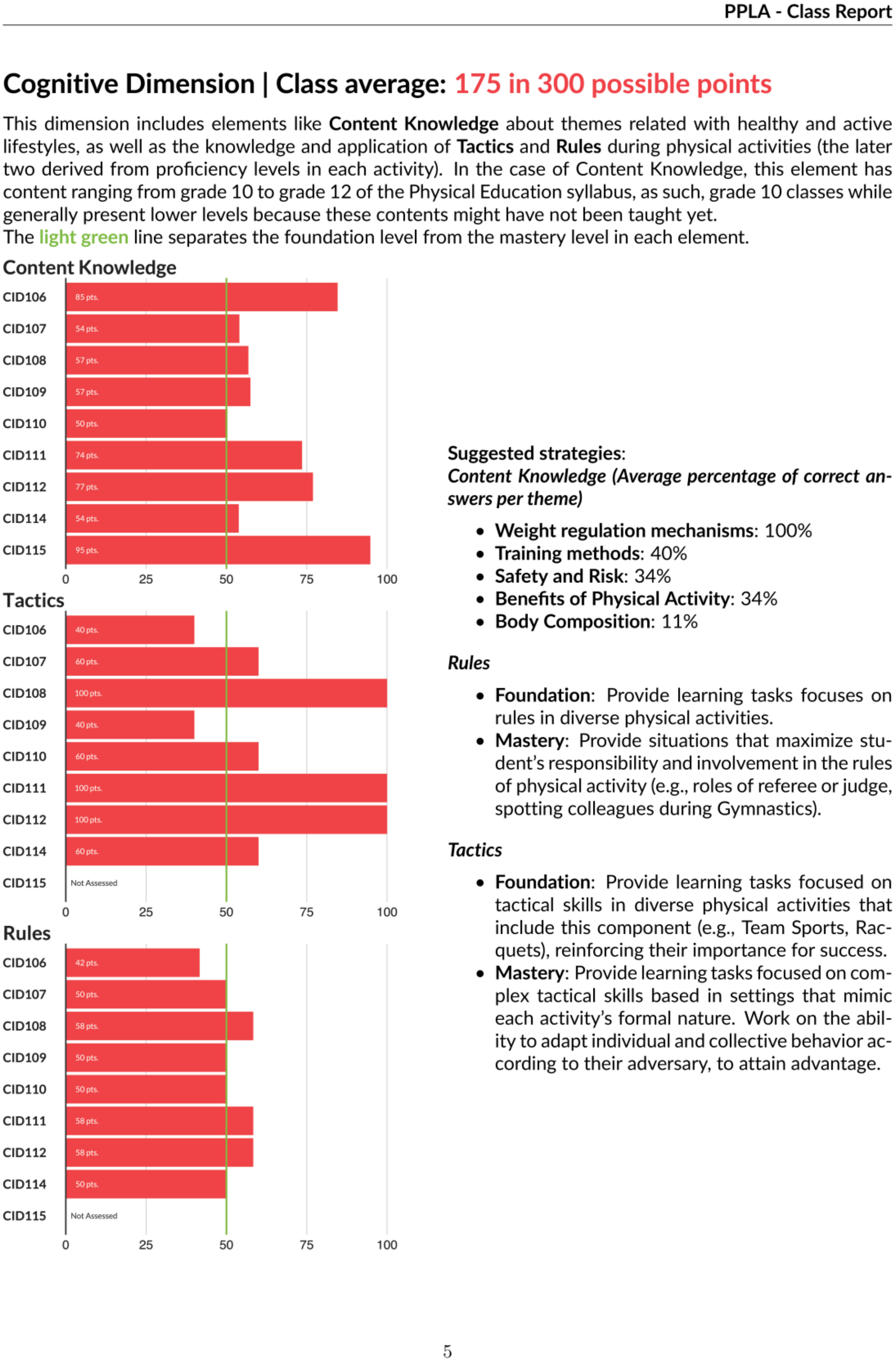
Object Manipulation

Suggested strategies::
Locomotion skills

- **Foundation**: Keep on providing learning opportunities in activities that require whole body movement across space, with diverse movement skills like jumping, rolling, and running (e.g., Gymnastics, Dance, Athletics, Rollerskating).
- **Mastery**: Provide complex exercises or practice situations that involve locomotion skills (e.g., Gymnastics, Dance, Athletics, Rollerskating), and learning tasks that capitalize on transferring student's competence from well-developed activities to other less-developed ones.

Object manipulation skills

- **Foundation**: Keep on providing learning opportunities in activities that require object manipulation (e.g., Team Sports, Racquets).
- **Mastery**: Provide complex exercises or practice situations that involve object manipulation (e.g., Team Sports, Racquets), and learning tasks that capitalize on transferring student's competence from well-developed activities to other less-developed ones.



Cognitive Dimension | Class average: 175 in 300 possible points

This dimension includes elements like **Content Knowledge** about themes related with healthy and active lifestyles, as well as the knowledge and application of **Tactics** and **Rules** during physical activities (the later two derived from proficiency levels in each activity). In the case of Content Knowledge, this element has content ranging from grade 10 to grade 12 of the Physical Education syllabus, as such, grade 10 classes while generally present lower levels because these contents might have not been taught yet.
The light green line separates the foundation level from the mastery level in each element.

Content Knowledge

Tactics

Rules

Suggested strategies:
Content Knowledge (Average percentage of correct answers per theme)

- **Weight regulation mechanisms**: 100%
- **Training methods**: 40%
- **Safety and Risk**: 34%
- **Benefits of Physical Activity**: 34%
- **Body Composition**: 11%

Rules

- **Foundation**: Provide learning tasks focuses on rules in diverse physical activities.
- **Mastery**: Provide situations that maximize student's responsibility and involvement in the rules of physical activity (e.g., roles of referee or judge, spotting colleagues during Gymnastics).

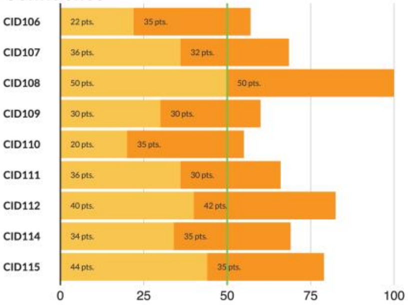
Tactics

- **Foundation**: Provide learning tasks focused on tactical skills in diverse physical activities that include this component (e.g., Team Sports, Racquets), reinforcing their importance for success.
- **Mastery**: Provide learning tasks focused on complex tactical skills based in settings that mimic each activity's formal nature. Work on the ability to adapt individual and collective behavior according to their adversary, to attain advantage.

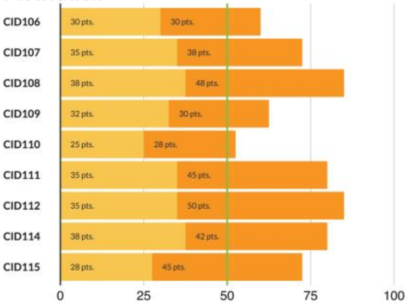
Psychological Dimension | Class average: 292 in 400 possible points

This dimension includes elements like **Motivation** and **Confidence** for practicing physical activity, as well as the way students manage their emotions (**Emotional Regulation**) and effort (**Physical Regulation**) during this practice.

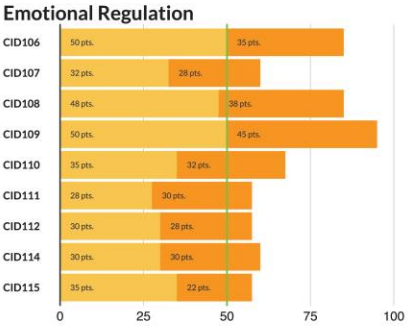
Confidence



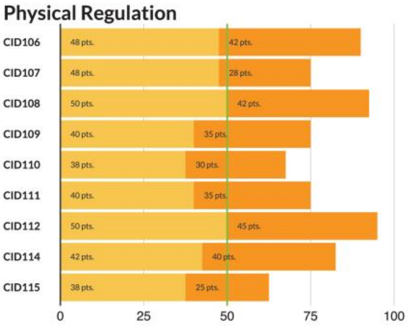
Motivation



Emotional Regulation



Physical Regulation



Suggested strategies:

Confidence

- **Foundation**: Provide opportunities for students to feel competent in diverse physical activities and learn how to identify areas of future improvement.
- **Mastery**: Provide challenging activities and opportunities for on-going evaluation of individual skills and definition of learning goals. Discuss the possibility of using these evaluation skills in other settings of their life.

Motivation

- **Foundation**: Support students in searching and participating in physical activities that they enjoy and reinforce its benefits (e.g., well-being and relaxation). Maximize opportunities where students can choose activities, and where they can bond with their colleagues.
- **Mastery**: Support students in reflecting on how physical activity fits into their identities and life goals. Continue to reinforce their sense of autonomy and involvement with others.

Emotional Regulation

- **Foundation**: Include opportunities for students to identify their emotions during physical activity (e.g., happiness, anger, fear), encouraging them to understand its causes.
- **Mastery**: Provide learning tasks focused on emotional regulation techniques (e.g., deep breathing, visualization). Discuss the possibility of using these skills in other settings of their life.

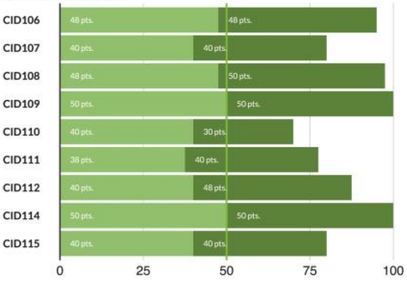
Physical Regulation

- **Foundation**: Include opportunities for students to identify physical indicators/signs like fatigue, pain, increase in breath and heart rate, and to learn about its physiological mechanisms.
- **Mastery**: Provide learning tasks focused on physical regulation/effort management in intense physical activities. Discuss the possibility of using these regulation skills in other settings of their life.

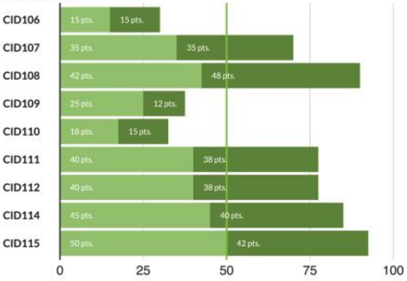
Social Dimension | Class average: 327 in 400 possible points

This dimension includes element pertaining to student's interaction with others. **Culture** represents the way students participate in sport/movement culture (e.g., through its specific terminology, rituals and values); **Ethics** represents the respect and adherence to fair-play/sport's ethics principles; **Collaboration** refers to the way students respect and collaborate with others; and **Relationships** represents the sense of connection and relationship with others.

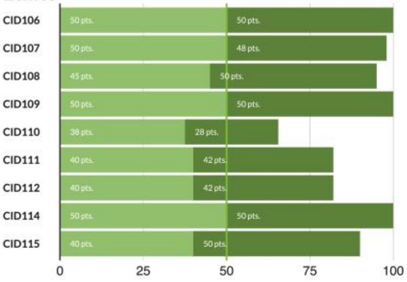
Collaboration



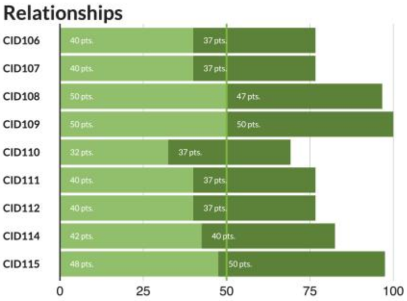
Culture



Ethics



Relationships



Suggested strategies:
Collaboration

- **Foundation**: Provide active cooperation opportunities in diverse physical activities, reinforcing its importance towards success, safety, and well-being.
- **Mastery**: Provide leadership (e.g., group/team leader) and mutual assistance (e.g., teaching styles that involve peer-supported learning) opportunities within class. Discuss the possibility of using these leadership and collaboration skills in other settings of their life.

Culture

- **Foundation**: Promote usage of technical terms referent to each physical activity and provide opportunities for students to create simple team rituals (e.g., cheers, applauses). Discuss cultural aspects of activities (e.g., their history, and values.
- **Mastery**: Support students' exploration of the values inherent to the sport/movement culture and sharing of its meaning with others (e.g., encouraging students to follow events of physical activities that they enjoy).

Ethics

- **Foundation**: Provide learning tasks focused on fair-play/sport ethics principles (formal, e.g., rules, and informal, e.g., etiquette guidelines), as well as opportunities to apply them during physical activity.
- **Mastery**: Provide learning tasks that promote self-regulation of fair-play/ethical behavior (e.g., games without supervision of a referee) and hetero-regulation among peers (e.g., roles of referee or judge, evaluation of adherence to informal rules of physical activity).

Relationships

- **Foundation**: Promote opportunities that reinforce student's sense of belonging to the class, and to the diverse practice groups.
- **Mastery**: Promote opportunities in which students can serve as role models and learning agents, making them responsible for being active developers of their relationship with others during physical activities. Discuss the possibility of using these skills in other settings of their life.