AKADÉMIAI KIADÓ

Journal of Behavioral Addictions

11 (2022) 2, 326–347

DOI:
10.1556/2006.2022.00014
© 2022 The Author(s)

REVIEW ARTICLE

Examining the reliability of the scores of self-report instruments assessing problematic exercise: A systematic review and meta-analysis

MANUEL ALCARAZ-IBÁÑEZ[1†], ADRIAN PATERNA[1*†], ÁLVARO SICILIA[1] and MARK D. GRIFFITHS[2]

[1] Health Research Centre and Department of Education, University of Almería, Spain

[2] Psychology Department, Nottingham Trent University, UK

Received: July 8, 2021 • Revised manuscript received: November 9, 2021; February 12, 2022 • Accepted: March 12, 2022
Published online: April 28, 2022

ABSTRACT

Background and aims: Problematic exercise (PE) has mainly been assessed with self-report instruments. However, summarized evidence on the reliability of the scores derived from such instruments has yet to be provided. The present study reports a reliability generalization meta-analysis of six well-known self-report measures of PE (Commitment to Exercise Scale, Compulsive Exercise Test, Exercise Addiction Inventory, Exercise Dependence Questionnaire, Exercise Dependence Scale, and Obligatory Exercise Questionnaire). *Methods:* Pooled effect sizes were computed using a random-effect model employing a restricted maximum likelihood estimation method. Univariable and multivariable meta-regressions analyses were employed for testing moderator variables. *Results:* Data retrieved from 255 studies (741 independent samples, $N = 254{,}174$) identified three main groups of findings: (i) pooled alpha values that, ranging from 0.768 to 0.930 for global scores and from 0.615 to 0.907 for subscale scores, were found to be sensitive to sociodemographic and methodological characteristics; (ii) reliability induction rates of 47.58%; and (iii) the virtually non-existent testing of the assumptions required for the proper applicability of alpha. Data unavailability prevented the provision of summarized reliability estimates in terms of temporal stability. *Discussion:* These findings highlight the need to improve reliability reporting of the scores of self-reported instruments of PE in primary studies. This implies providing both prior justification for the appropriateness of the index employed and reliability data for all the subpopulation of interest. The values presented could be used as a reference both for comparisons with those obtained in future primary studies and for correcting measurement-related artefacts in quantitative meta-analytic research concerning PE.

KEYWORDS

internal consistency, alpha, psychometric properties, morbid exercise, exercise dependence

INTRODUCTION

Promotion of regular physical activity has been proposed as a comprehensive and valid strategy to reduce cardiovascular risk (Ding et al., 2016). One of the domains in which physical activity is more frequently undertaken is leisure time, in particular, throughout recreational participation in sports activities or by engaging in exercise conditioning/training (Bull et al., 2020). However, a small proportion of the population may develop a potentially dysfunctional pattern of exercise behaviour (Marques et al., 2019). This is a complex and multifaceted phenomenon that, irrespective of the different umbrella terms used to refer to it (e.g., problematic exercise; Scharmer, Gorrell, Schaumberg, & Anderson, 2020; or morbid exercise behaviour; Szabo, Demetrovics, & Griffiths, 2018) implies losing control over exercise behaviour to the point of experiencing harm at a physical level (e.g., injuries or immune

[†]AP and MAI contributed equally to this work and should be considered co-first authors.

*Corresponding author.
E-mail: a.paterna@ual.es

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

problems), psychological level (e.g., altered mood states or inability to concentrate), or social level (e.g., loss of social relationships or job) (Juwono & Szabo, 2021; Szabo et al., 2018).

Existing research on the phenomenon – hereafter referred to as 'problematic exercise' (PE) – has been mainly approached using quantitative techniques and, more specifically, self-report instruments (Marques et al., 2019; Szabo, Griffiths, deLa Vega Marcos, Mervó, & Demetrovics, 2015). To date, much research has been devoted to examining the psychometric properties of scores obtained from translations of the original English versions of such instruments in non-English speaking countries from Europe (Mónok et al., 2012; Sauchelli et al., 2016; Sicilia, Alías-García, Ferriz, & Moreno-Murcia, 2013; Zeeck et al., 2017), South America (Alchieri et al., 2015; Sicilia et al., 2017), and Asia (Li, Nie, & Ren, 2016; Shin & You, 2015). However, much less effort has been spent on examining the psychometric properties of these PE scores among specific populations (e.g., in terms of their clinical condition [Formby, Watson, Hilyard, Martin, & Egan, 2014] or the exercise modality practised [Lichtenstein & Jensen, 2016]), as well as whether these properties can be generalized across different countries or languages (Griffiths et al., 2015). This is an important limitation in the case of a psychometric property that, such as reliability (i.e., measurement precision), is highly dependent on both the test application conditions and the characteristics of the sample under consideration (Slaney, 2017). A main practical implication of the extant literature concerns cross-group comparisons, because unequal reliability between groups can lead to wrong conclusions when comparing their respective scores (Graham & Unterschute, 2015). This is a matter of relevance in PE research because sample characteristics (e.g., exercise modality practised or being at-risk of an eating disorder) are frequently used for comparison purposes (Di Lodovico, Poulnais, & Gorwood, 2019; Trott et al., 2020). Having a comprehensive understanding of the effect of the sample and application characteristics on the score reliability of self-report instruments assessing PE is likely to contribute to advancing the science in this field. For example, this knowledge may assist practitioners and researchers in choosing an assessment tool capable of producing reliable scores across a range of circumstances. However, there is no summarized evidence on the reliability of scores derived from self-report instruments assessing PE across populations and application conditions.

Reliability Generalization (RG) meta-analysis provides cumulative evidence on elements contributing to the variability of test score reliability across studies (Vacha-Haase, Kogan, & Thompson, 2000; Vacha-Haase, Henson, & Caruso, 2002). Despite many reliability indices being available (Cho, 2016), it is often the case that RG meta-analysis only presents information concerning Cronbach's alpha coefficients (e.g., Graham & Unterschute, 2015; Vicent, Rubio-Aparicio, Sánchez-Meca, & Gonzálvez, 2019). This is due to an overwhelming use of alpha in primary studies (Hoekstra, Vugteveen, Warrens, & Kruyen, 2019). However, it has been suggested that this prevalent use of alpha is more due to compliance reasons such as it being perceived as a common and required practice (Hoekstra et al., 2019) rather than to its superiority over other reliability indexes or, as it would be methodologically sound, its adequacy according to the nature of the data (Cho, 2016). Indeed, the fact that alpha functions as an unbiased reliability estimator is dependent on the fulfilment of three main assumptions: (i) the unidimensionality of the test, (ii) the equality of the factor loadings of the items (i.e., tau-equivalence; if not met, alpha will underestimate reliability), and (iii) the independency of the error terms of the items (if not met, alpha will overestimate reliability) (Cho & Kim, 2015).

Based on these considerations, it follows that providing evidence on whether reported alpha values have been obtained after testing the assumptions required for the unbiased use of such a coefficient may be of interest from the perspective of RG meta-analysis. Similar ways of proceeding are common in RG meta-analysis (e.g., Graham & Unterschute, 2015; Vicent et al., 2019) with regard to another questionable reporting practice that may also influence the scope of the results, namely, *reliability induction* (i.e., the fact of not reporting reliability estimates for the data at hand; Vacha-Haase et al., 2000). Moreover, almost no attention has been paid to date in RG meta-analysis to alpha reporting practices in terms of their application assumptions (Vacha-Haase & Thompson, 2011). In view of these considerations, it is reasonable to suggest that examining both the rate of reliability induction and the extent to which the assumptions underlying the unbiased performance of alpha may lead to a more accurate and comprehensive interpretation of the results provided in RG meta-analysis.

Within this context, the present RG meta-analysis addresses three objectives concerning several widely used instruments proposed in the self-reported assessment of PE. More specifically, these are to (i) estimate the average reliability of the test scores under consideration; (ii) examine the sociodemographic and methodological characteristics that may affect the reliability estimates of the test scores of interest; and (iii) examine the reliability reporting practices of studies employing these instruments. The latter will be done (a) by examining the reliability induction rates; and (b) in view of the very likely possibility that alpha will be the most frequently reported index (Cho, 2016), by examining the extent to which the assumptions for unbiased estimates of such coefficient are tested and met.

METHOD

The systematic review and meta-analysis was conducted in accordance with the checklist from Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) (Moher, Liberati, Tetzlaff, & Altman, 2009) and was registered on PROSPERO (CRD42021237100) (see Supplementary material A).

Locating studies

Electronic bibliographic databases MEDLINE, PsycINFO, Web of Science, Current Contents Connect, SciELO, and

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Dissertations & Theses Global were searched for eligible studies from inception to January 30, 2020 (see Supplementary material B for the full search strategy). No geographical or cultural restrictions were applied. Reference lists of all retrieved studies were hand-searched to identify further potentially eligible studies.

The references of the retrieved studies were managed in EndnoteX9. Studies were independently selected by two of the authors in two stages by examining (a) their titles and abstracts, and (b) their full-texts. Disagreements were discussed and resolved on a consensual basis with the assistance of a third author if needed.

Eligibility criteria

The review collated data from studies employing the most widely used self-report instruments for the assessment of symptoms of PE (i.e., exercising to the point of losing the control over such a behaviour, so that it may leads to physical, psychological, or social damage; Szabo et al., 2018). According to the findings from previous reviews conducted in the field of PE (e.g., Alcaraz-Ibáñez, Paterna, Sicilia, & Griffiths, 2020, 2021), the following six key instruments were considered eligible: *Commitment to Exercise Scale* (CES), that assesses the extent to which (i) individuals' well-being are influenced by exercising, (ii) adherence to exercise is maintained in the face of adverse conditions, and (iii) exercise regimen interferes with social commitments (Davis, Brewer, & Ratusny, 1993); *Compulsive Exercise Test* (CET), which assesses the primary factors operating in the maintenance of excessive exercise within the eating disorders domain (Taranis, Touyz, & Meyer, 2011); *Exercise Addiction Inventory* (EAI), which assesses six common criteria proposed for behavioural addictions (Terry, Szabo, & Griffiths, 2004); *Exercise Dependence Questionnaire* (EDQ), which assesses elements employed in traditional models of addiction and both psychologically-related and socially-related consequences of exercise behaviour (Ogden, Veale, & Summers, 1997); *Exercise Dependence Scale* (EDS-21), which assesses seven criteria adapted from substance abuse defined in the *Diagnostic and Statistical Manual for Mental Disorders* (American Psychiatric Association, 1994) applied to the exercise domain (Downs, Hausenblas, & Nigg, 2004); and *Obligatory Exercise Questionnaire* (OEQ), which assesses the subjective need to engage in repetitive exercise behaviours (Pasman & Thompson, 1988). The eligibility of these instruments was also supported by the findings derived from a search on *Google Scholar* performed by the present authors for all the 17 measures previously identified within the field (Sicilia, Paterna, Alcaraz-Ibáñez, & Griffiths, 2021). In particular, these instruments were shown to be the ones with the highest number of citations (see Supplementary material C).

Inclusion criteria. Studies were considered eligible if the following criteria were met: (a) at least one of the following six self-report instrument of PE was used: CES, CET, EAI, EDQ, EDS-21, OEQ; (b) they were written in English, Spanish, French, or Portuguese (the working languages of the review team); and (c) some estimate of reliability was provided (e.g., Cronbach's alpha [α], intra-class correlation index [ICC], or Pearson's correlation index [r]).

Exclusion criteria. Studies were excluded on the basis of the following criteria: (a) only composite scores comprising two or more instruments assessing PE were provided so that individual scores were not available; (b) specific items were excluded when obtaining global scores of PE and sub-domains scores were not available; (c) specific items were excluded when obtaining sub-scale scores of PE; (d) the scores of PE were obtained using a partially/completely altered factorial structure from the one originally proposed for the instrument; and (e) studies with less than 30 participants. The first four exclusion criteria were implemented with the aim of fulfilling one of the main assumptions of meta-analytic research (i.e., the application of a similar statistical configuration) (Lipsey & Wilson, 2001). The final exclusion criterion was implemented on the basis of the increased sampling error and variations in the assessment of heterogeneity likely introduced by studies with small sample sizes (Lin, 2018).

Coding procedure. A coding frame was developed taking into account the common features of the studies retrieved in a preliminary search. After being pilot-tested, the coding sheet was used by two of the present authors when extracting the relevant data from the retrieved studies (see Supplementary material D). Disagreements between the reviewers were discussed and resolved on a consensual basis with the assistance of a third author if necessary. The following coding categories were considered: (i) citation and year of publication; (ii) sample size; (iii) exercise modality; (iv) eating disorders (EDs); (v) report of leisure time exercise; (vi) regular exercisers; (vii) region (geographic location); (viii) test version; (ix) type of survey; (x) publication status; (xi) study design; (xii) mean and standard deviation (*SD*) of test scores; (xiii) mean and *SD* of age; (xiv) % of Whites; (xv) % of females; and (xvi) PE measure. These coded features were considered for descriptive purposes and – where appropriate – as potential moderator variables (Rosenthal, 1995).

Statistical analysis

Effect size calculations. Cronbach's alpha (α) was employed as the effect size index. In order to normalize their distributions and stabilize their variances, the reliability coefficients were (α)-to-($\tilde{\alpha}$) transformed by applying the formula proposed by Bonett (2002) before conducting the statistical analyses. In the interest of facilitating interpretation of the results, effect sizes and their 95% confidence intervals (CIs) were subsequently ($\tilde{\alpha}$)-to-(α) transformed (Sánchez-Meca, López-López, & López-Pina, 2013).

Due to the expected heterogeneity between studies in terms of participants' characteristics, and assuming that variations in the distribution and sampling errors of effect sizes may contribute to explain differences between them,

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

the pooled effect sizes were computed using a random-effect model using an estimation method robust to the normality (i.e., restricted maximum likelihood, REML) (Pigott, 2012). The I^2 statistic was used to assess statistical heterogeneity, with values of 25%, 50%, and 75% indicating low, moderate, and high heterogeneity, respectively (Higgins, Thompson, Deeks, & Altman, 2003). The robustness of the summarized estimates was examined through sensitivity analyses (i.e., by conducting systematic reanalysis while removing studies one at a time). Results from sensitivity analyses (see Supplementary material E) were considered meaningful when corrected estimates were beyond the 95% CI of the original ones.

Consistent with previous RG meta-analyses (Rubio-Aparicio, Badenes-Ribera, Sánchez-Meca, Fabris, & Longobardi, 2020), moderator analyses for categorical and continuous variables were conducted provided that at least 15 effect sizes were available. Meta-regression analyses employed for testing moderator variables were conducted in two stages. Firstly, by employing univariable models (i.e., considering each potential moderator in isolation). Secondly, by employing multivariable models in which all significant moderators identified in the first stage were simultaneously introduced. For a better control of Type I error rate, meta-regressions were conducted using the method proposed by Knapp and Hartung (2003). Given constraints due to available sample size, non-significant categorical predictors were sequentially dropped from the full starting multivariable models in order to obtain the most parsimonious and accurate representation of the data. The tenability of the reduced vs. the full model was judged through a likelihood ratio test (LRT). Explained variance by the moderators was quantified as a percentage and expressed by R^2. Provided that at least 10 effect sizes were available (Page, Higgins, & Sterne, 2019), publication bias was examined by visual inspection of funnel plot symmetry, Egger's test, and the 'trim and fill' procedure (see Supplementary material F). The statistical analyses described in this section were conducted in R using the *metafor* package.

RESULTS

Selection of studies

A total of 3,852 studies were identified from multiple database searches. The study selection procedure was conducted in two stages. Firstly, the eligibility criteria were applied to the studies considered for full text assessment (see Fig. 1). Secondly, the report of reliability indices was examined. Despite the intention of including data on temporal stability (e.g., Pearson's correlation), the number of studies reporting this information was too low to meta-analytical techniques to be applied (i.e., EAI, Griffiths, Szabo, & Terry, 2005; Li et al., 2016; EDQ, Kern & Baudin, 2011; EDS-21, Downs et al., 2004; Kern, 2007). As a result of this process, 255

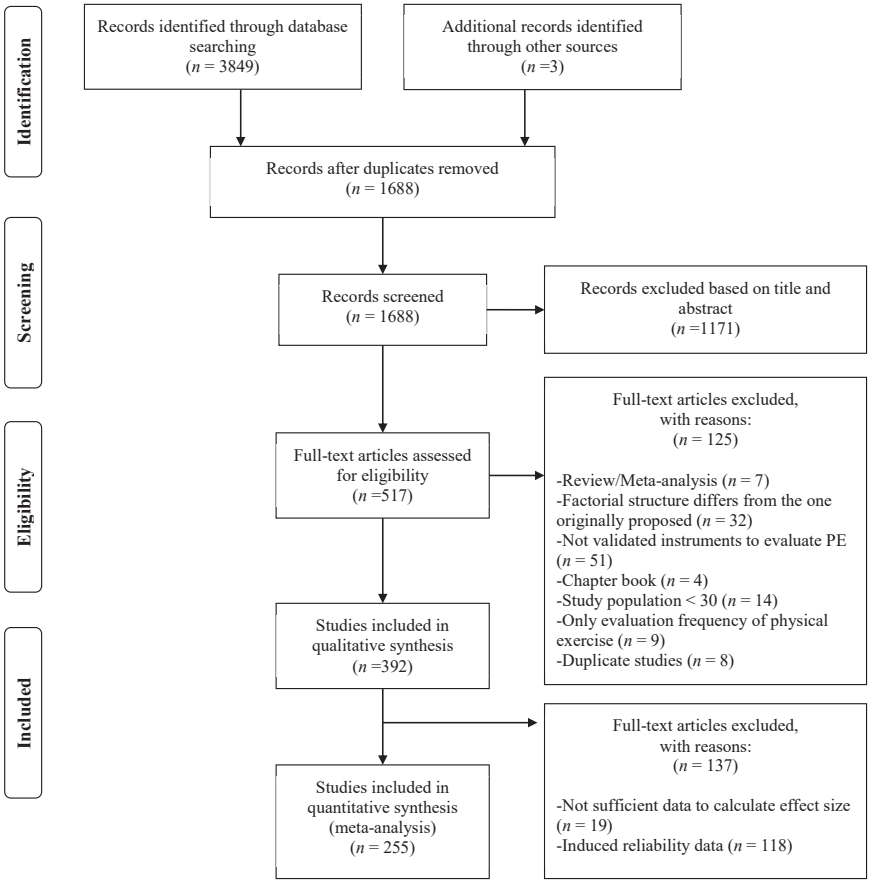

Fig. 1. PRISMA flow diagram of study selection

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

studies that reported reliability in terms of alpha coefficient were included in the RG meta-analysis. The study characteristics and their corresponding effect sizes were grouped according to PE measures. Consequently, 741 effect sizes from 255 studies (N = 254,174) were examined in 27 different meta-analyses (see Table 1).

Commitment to Exercise Scale

Two different response procedures were employed in the retrieved studies using the CES (i.e., Likert scales or visual analogue scales [VAS]). Given that the homogeneity of statistical configuration across studies is one of the main underlying assumptions of meta-analysis (Lipsey & Wilson, 2001), the scores of the CES (Likert) and CES (VAS) were examined independently.

Commitment to Exercise Scale using Likert scales. The analysis examining alpha estimates for the global score on the CES-Likert (see Forest plot in Supplementary material G) included 10 effect sizes from nine studies involving a total (N_{total}) of 2,891 participants. Results from the random effects model showed a pooled alpha estimate of 0.872 (P < 0.001; 95% CI = 0.853 to 0.889, I^2 = 81.29). Since the number of effect sizes retrieved was <15, moderation analyses were not conducted.

Commitment to Exercise Scale using visual analogue scales. The analysis examining alpha estimates for the global score on the CES-VAS (see Forest plot in Supplementary material G) included 30 effect sizes from 23 studies (N_{total} = 6,529). Results from the random effects model showed a pooled alpha estimate of 0.842 (P < 0.001; 95% CI = 0.816 to 0.864, I^2 = 93.60). Results from the univariate meta-regression analysis for categorical variables (see Table 2) identified the following significant moderators: (a) eating disorders (omnibus-test [2, 27] = 7.451; P = 0.003; R^2 = 33.59); (b) report of leisure time exercise (omnibus-test [1, 28] = 6.096; P = 0.020; R^2 = 16.93); (c) region (omnibus-test [4, 25] = 3.850; P = 0.014; R^2 = 28.21); (d) test version (omnibus-test [1, 28] = 5.621; P = 0.025; R^2 = 13.48); and (e) type of survey (omnibus-test [3, 26] = 3.990; P = 0.018; R^2 = 25.87). Results from the univariate meta-regression analysis for continuous variables (see Table 3) did not identify any significant moderator. Results from the multivariate meta-regression analysis showed that eating disorders, report of leisure time exercise, test

Table 1. Alpha estimates for the scores of instruments assessing problematic exercise

				Meta-analysis report					
						95% CI			
Measure (Subscale)	Items	Range	Original α	k	$\bar{\alpha}$	Lo	Up	Q	I^2
CES-Likert	8	1–10	N.R.	10	0.872	0.853	0.889	47.856	81.29
CES-VAS	8	0–155	0.770	30	0.842	0.816	0.864	401.834	93.60
CET	24	0–5	0.850, 0.830	48	0.880	0.868	0.891	450.903	92.99
CET (Avoidance)	8	0–5	0.880, 0.880	27	0.907	0.888	0.923	601.459	95.98
CET (Weight control)	5	0–5	0.860, 0.850	21	0.817	0.787	0.842	175.464	90.72
CET (Mood improvement)	5	0–5	0.750, 0.720	20	0.801	0.779	0.836	187.271	90.71
CET (Lack of enjoyment)	3	0–5	0.840, 0.820	18	0.777	0.739	0.810	155.376	88.08
CET (Rigidity)	3	0–5	0.730, 0.820	23	0.771	0.748	0.793	92.048	76.36
EAI	6	1–5	0.840	42	0.768	0.739	0.794	2,258.405	97.27
EDQ	29	1–7	0.843	12	0.862	0.842	0.879	70.101	84.26
EDQ (Interference)	5	1–7	0.814	7	0.743	0.676	0.795	49.772	86.57
EDQ (Positive reward)	4	1–7	0.795	6	0.789	0.688	0.857	75.291	94.89
EDQ (Withdrawal)	4	1–7	0.799	7	0.772	0.719	0.815	35.498	82.67
EDQ (Weight control)	4	1–7	0.781	6	0.721	0.670	0.764	18.925	71.44
EDQ (Insight into problem)	4	1–7	0.756	6	0.690	0.625	0.744	24.952	78.19
EDQ (Social reasons)	3	1–7	0.755	6	0.615	0.489	0.710	53.587	88.86
EDQ (Health reasons)	3	1–7	0.701	6	0.774	0.692	0.834	56.772	90.64
EDQ (Stereotyped behaviour)	2	1–7	0.516	6	0.670	0.561	0.736	25.358	81.63
EDS-21	21	1–6	N.R.	90	0.930	0.923	0.937	3,906.857	97.76
EDS-21 (Tolerance)	3	1–6	0.780, 0.780	43	0.857	0.840	0.872	673.810	93.94
EDS-21 (Withdrawal)	3	1–6	0.930, 0.900	42	0.828	0.809	0.845	603.767	92.86
EDS-21 (Intention effects)	3	1–6	0.920, 0.890	43	0.881	0.865	0.895	906.013	95.48
EDS-21 (Lack of control)	3	1–6	0.820, 0.820	44	0.823	0.803	0.841	691.373	93.80
EDS-21 (Time)	3	1–6	0.880, 0.860	43	0.848	0.833	0.862	549.977	91.82
EDS-21 (Reduction in other activities)	3	1–6	0.670, 0.750	53	0.704	0.675	0.730	692.150	92.53
EDS-21 (Continuance)	3	1–6	0.890, 0.900	43	0.834	0.816	0.851	611.499	93.26
OEQ 20	20	1–4	0.960	38	0.870	0.853	0.885	556.527	94.43

Note. α = alpha value(s) reported in the original validation studies; $\bar{\alpha}$ = Estimated effect size (corrected coefficient alpha); CI = Confidence interval; Lo = Lower; Up = Upper; N.R. = non-reported; CES-VAS = Commitment Exercise Scale; CET = Compulsive Exercise Test; EAI = Exercise Addiction Inventory; EDS-21 = Exercise Dependence Scale-21; OEQ = Obligatory Exercise Questionnaire

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 2. Results of univariable meta-regression analyses for categorical variables (global scores)

	CES-VAS					CET					EAI					EDS-21					OEQ				
			95% CI					95% CI					95% CI					95% CI					95% CI		
Subgroups	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2
Exercise modality																									
Unknown (RC)	19	0.843	0.805	0.874	95.23	38	0.887	0.876	0.897	92.07	16	0.783	0.740	0.819	97.46	39	0.946	0.937	0.953	96.57	25	0.867	0.849	0.883	90.02
Unclear	2	0.800	0.746	0.842	71.77	8	0.843	0.790	0.883	90.36	8	0.769	0.710	0.815	94.69	18	0.920	0.896	0.939	98.60	6	0.863	0.829	0.890	89.39
Power disciplines	–	–	–	–	–	–	–	–	–	–	2	0.733	0.705	0.759	0.00	3	0.918	0.889	0.939	76.70	–	–	–	–	–
Non-endurance	1	0.770	0.642	0.852	–	1	0.850	0.805	0.885	–	2	0.708	0.523	0.821	85.27	4	0.917	0.838	0.957	98.09	–	–	–	–	–
Multiple sports	7	0.871	0.860	0.855	13.50	–	–	–	–	–	6	0.796	0.647	0.882	98.49	9	0.925	0.903	0.941	95.89	2	0.954	0.943	0.962	34.26
Fitness and health	1	0.770	0.726	0.807	–	–	–	–	–	–	4	0.720	0.661	0.770	92.09	7	0.924	0.903	0.942	95.89	–	–	–	–	–
Endurance	–	–	–	–	–	1	0.850	0.822	0.874	–	4	0.764	0.581	0.867	96.99	10	0.913	0.899	0.925	93.52	5	0.837	0.805	0.864	89.47
Eating disorders																									
Unknown (RC)	23	0.824	0.799	0.846	90.69	32	0.863	0.850	0.875	91.31	39	0.770	0.739	0.798	97.35	69	0.930	0.921	0.938	0.97.90	37	0.869	0.852	0.884	94.61
At-risk	–	–	–	–	–	2	0.874	0.801	0.921	91.04	–	–	–	–	–	4	0.963	0.945	0.974	78.37	–	–	–	–	–
Not at-risk	–	–	–	–	–	1	0.900	0.861	0.928	–	–	–	–	–	–	5	0.917	0.894	0.934	84.06	–	–	–	–	–
Mixed	6	0.857	0.827	0.911	90.33	3	0.902	0.853	0.934	84.62	3	0.745	0.695	0.785	80.49	11	0.922	0.896	0.942	98.00	1	0.900	0.863	0.927	–
Clinical	1	0.950	0.930	0.964	–	10	0.927	0.920	0.934	0.01	–	–	–	–	–	1	0.930	0.913	0.944	–	–	–	–	–	–
Report of LTE																									
No (RC)	13	0.808	0.766	0.842	92.11	31	0.887	0.875	0.899	92.81	20	0.786	0.740	0.824	98.06	42	0.931	0.920	0.941	98.30	23	0.868	0.849	0.884	90.84
Yes	17	0.864	0.836	0.887	92.50	17	0.864	0.839	0.885	90.77	22	0.751	0.714	0.782	95.19	48	0.929	0.919	0.939	96.95	15	0.873	0.839	0.900	96.96
Regular exercisers																									
Unknown (RC)	21	0.838	0.803	0.867	95.06	40	0.886	0.874	0.896	92.08	28	0.784	0.751	0.813	97.57	55	0.934	0.925	0.942	97.93	10	0.883	0.866	0.898	93.10
Yes	9	0.851	0.821	0.877	83.40	8	0.842	0.792	0.881	89.36	13	0.729	0.673	0.775	94.91	34	0.922	0.909	0.934	97.24	28	0.827	0.800	0.850	88.61
Region																									
Unknown (RC)	2	0.815	0.719	0.878	88.12	7	0.905	0.872	0.929	92.18	8	0.790	0.756	0.819	90.96	16	0.941	0.928	0.951	94.85	13	0.891	0.863	0.914	94.59
South America	18	0.820	0.789	0.847	91.29	–	–	–	–	–	2	0.640	0.527	0.726	82.69	4	0.880	0.856	0.900	75.04	–	–	–	–	–
Oceania	–	–	–	–	–	7	0.890	0.861	0.913	84.45	2	0.704	0.649	0.750	0.00	1	0.930	0.911	0.945	–	7	0.854	0.804	0.892	94.19
North America	7	0.875	0.832	0.907	90.63	12	0.864	0.841	0.884	87.56	5	0.837	0.795	0.871	87.39	29	0.938	0.927	0.947	97.52	15	0.858	0.833	0.879	93.31
Mixed	1	0.950	0.930	0.964	–	4	0.890	0.795	0.941	93.92	–	–	–	–	–	4	0.938	0.889	0.965	98.47	–	–	–	–	–
Europe	2	0.834	0.803	0.862	67.28	18	0.875	0.861	0.887	89.97	24	0.745	0.706	0.779	96.82	34	0.920	0.905	0.932	97.69	3	0.855	0.787	0.901	85.50
Asia	–	–	–	–	–	–	–	–	–	–	1	0.920	0.914	0.926	–	2	0.942	0.525	0.993	99.67	–	–	–	–	–
Test version																									
Original (RC)	11	0.874	0.834	0.904	93.81	44	0.877	0.865	0.888	92.89	21	0.795	0.765	0.821	95.19	58	0.936	0.928	0.944	97.38	38	0.870	0.853	0.855	94.43
Linguistically adapted	19	0.820	0.791	0.846	81.28	4	0.905	0.857	0.937	85.16	21	0.739	0.687	0.782	97.73	32	0.918	0.902	0.930	97.96	–	–	–	–	–
Type of survey																									
Unknown (RC)	15	0.871	0.844	0.893	91.89	36	0.884	0.871	0.896	92.54	25	0.786	0.751	0.816	97.45	46	0.929	0.919	0.938	97.18	24	0.863	0.848	0.878	90.19
Paper-pencil	2	0.708	0.592	0.791	55.12	9	0.878	0.859	0.895	54.25	5	0.727	0.634	0.796	96.08	26	0.928	0.914	0.940	97.66	8	0.867	0.822	0.901	94.57
On-line	12	0.820	0.780	0.852	91.66	3	0.861	0.822	0.893	94.61	11	0.748	0.676	0.803	96.12	16	0.930	0.905	0.948	98.26	4	0.881	0.788	0.933	96.45
Both	1	0.770	0.714	0.815	–	–	–	–	–	–	1	0.710	0.669	0.746	–	2	0.968	0.923	0.987	97.81	2	0.920	0.687	0.979	98.31
Publication status																									
Published (RC)	24	0.830	0.800	0.855	93.76	41	0.881	0.868	0.893	94.10	38	0.759	0.728	0.785	97.25	79	0.931	0.923	0.938	97.95	30	0.870	0.849	0.888	95.59
Unpublished	6	0.882	0.846	0.910	85.65	7	0.870	0.849	0.899	70.78	4	0.843	0.795	0.879	89.55	11	0.920	0.896	0.939	94.01	8	0.870	0.848	0.889	80.17
Study design																									
Psychometric (RC)	5	0.859	0.767	0.914	96.57	8	0.848	0.797	0.887	93.70	12	0.784	0.714	0.837	97.74	9	0.933	0.902	0.954	98.69	6	0.878	0.805	0.924	96.13
Applied	25	0.838	0.811	0.861	92.81	40	0.885	0.873	0.895	91.62	30	0.761	0.730	0.789	96.68	81	0.930	0.922	0.937	97.60	32	0.868	0.852	0.883	93.68

Note. $\bar{\alpha}$ = Corrected coefficient alpha; CI = Confidence interval; Lo = Lower; Up = Upper; RC = Reference category; LTE = Leisure time exercise; CES-VAS = Commitment Exercise Scale; CET = Compulsive Exercise Test; EAI = Exercise Addiction Inventory; EDS-21 = Exercise Dependence Scale-21; OEQ = Obligatory Exercise Questionnaire.

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 3. Results of univariable meta-regression analyses for continuous variables (global scores)

Moderators	CES-VAS					CET					EAI					EDS-21					OEQ				
	K	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2
Mean of test score	29	0.000	0.001	0.971	0.00	40	−0.025	0.998	0.324	0.36	31	**−0.250**	4.993	0.033	13.08	68	**−0.243**	4.895	0.030	6.37	33	−0.047	0.150	0.701	0.00
SD of test score	29	−0.001	0.004	0.948	0.00	39	−0.301	2.690	0.109	5.34	31	0.398	1.304	0.263	2.18	66	**0.618**	5.836	0.019	6.88	33	−0.166	0.402	0.531	0.00
Mean age	26	0.011	1.396	0.249	1.41	43	−0.008	0.785	0.381	0.00	40	−0.003	0.076	0.785	0.00	78	−0.001	0.012	0.913	0.00	32	−0.005	0.233	0.633	0.00
SD age	26	0.029	0.734	0.400	0.00	42	0.008	0.270	0.606	0.00	37	0.000	0.001	0.982	0.00	76	−0.007	0.189	0.666	0.00	31	0.004	0.034	0.855	0.00
% of Whites	14	0.001	0.138	0.717	0.00	17	−0.003	0.133	0.720	0.00	7*	−0.003	0.279	0.620	0.00	38	−0.003	2.379	0.132	4.50	18	0.003	0.155	0.699	0.00
% of Females	30	0.001	0.273	0.605	0.00	47	0.002	0.948	0.336	0.00	40	0.001	0.167	0.685	0.00	89	0.002	1.544	0.217	0.44	34	−0.001	0.156	0.695	0.00
Year of publication	30	0.002	0.040	0.843	0.00	48	0.027	3.821	0.057	5.95	42	−0.005	0.091	0.765	0.00	90	0.008	0.442	0.508	0.00	38	−0.012	1.688	0.202	1.40

Note. β_1 = estimated regression coefficient; R^2 = Explained variance; *F* = Omnibus test; RC = Reference category; CES-VAS = Commitment Exercise Scale; CET = Compulsive Exercise Test; EAI = Exercise Addiction Inventory; EDS-21 = Exercise Dependence Scale-21; OEQ = Obligatory Exercise Questionnaire. Statistically significant effects ($P < 0.05$) appear highlighted in bold.
*Correspond to *K*< 10 and should therefore not be interpreted (Fu et al., 2011)

version, and type of survey explained together 68.73% of variance in pooled alpha estimate (see Table 4).

Compulsive Exercise Test

The analysis examining the alpha estimates for the global score on the CET (see Forest plot in Supplementary material G) included 48 effect sizes from 42 studies (N_{total} = 14,675). Results from the random effects model showed a pooled alpha estimate of 0.880 ($P < 0.001$; 95% CI = 0.868 to 0.891, I^2 = 92.99). Results from the univariate meta-regression analysis for continuous categorical variables (see Table 2) identified the following significant moderators: (a) eating disorders (omnibus-test [4, 43] = 8.737; $P < 0.001$; R^2 = 43.48); (b) regular exercisers (omnibus-test [1, 46] = 6.482; P = 0.014; R^2 = 11.63); and (c) study design (omnibus-test [1, 46] = 4.723; P = 0.035; R^2 = 7.47). Results from the univariate meta-regression analysis for continuous variables (see Table 3) did not identify any significant moderators. Results from the multivariate meta-regression analysis showed that eating disorders and regular exercisers together explained 57.55% of variance in pooled alpha estimate (see Table 4).

Compulsive Exercise Test subscales. The analysis examining the alpha estimates for the subscale scores on the CET (see Forest plot in Supplementary material G) included 109 effect sizes. Considering the different subscales, the effect sizes available ranged from 18 (lack of exercise enjoyment, N_{total} = 4,302) to 27 (avoidance, N_{total} = 6,888). Findings from the random effects model showed pooled alpha estimates ranging from 0.771 (exercise rigidity; $P < 0.001$; 95% CI = 0.748 to 0.793, I^2 = 76.36) to 0.907 (avoidance; $P < 0.001$; 95% CI = 0.888 to 0.923, I^2 = 95.98). Results from the univariate meta-regression analysis for categorical variables (see Table 5) identified the following significant moderators: (a) avoidance: exercise modality (omnibus-test [3, 23] = 3.222, P = 0.041, R^2 = 20.10), eating disorders (omnibus-test [2, 24] = 33.606, $P < 0.001$, R^2 = 75.04), report of leisure time exercise (omnibus-test [1, 25] = 5.833, P = 0.023, R^2 = 16.40), regular exercisers (omnibus-test [1, 25] = 5.429, P = 0.028, R^2 = 14.24), and test version (omnibus-test [1, 25] = 5.455, P = 0.028, R^2 = 16.21); (b) weight control: (type of survey, omnibus-test [2, 18] = 5.322, P = 0.015, R^2 = 35.20); and (c) exercise rigidity: region (omnibus-test [4, 18] = 4.535, P = 0.010, R^2 = 41.51), and study design (omnibus-test [1, 21] = 5.334, P = 0.031, R^2 = 17.36). The results of the univariate meta-regression analysis for continuous variables (see Table 6) identified the following significant moderators: (a) mean of test score (avoidance and mood improvement); (b) age (avoidance); (c) *SD* of age (avoidance and mood improvement); (d) year of publication (avoidance and weight control; and percentage of females (weight control and exercise rigidity). However, the results of the multivariate meta-regression analysis (see Table 7) supported the moderating role of the variables under examination just for the following cases: (a) eating disorders and *SD* of test score (avoidance); (b) percentage of females and year of publication (weight control); (c) *SD* of test score and

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 4. Results of multivariable meta-regression analyses (global scores)

Moderators	*K*	β_0	β_1	*SE*	*F*	*P*	R^2
CES-VAS	30				51.844	<0.001	68.73
		1.779	–	0.117			
Eating disorders (Mixed)			**0.281**	0.133			
Eating disorders (Clinical)			**0.931**	0.298			
Report of LTE (Yes)			**0.286**	0.117			
Test version (linguistically adapted)			**–0.268**	0.125			
Type of survey (Paper-pencil)			**–0.476**	0.222			
Type of survey (Online)			0.110	0.142			
Type of survey (Both)			**–0.595**	0.267			
CET	48				49.917	<0.001	57.55
		2.039	–	0.043			
Eating disorders (At risk)			0.041	0.163			
Eating disorders (Not at risk)			0.263	0.264			
Eating disorders (Mixed)			0.255	0.147			
Eating disorders (Clinical)			**0.564**	0.093			
Regular exercisers (Yes)			**–0.257**	0.094			
EAI	31				38.281	<0.001	59.22
		2.251	–	0.282			
Region (South America)			**–0.334**	0.168			
Region (Oceania)			**–0.337**	0.166			
Region (North America)			0.023	0.145			
Region (Europe)			–0.139	0.102			
Test version (linguistically adapted)			**–0.248**	0.091			
Mean total score*			**–0.223**	0.094			
EDS-21	66				37.410	<0.001	38.02
		2.938	–	0.323			
Exercise modality (Unclear)			**–0.380**	0.137			
Exercise modality (Power disciplines)			–0.437	0.287			
Exercise modality (Non-endurance)			**–0.684**	0.247			
Exercise modality (Multiple sports)			**–0.382**	0.169			
Exercise modality (Fitness and health)			**–0.645**	0.214			
Exercise modality (Endurance)			**–0.488**	0.159			
Mean total score*			–0.078	0.106			
SD total score*			0.203	0.228			
OEQ	38				64.660	<0.001	68.55
		2.096	–	0.050			
Exercise modality (Unclear)			0.156	0.114			
Exercise modality (Multiple sports)			**0.997**	0.174			
Exercise modality (Endurance)			0.295	0.160			
Regular exercisers (Yes)			**–0.463**	0.124			
Publication status (Unpublished)			**–0.197**	0.093			

Note. β_0 = intercept/mean effect size; β_1 = estimated regression coefficient; R^2 = Explained variance; *F* = Omnibus test of moderators; CES-VAS = Commitment Exercise Scale; CET = Compulsive Exercise Test; EAI = Exercise Addiction Inventory; EDS-21 = Exercise Dependence Scale-21; OEQ = Obligatory Exercise Questionnaire; LTE = Leisure time exercise. The reference categories were: Unknown (Eating disorders, Exercise modality, and Region), Original version (Test version), and Published (Publication status). Statistically significant effects ($P < 0.05$) appear highlighted in bold.

* Continuous moderator.

SD of age (mood improvement); and (d) region and percentage of females (exercise rigidity). The amount of variance in pooled alpha estimates explained by the retained models in the multivariate meta-regression analyses ranged from 63.26% (weight control) to 86.08% (avoidance).

Exercise Addiction Inventory

The retrieved studies included multiple versions of the EAI. Since only one study reported alpha scores for the EAI-R (Szabo, Pinto, Griffiths, Kovácsik, & Demetrovics, 2019) (α = 0.90), this was excluded from the analyses. The analysis examining the alpha estimates for the global score on the EAI (see Forest plot in Supplementary material G) included 42 effect sizes from 40 studies (N_{total} = 26,565). Results from the random effects model showed a pooled alpha estimate of 0.768 ($P < 0.001$; 95% CI = 0.739 to 0.810, I^2 = 97.27). Results from the univariate meta-regression analysis for categorical variables (see Table 2) identified the following significant moderators: (a) region (omnibus-test

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 5. Results of univariable meta-regression analyses for categorical variables (subscale scores of the Compulsive Exercise Test)

	Avoidance					Weight control					Mood improvement					Lack of enjoyment					Exercise rigidity				
			95% CI					95% CI					95% CI					95% CI					95% CI		
Subgroups	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2
Exercise modality																									
Unknown (RC)	18	0.922	0.901	0.938	96.19	12	0.797	0.748	0.836	92.92	11	0.830	0.793	0.860	91.25	11	0.787	0.727	0.833	93.27	16	0.764	0.736	0.789	77.44
Unclear	6	0.857	0.827	0.880	68.90	6	0.864	0.846	0.879	22.90	6	0.796	0.728	0.846	85.18	6	0.758	0.731	0.783	0.01	6	0.800	0.756	0.836	62.75
Power disciplines	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Non-endurance	1	0.870	0.831	0.900	–	1	0.750	0.670	0.810	–	1	0.770	0.697	0.826	–	1	0.770	0.689	0.830	–	1	0.720	0.621	0.793	–
Multiple sports	2	0.890	0.843	0.924	88.14	2	0.818	0.798	0.936	0.00	2	0.736	0.670	0.789	67.36	–	–	–	–	–	–	–	–	–	–
Fitness and health	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Endurance	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Eating disorders																									
Unknown (RC)	15	0.876	0.856	0.893	90.17	12	0.818	0.778	0.851	92.22	12	0.806	0.775	0.832	84.76	12	0.770	0.726	0.807	88.23	15	0.764	0.732	0.791	82.02
At risk	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Not at risk	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Mixed	4	0.893	0.864	0.918	79.73	4	0.808	0.780	0.833	36.92	3	0.744	0.695	0.784	55.30	1	0.770	0.689	0.830	–	1	0.720	0.621	0.793	–
Clinical	8	0.953	0.947	0.959	44.96	5	0.828	0.730	0.890	91.46	5	0.849	0.769	0.901	90.32	5	0.800	0.698	0.867	87.42	7	0.798	0.767	0.825	34.89
Report of LTE																									
No (RC)	16	0.921	0.900	0.939	96.07	10	0.809	0.761	0.847	91.37	9	0.823	0.770	0.865	93.77	8	0.781	0.730	0.838	88.71	13	0.768	0.739	0.793	70.21
Yes	11	0.880	0.852	0.903	91.86	11	0.824	0.784	0.856	89.97	11	0.796	0.761	0.826	83.04	10	0.766	0.714	0.809	87.17	10	0.778	0.734	0.814	82.07
Regular exercisers																									
Unknown (RC)	19	0.919	0.898	0.935	96.59	13	0.804	0.769	0.834	89.48	12	0.822	0.782	0.855	93.01	11	0.797	0.747	0.837	90.97	16	0.766	0.740	0.790	74.34
Yes	8	0.873	0.844	0.897	83.62	8	0.834	0.783	0.873	89.44	8	0.788	0.738	0.829	82.36	7	0.741	0.689	0.785	69.09	7	0.783	0.730	0.826	76.72
Region																									
Unknown (RC)	6	0.935	0.914	0.951	88.06	3	0.789	0.734	0.833	76.58	3	0.827	0.763	0.874	87.53	3	0.764	0.596	0.862	94.83	6	0.755	0.707	0.795	64.68
South America	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Oceania	5	0.904	0.864	0.932	87.25	5	0.869	0.832	0.898	66.02	5	0.777	0.723	0.820	54.50	4	0.776	0.725	0.817	15.30	4	0.825	0.788	0.855	0.00
North America	2	0.896	0.819	0.940	92.29	2	0.774	0.619	0.866	85.93	1	0.850	0.816	0.878	–	1	0.770	0.713	0.816	–	1	0.800	0.750	0.840	–
Mixed	2	0.932	0.879	0.962	88.46	2	0.858	0.790	0.904	76.18	2	0.860	0.838	0.879	0.00	2	0.785	0.689	0.851	68.86	2	0.842	0.814	0.865	0.00
Europe	12	0.887	0.847	0.917	97.50	9	0.791	0.743	0.831	91.92	9	0.800	0.739	0.847	95.18	8	0.800	0.708	0.834	93.13	10	0.746	0.713	0.777	75.82
Asia	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Test version																									
Original (RC)	23	0.899	0.878	0.916	95.75	21	0.817	0.787	0.842	90.72	20	0.809	0.779	0.836	90.71	18	0.777	0.739	0.810	88.08	19	0.776	0.750	0.800	79.31
Linguistically adapted	4	0.943	0.920	0.960	86.80	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4	0.748	0.691	0.795	49.57
Type of survey																									
Unknown (RC)	17	0.914	0.889	0.934	97.04	17	0.832	0.807	0.853	84.49	13	0.809	0.774	0.840	87.15	11	0.778	0.727	0.820	89.17	18	0.769	0.744	0.791	68.54
Paper-pencil	5	0.895	0.840	0.931	93.83	1	0.620	0.536	0.689	–	3	0.854	0.728	0.922	96.73	4	0.770	0.644	0.852	94.86	3	0.746	0.593	0.842	91.87
On-line	5	0.892	0.857	0.918	85.85	3	0.767	0.678	0.831	90.90	4	0.778	0.754	0.800	17.01	3	0.783	0.735	0.823	–	2	0.807	0.776	0.834	19.73
Both	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Publication status																									
Published (RC)	23	0.912	0.891	0.928	96.45	17	0.816	0.781	0.846	92.43	16	0.821	0.787	0.849	91.73	14	0.766	0.719	0.805	89.12	19	0.776	0.750	0.798	76.70
Unpublished	4	0.873	0.842	0.898	69.71	4	0.819	0.766	0.859	75.08	4	0.761	0.714	0.800	50.14	4	0.815	758	0.859	74.17	4	0.752	0.673	0.812	74.92
Study design																									
Psychometric (RC)	11	0.904	0.866	0.931	97.23	10	0.828	0.791	0.857	89.21	10	0.817	0.763	0.858	94.15	8	0.780	0.726	0.823	85.34	8	0.802	0.769	0.831	70.24
Applied	16	0.909	0.887	0.927	94.64	11	0.806	0.757	0.846	91.20	10	0.805	0.770	0.831	81.90	10	0.775	0.717	0.822	89.44	15	0.753	0.724	0.780	72.36

Note: $\bar{\alpha}$ = Corrected coefficient alpha. CI = Confidence interval; Lo = Lower; Up = Upper; RC = Reference category. LTE = Leisure time exercise.

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 6. Results of univariable meta-regression analyses for continuous variables (subscale scores of the Compulsive Exercise Test)

Moderators	Avoidance					Weight control					Mood improvement					Lack of enjoyment					Exercise rigidity				
	K	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2
Mean total scores	20	0.314	2.627	0.122	8.29	18	−0.200	3.473	0.080	17.27	17	−0.159	0.696	0.417	0.00	15	−0.118	1.345	0.267	1.84	17	−0.087	0.840	0.374	0.00
SD total scores	20	**1.383**	41.712	<0.001	70.92	18	−0.454	0.739	0.403	0.00	17	**1.912**	30.996	<0.001	71.45	15	−0.072	0.057	0.814	0.00	17	0.371	1.675	0.215	2.51
Mean age	27	**0.038**	4.748	0.039	14.91	21	−0.024	0.013	0.081	8.75	20	0.026	4.357	0.051	14.75	18	0.021	2.182	0.159	7.12	23	0.007	0.330	0.572	0.00
SD age	27	**0.071**	9.548	0.005	27.69	21	−0.029	1.851	0.190	3.14	20	**0.045**	4.916	0.040	17.60	18	0.033	1.930	0.184	4.68	23	0.006	0.140	0.712	0.00
% of Whites	7*	−0.002	0.062	0.813	0.00	7*	0.013	2.477	0.176	23.21	6*	−0.006	2.714	0.175	32.78	6*	0.002	0.078	0.794	0.00	6*	−0.006	2.188	0.213	61.24
% of Females	27	0.003	1.049	0.316	0.12	21	**0.006**	8.154	0.010	29.83	20	0.003	1.202	0.287	2.56	18	0.001	0.150	0.703	0.00	23	**0.004**	8.807	0.007	38.16
Year of publication	27	**0.071**	10.694	0.003	28.75	21	**−0.046**	6.218	0.022	27.52	20	0.026	1.599	0.222	3.60	18	0.001	0.001	0.971	0.00	23	−0.011	0.654	0.428	0.00

Note. β_1 = estimated regression coefficient; R^2 = Explained variance; *F* = Omnibus test of moderators; Statistically significant effects ($P < 0.05$) appear highlighted in bold.
* Correspond to *K*< 10 and should therefore not be interpreted (Fu et al., 2011).

[5, 36] = 5.182; $P = 0.001$; $R^2 = 35.78$); (b) test version (omnibus-test [1, 40] = 4.264; $P = 0.046$; $R^2 = 7.46$); and (c) publication status (omnibus-test [1, 40] = 4.720; $P = 0.036$; $R^2 = 8.50$). Results from the univariate meta-regression analysis for continuous variables (see Table 3) identified the mean of test score as a significant moderator. Results from the multivariate meta-regression analysis (see Table 4) showed that region, test version, and mean of test score together explained 59.22% of variance in pooled alpha estimate.

Exercise Dependence Questionnaire

The analysis examining the alpha estimates for the global score on the EDQ (see Forest plot in Supplementary material G) included 12 effect sizes from 11 studies (N_{total} = 2,961). Results from the random effects model showed a pooled alpha estimate of 0.862 ($P < 0.001$; 95% CI = 0.842 to 0.879, $I^2 = 84.26$). Since the number of effect sizes available was <15, moderation analyses were not performed.

Exercise Dependence Questionnaire subscales. The analyses examining the alpha estimates for the subscale scores on the EDQ (see Forest plot in Supplementary material G) included 50 single alpha scores. The effect sizes available ranged from six (positive reward, N_{total} = 1,405) to seven (interference, N_{total} = 1,498). Findings from the random effects model showed pooled alpha estimates ranging from 0.615 (social reasons; $P < 0.001$; 95% CI = 0.489 to 0.710, $I^2 = 88.86$) to 0.789 (positive reward; $P < 0.001$; 95% CI = 0.688 to 0.857, $I^2 = 94.89$). Since the number of effect sizes available was <15, moderation analyses were not performed.

Exercise Dependence Scale-21

The analysis examining the reliability estimates for the global score on the EDS-21 (see Forest plot in Supplementary material G) included 90 effect sizes from 84 studies (N_{total} = 35,918). Results from the random effects model showed a pooled alpha estimate of 0.930 ($P < 0.001$; 95% CI = 0.923 to 0.937, $I^2 = 97.96$). Results from the univariate meta-regression analysis for categorical variables (see Table 2) identified both exercise modality (omnibus-test [6, 83] = 4.100; $P = 0.001$; $R^2 = 18.00$) and test version (omnibus-test [1, 88] = 5.930; $P = 0.017$; $R^2 = 5.24$) as significant moderators. Results from the univariate meta-regression analysis for continuous variables (see Table 3) identified both mean test score and *SD* of test score as significant moderators. Results from the multivariate meta-regression analysis showed that exercise modality, test version, and mean test score and *SD* of these scores together explained 38.02% of variance in pooled alpha estimates (see Table 4).

Exercise Dependence Scale-21 subscales. The analyses examining the reliability estimates for the subscale scores on the EDS-21 (see Forest plot in Supplementary material G) included a total of 311 effect sizes. The effect sizes available ranged from 42 (withdrawal, N_{total} = 15,457) to 53 (reduction in other activities, N_{total} = 18,755). Findings from the random

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 7. Results of multivariable meta-regression analyses (subscale scores of the Compulsive Exercise Test)

Moderators	*K*	β_0	β_1	*SE*	*F*	*P*	R^2
Avoidance	27				26.516	<0.001	86.08
		1.300	–	0.263			
Eating disorders (Mixed)			−0.020	0.132			
Eating disorders (Clinical)			**0.615**	0.182			
SD total score*			**0.806**	0.245			
Weight control	21				9.335	0.002	63.26
		2.418	–	0.436			
% of Females*			**0.005**	0.002			
Year of publication*			**−0.042**	0.015			
Mood improvement	20				20.014	<0.001	81.45
		−0.325	–	0.340			
SD total score*			**1.777**	0.321			
SD age*			0.0264	0.013			
Exercise rigidity	23				5.427	0.004	73.70
		1.144	–	0.132			
Region (Oceania)			**0.289**	0.135			
Region (North America)			0.228	0.172			
Region (Mixed)			**0.407**	0.139			
Region (Europe)			0.030	0.090			
% of Females*			**0.003**	0.001			

Note. β_0 = intercept/mean effect size; β_1 = estimated regression coefficient; R^2 = Explained variance; *F* = Omnibus test of moderators. Unknown was considered as the reference category both for Eating disorders and Region. Statistically significant effects ($P < 0.05$) appear highlighted in bold.
* Continuous moderator.

effects model showed pooled alpha estimates ranging from 0.704 (reduction in other activities; $P < 0.001$; 95% CI = 0.675 to 0.730, $I^2 = 92.53$) to 0.881 (intention effects; $P < 0.001$; 95% CI = 0.865 to 0.895, $I^2 = 95.48$). Results from the univariate meta-regression analysis for categorical variables (see Table 8) identified the following significant moderators: (a) tolerance: region (omnibus-test [5, 37] = 4.528, $P = 0.003$, $R^2 = 31.52$), test version (omnibus-test [1, 41] = 6.763, $P = 0.013$, $R^2 = 13.49$), and publication status (omnibus-test [1, 41] = 4.440, $P = 0.041$, $R^2 = 8.69$); (b) withdrawal: region (omnibus-test [5, 36] = 10.317, $P < 0.001$, $R^2 = 61.22$), and test version (omnibus-test [1, 40] = 18.992, $P < 0.001$, $R^2 = 34.95$); (c) intention: report of leisure time (omnibus-test [1, 41] = 4.465, $P = 0.041$, $R^2 = 7.92$), regular exercisers (omnibus-test [1, 41] = 5.434, $P = 0.025$, $R^2 = 10.36$), region (omnibus-test [5, 37] = 10.661, $P < 0.001$, $R^2 = 55.86$), test version (omnibus-test [1, 41] = 28.574, $P < 0.001$, $R^2 = 42.29$), and publication status (omnibus-test [1, 41] = 8.651, $P = 0.005$, $R^2 = 16.05$); (d) lack of control: region (omnibus-test [5, 37] = 10.661, $P < 0.001$, $R^2 = 54.87$), test version (omnibus-test [1, 42] = 28.574, $P < 0.001$, $R^2 = 42.99$), publication status (omnibus-test [1, 42] = 4.475, $P = 0.040$, $R^2 = 8.40$), and study design (omnibus-test [1, 42] = 5.792, $P = 0.021$, $R^2 = 9.99$); (e) time: region (omnibus-test [5, 37] = 5.849, $P < 0.001$, $R^2 = 41.55$), and test version (omnibus-test [1, 41] = 7.396, $P = 0.010$, $R^2 = 15.06$); (f) continuance: region (omnibus-test [5, 37] = 6.759, $P < 0.001$, $R^2 = 45.41$), and test version (omnibus-test [1, 41] = 7.716, $P = 0.008$, $R^2 = 15.95$). The results of the univariate meta-regression analysis for continuous variables (see Table 9) identified of the following significant moderators: (a) test mean score (lack of control); (b) *SD* of test score (tolerance); and (c) percentage of females (tolerance, intention effects, lack of control, time, and continuance). The results of the multivariate meta-regression analysis (see Table 10) supported the moderating role of the following variables: (a) *SD* of test scores and percentage of females, (tolerance); (b) region and percentage of females (intention effects); (c) region and percentage of females (lack of control); (d) test version and percentage of females (Time); and (e) region, test version, and percentage of females (continuance). The amount of variance in pooled alpha estimates explained by the retained models the multivariate meta-regression analyses ranged from 27.97% (tolerance) to 67.73% (intention effects).

Obligatory Exercise Questionnaire

The analysis examining the reliability estimates for the global score on the OEQ (see Forest plot in Supplementary material G) included 38 effect sizes from 33 primary studies (N_{total} = 10,548). Results from the random effects model showed a pooled alpha estimate of 0.870 ($P < 0.001$; 95% CI = 0.853 to 0.885, $I^2 = 84.43$). Results from the univariate meta-regression analysis for categorical variables (see Table 2) identified both exercise modality (omnibus test [3, 34] = 9.568; $P < 0.001$; $R^2 = 43.48$) and (b) regular exercisers (omnibus-test [1, 36] = 10.087; $P = 0.003$; $R^2 = 22.55$) as significant moderators. Results from the univariate meta-regression analysis for continuous variables (see Table 3) did not identify any significant moderators.

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 8. Results of univariable meta-regression analyses for categorical variables (subscale scores of the Exercise Dependence Scale-21)

	Tolerance					Withdrawal					Intention effects					Lack of control					Time					Reduction in other activities					Continuance				
			95% CI					95% CI					95% CI					95% CI						95% CI					95% CI				
Subgroups	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2	*K*	$\bar{\alpha}$	Lo	Up	I^2
Exercise modality																																		
Unknown (RC)	8	0.892	0.859	0.917	91.34	8	0.838	0.793	0.874	90.27	9	0.909	0.877	0.933	94.42	9	0.829	0.762	0.878	95.85	8	0.849	0.811	0.800	88.67	13	0.720	0.639	0.782	93.67	9	0.811	0.748	0.858	94.43
Unclear	18	0.849	0.823	0.870	93.48	17	0.805	0.776	0.829	90.77	17	0.872	0.845	0.894	95.32	17	0.824	0.789	0.853	94.99	18	0.854	0.825	0.878	95.41	18	0.707	0.667	0.741	90.22	17	0.838	0.807	0.863	94.42
Power disciplines	2	0.784	0.690	0.849	69.70	2	0.835	0.799	0.865	0.00	1	0.890	0.854	0.817	–	2	0.765	0.714	0.807	0.00	2	0.805	0.763	0.840	0.00	3	0.762	0.718	0.799	7.61	2	0.844	0.693	0.921	91.47
Non-endurance	2	0.822	0.791	0.848	0.00	2	0.803	0.760	0.838	34.28	2	0.808	0.775	0.836	0.00	2	0.839	0.755	0.895	85.74	2	0.834	0.806	0.859	0.00	2	0.606	0.496	0.692	58.10	2	0.790	0.754	0.821	0.00
Multiple sports	6	0.853	0.798	0.892	94.86	6	0.830	0.779	0.869	92.59	6	0.881	0.833	0.915	95.56	6	0.811	0.750	0.857	0.93.41	6	0.844	0.805	0.875	89.79	6	0.749	0.646	0.822	95.75	6	0.843	0.817	0.865	76.98
Fitness and health	4	0.836	0.751	0.892	96.38	4	0.869	0.764	0.927	98.17	4	0.884	0.843	0.915	93.26	4	0.836	0.802	0.864	81.93	3	0.868	0.838	0.893	83.71	5	0.703	0.617	0.769	88.71	4	0.876	0.830	0.909	93.58
Endurance	3	0.891	0.859	0.915	73.57	3	0.865	0.830	0.892	67.62	4	0.871	0.774	0.926	97.30	4	0.813	0.761	0.855	85.39	4	0.825	0.806	0.843	23.85	6	0.614	0.551	0.667	77.38	3	0.806	0.740	0.855	80.53
Eating disorders																																		
Unknown (RC)	41	0.858	0.841	0.874	94.13	40	0.831	0.811	0.848	93.02	40	0.882	0.865	0.897	95.43	42	0.823	0.802	0.842	94.11	41	0.849	0.834	0.863	92.11	48	0.706	0.676	0.734	92.79	41	0.837	0.819	0.854	93.14
At risk	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Not at risk	1	0.820	0.788	0.847	–	1	0.770	0.729	0.805	–	1	0.850	0.823	0.873	–	1	0.840	0.811	0.864	–	1	0.800	0.764	0.831	–	1	0.680	0.622	0.729	–	1	0.720	0.700	0.763	–
Mixed	1	0.810	0.759	0.851	–	1	0.780	0.721	0.827	–	2	0.871	0.668	0.950	97.92	1	0.800	0.746	0.843	–	1	0.840	0.797	0.874	–	2	0.643	0.451	0.768	89.98	1	0.790	0.733	0.835	–
Clinical	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Report of LTE																																		
No (RC)	12	0.874	0.842	0.900	96.27	11	0.845	0.816	0.869	92.42	12	0.903	0.870	0.928	97.62	13	0.841	0.799	0.874	96.77	13	0.863	0.841	0.883	92.28	19	0.701	0.636	0.754	95.87	12	0.829	0.792	0.860	94.96
Yes	31	0.849	0.829	0.866	91.93	31	0.822	0.797	0.843	92.45	31	0.871	0.854	0.886	92.65	31	0.814	0.792	0.834	90.68	30	0.840	0.821	0.858	90.79	34	0.705	0.674	0.733	88.82	31	0.836	0.814	0.855	92.36
Regular exercisers																																		
Unknown (RC)	16	0.873	0.847	0.895	95.26	15	0.840	0.815	0.861	90.91	17	0.900	0.875	0.920	96.73	17	0.839	0.806	0.866	95.40	17	0.864	0.846	0.880	89.72	25	7.02	0.650	0.746	95.13	16	0.831	0.802	0.855	93.08
Yes	27	0.846	0.824	0.865	92.15	27	0.821	0.794	0.845	93.22	26	0.866	0.847	0.884	92.68	27	0.812	0.786	0.834	91.82	26	0.836	0.813	0.856	91.64	28	0.706	0.673	0.736	88.11	27	0.836	0.811	0.858	93.27
Region																																		
Unknown (RC)	6	0.881	0.846	0.907	87.45	6	0.854	0.824	0.879	76.96	7	0.909	0.880	0.931	91.16	8	0.847	0.807	0.879	91.41	7	0.866	0.838	0.889	83.77	13	0.726	0.634	0.795	95.56	7	0.865	0.839	0.886	79.85
South America	4	0.780	0.737	0.816	67.21	4	0.748	0.646	0.820	90.94	3	0.838	0.790	0.875	82.54	4	0.754	0.712	0.791	59.47	4	0.779	0.721	0.824	79.95	5	0.743	0.639	0.817	91.94	4	0.834	0.772	0.878	89.22
Oceania	1	0.920	0.903	0.934	–	1	0.890	0.866	0.910	–	1	0.930	0.915	0.943	–	1	0.920	0.903	0.934	–	1	0.940	0.927	0.951	–	1	0.760	0.708	0.803	–	1	0.930	0.915	0.943	–
North America	8	0.891	0.854	0.918	95.52	8	0.885	0.860	0.906	90.08	9	0.924	0.912	0.935	85.80	8	0.862	0.832	0.887	90.28	8	0.870	0.845	0.891	87.76	10	0.674	0.625	0.717	84.51	8	0.871	0.847	0.892	86.75
Mixed	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Europe	22	0.847	0.827	0.864	90.65	21	0.809	0.795	0.823	72.08	9	0.845	0.823	0.864	91.68	21	0.797	0.766	0.823	92.74	22	0.838	0.820	0.854	87.73	22	0.688	0.648	0.723	90.90	21	0.796	0.770	0.819	90.36
Asia	2	0.807	0.752	0.850	60.93	2	0.749	0.707	0.786	0.00	2	0.886	0.866	0.902	0.00	2	0.832	0.762	0.882	79.59	1	0.840	0.802	0.871	–	2	0.741	0.697	0.779	0.00	2	0.841	0.814	0.864	0.00
Test version																																		
Original (RC)	18	0.878	0.853	0.899	94.05	18	0.863	0.840	0.882	90.84	19	0.912	0.896	0.926	93.34	20	0.849	0.824	0.871	92.59	19	0.868	0.847	0.885	90.64	25	0.712	0.669	0.749	92.59	19	0.858	0.830	0.881	93.73
Linguistically adapted	25	0.839	0.819	0.857	91.33	24	0.798	0.777	0.816	86.27	24	0.849	0.830	0.866	91.18	24	0.797	0.769	0.821	92.47	24	0.831	0.811	0.849	90.09	28	0.697	0.656	0.732	92.43	24	0.812	0.791	0.831	88.98
Type of survey																																		
Unknown (RC)	18	0.859	0.831	0.882	95.00	24	0.836	0.809	0.859	94.41	12	0.896	0.866	0.919	95.59	18	0.807	0.769	0.839	94.61	22	0.838	0.817	0.856	91.16	32	0.702	0.658	0.740	93.58	22	0.851	0.831	0.869	91.33
Paper-pencil	15	0.863	0.835	0.886	93.46	9	0.807	0.764	0.842	90.27	17	0.886	0.862	0.905	95.44	11	0.830	0.788	0.864	94.33	11	0.856	0.828	0.880	90.06	9	0.690	0.616	0.749	93.50	8	0.808	0.751	0.852	94.16
On-line	7	0.813	0.775	0.845	86.20	8	0.823	0.784	0.855	86.25	12	0.862	0.822	0.893	95.65	12	0.823	0.801	0.842	77.67	9	0.858	0.816	0.890	93.50	11	0.707	0.676	0.735	69.15	11	0.819	0.771	0.857	94.49
Both	3	0.896	0.863	0.921	70.05	1	0.850	0.809	0.882	–	2	0.842	0.795	0.877	0.00	3	0.878	0.779	0.933	94.35	1	0.900	0.859	0.929	–	1	0.840	0.796	0.875	–	2	0.818	0.704	0.888	81.46
Publication status																																		
Published (RC)	40	0.852	0.835	0.868	93.56	39	0.825	0.804	0.843	93.02	39	0.874	0.857	0.889	94.93	41	0.817	0.797	0.836	93.13	40	0.849	0.833	0.863	92.03	49	0.707	0.676	0.734	92.83	40	0.833	0.813	0.850	93.57
Unpublished	3	0.906	0.863	0.936	88.89	3	0.876	0.854	0.894	38.59	4	0.931	0.913	0.946	83.59	3	0.882	0.801	0.931	94.26	3	0.837	0.751	0.893	91.07	4	0.669	0.578	0.741	84.99	3	0.855	0.798	0.895	84.94
Study design																																		
Psychometric (RC)	15	0.843	0.812	0.869	95.02	15	0.815	0.771	0.852	96.66	16	0.873	0.846	0.894	95.73	15	0.789	0.748	0.824	95.04	15	0.838	0.810	0.863	94.11	16	0.712	0.664	0.754	93.49	15	0.832	0.797	0.860	95.34
Applied	28	0.864	0.843	0.882	92.78	27	0.835	0.816	0.852	86.96	27	0.886	0.864	0.903	95.07	29	0.838	0.817	0.857	91.50	28	0.853	0.835	0.869	89.90	37	0.700	0.663	0.733	91.87	28	0.835	0.812	0.855	91.60

Note. $\bar{\alpha}$ = Corrected coefficient alpha. CI = Confidence interval; Lo = Lower; Up = Upper; RC = Reference category. LTE = Leisure time exercise.

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 9. Results of univariable meta-regression analyses for continuous variables (subscale scores of the Exercise Dependence Scale-21)

Moderators	Tolerance				Withdrawal				Intention effects				Lack of control				Time				Reduction in other activities				Continuance			
	K	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2	*K*	β_1	*F*	*P*	R^2
Mean total scores	36	−0.183	3.573	0.067	8.02	36	−0.035	0.312	0.580	9.59	37	−0.170	1.266	0.268	1.19	38	**−0.294**	8.745	0.006	19.17	36	−0.129	2.629	0.114	5.44	41	−0.096	0.748	0.393	0.00	37	0.060	0.256	0.616	0.00
SD total scores	36	**0.623**	4.524	0.041	9.22	36	−0.060	0.054	0.818	9.58	37	0.082	0.041	0.840	0.00	38	−0.075	0.097	0.758	0.00	36	0.210	0.539	0.468	0.00	41	−0.008	0.002	0.967	0.00	37	−0.458	2.319	0.137	4.02
Mean age	41	−0.001	0.003	0.957	0.00	40	−0.003	0.179	0.675	14.64	40	−0.004	0.134	0.716	0.00	42	0.003	0.140	0.710	0.00	41	−0.003	0.198	0.659	0.00	49	0.006	0.519	0.475	0.00	41	−0.011	2.253	0.141	3.84
SD age	40	−0.010	0.430	0.516	0.00	39	−0.019	1.616	0.212	14.32	39	−0.001	0.002	0.966	0.00	41	−0.004	0.055	0.815	0.00	40	−0.001	0.009	0.927	0.00	48	0.022	2.342	0.133	2.34	40	−0.014	0.986	0.327	0.87
% of Whites	8*	0.010	2.638	0.156	19.41	8*	0.008	3.708	0.103	9.21	10	0.003	0.401	0.544	0.00	9*	0.001	0.044	0.841	0.00	7*	−0.007	0.494	0.513	0.00	12	0.002	0.117	0.740	0.00	9*	0.001	0.112	0.748	0.00
% of Females	40	**0.005**	4.256	0.046	8.58	39	0.003	2.242	0.143	13.88	40	**0.006**	5.420	0.025	11.17	41	**0.007**	12.342	0.001	24.97	40	**0.008**	17.577	<0.001	32.19	50	0.002	0.646	0.426	0.00	40	**0.005**	6.018	0.019	12.29
Year of publication	43	0.005	0.126	0.725	0.00	42	−0.022	2.740	0.106	12.99	43	−0.012	0.559	0.459	0.00	44	−0.004	0.065	0.800	0.00	43	−0.003	0.041	0.842	0.00	53	−0.001	0.012	0.913	0.00	43	−0.007	0.258	0.614	0.00

Note. β_1 = estimated regression coefficient; R^2 = Explained variance; *F* = Omnibus test of moderators; Statistically significant effects ($P < 0.05$) appear highlighted in bold.

Results from the multivariate meta-regression analysis showed that exercise modality and regular exercisers together explained 68.55% of variance in pooled alpha estimates (see Table 4).

Reliability reporting practices

A total of 118 studies reported induced reliability (e.g., based on other studies), eleven studies reported unusable reliability indices (i.e., reliability ranges), and eight studies did not report alpha or Pearson's correlation but other reliability indices (i.e., ω, Meule et al., 2021; ρ, Alcaraz-Ibáñez, Aguilar-Parra, & Álvarez-Hernández, 2018; Sicilia, Alcaraz-Ibáñez, Lirola, Burgueño, & Maher, 2018; *ave*, Egan et al., 2017; or *ICC*, Parastatidou, Doganis, Theodorakis, & Vlachopoulos, 2012; Sicilia et al., 2013, 2017; Sicilia & González-Cutre, 2011). A global reliability induction rate of 47.58% was found. This ranged from 18.64% to 57.14% in the case of the global scores and from 14.93% to 66.67% in the case of subscale scores (see Table 11).

Concerning the assumptions required for the unbiased performance of alpha, the first one (i.e., the unidimensionality of the test) was in no case used as an argument to justify the employment of alpha against other reliability indices. Despite the theoretically multidimensional nature of three of the instruments under consideration (CET, EDQ, EDS-21), alpha was frequently used as the reliability index of their global scores (see Table 1). The second assumption (the equality of the factor loadings of the items) was not examined in any of the retrieved studies. The third assumption (i.e., the independency of the error terms), was found to be tested just in the context of improving model fit (e.g.; Zeeck et al., 2017) but in no case to justify the use of alpha or to comment on the implications of using it in such circumstances.

DISCUSSION

The present RG meta-analysis provides summarized evidence on the reliability scores in terms of coefficient alpha of six of the most commonly used self-report instruments assessing PE. Data retrieved from 255 studies (741 independent samples) showed alpha values that ranged from 0.768 to 0.930 for global scores and from 0.615 to 0.907 for subscale scores. The alpha estimates of both global and subscales test scores were affected by several sociodemographic and methodological characteristics. The main implications of these findings are discussed in detail below.

Alpha estimates for total and subscale scores

Interpretation of alpha values has generally been carried out adopting a *more is better* and cut-off-based approach. This implies that the level of reliability of the scores of a given instrument in terms of alpha would dictate the use for which it may be recommended (Cicchetti, 1994; Nunnally & Bernstein, 1994). According to this approach, the alpha estimates found for the global scores of the instruments under consideration may lead to judging them as suitable for (a)

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 10. Results of multivariable meta-regression analyses (subscale scores of the Exercise Dependence Scale-21)

Moderators	*K*	β_0	β_1	*SE*	*F*	*P*	R^2
Tolerance	43				5.591	0.008	27.97
		0.825	–	0.387			
SD total scores*			**0.697**	0.277			
% of Females*			**0.006**	0.002			
Withdrawal	42				10.550	<0.001	67.73
		1.925	–	0.099			
Region (South America)			**−0.569**	0.154			
Region (Oceania)			0.283	0.251			
Region (North America)			0.243	0.128			
Region (Europe)			**−0.270**	0.111			
Region (Asia)			**−0.539**	0.196			
Intention effects	43				9.240	<0.001	69.91
		2.596	–	0.188			
Report of LTE (Yes)			**−0.306**	0.107			
Region (South America)			−0.339	0.217			
Region (Oceania)			0.414	0.322			
Region (North America)			0.216	0.139			
Region (Europe)			**−0.482**	0.123			
Region (Asia)			−0.090	0.241			
% of Females*			−0.000	0.002			
Lack of control	44				4.592	0.002	47.07
		1.661	–	0.146			
Region (South America)			**−0.440**	0.205			
Region (Oceania)			0.375	0.337			
Region (North America)			0.032	0.152			
Region (Europe)			**−0.263**	0.126			
Region (Asia)			−0.264	0.250			
% of Females*			**0.005**	0.002			
Time	43				14.198	<0.001	47.48
		1.683	–	0.100			
Test version (Linguistically adapted)			**−0.218**	0.078			
% of Females*			**0.007**	0.002			
Continuance	43				6.847	<0.001	65.81
		2.004	–	0.148			
Region (South America)			**−0.567**	0.257			
Region (Oceania)			**0.665**	0.290			
Region (North America)			0.057	0.133			
Region (Europe)			**−0.955**	0.248			
Region (Asia)			**−0.770**	0.292			
Test version (Linguistically adapted)			**0.600**	0.226			
% of Females*			−0.000	0.002			

Note. β_0 = intercept/mean effect size; β_1 = estimated regression coefficient; R^2 = Explained variance; *F* = Omnibus test of moderators; LTE = Leisure time exercise. The reference categories were: No (Report of LTE), Unknown (Region), and Original version (Test version). Statistically significant effects ($P < 0.05$) appear highlighted in bold.
* Continuous moderator.

exploratory research (EAI), (b) basic research purposes (CES, CET, EDQ, and OEQ), and (c) applied research and clinical practice (EDS-21). In the case of the subscale scores, applying this same criterion implies considering them as (a) unacceptable for research purposes (insight into problem, social reasons, and stereotyped subscales of the EDQ), (b) acceptable for exploratory research (lack of control and rigidity subscales of the CET; interference, positive reward, withdrawal, weight control, and health reasons subscales of the EDQ; and reduction in other activities subscale of the EDS), (c) suitable for basic research purposes (weight control and mood subscales of the CET; and tolerance, withdrawal, intention effects, lack of control, time, and continuance subscales of the EDS-21), and (d) suitable for applied research and clinical practice (avoidance subscale of the CET). However, the automatic application of cut-off points inherent to this purely quantitative approach of interpreting alpha has been strongly criticised by arguing that they do not emerge as a result of empirical evidence but from researchers' intuition (Cho & Kim, 2015; Hoekstra et al., 2019; Panayides, 2013). Alternatively, it has been suggested that alpha values should be interpreted also taking into account both instrument length and complexity of the construct being assessed (Cho & Kim, 2015). The implications derived from the latter

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Table 11. Reliability reporting practices of in studies using self-report instruments assessing problematic exercise

Measure (Subscale)	Induced reliability				Reported reliability	
	By omission *K* (%)	Vague report *K* (%)	Precise report *K* (%)	Induction rate %	Unusable *K* (%)	Usable *K* (%)
CES-Likert	5 (31.25)	–	–	31.25	1 (6.25)	10 (62.50)
CES-VAS	14 (27.45)	2 (3.92)	5 (9.80)	41.18	–	30 (58.82)
CET	7 (11.86)	3 (5.08)	1 (1.69)	18.64	–	48 (81.36)
CET (Avoidance)	5 (13.16)	4 (10.53)	1 (2.63)	26.32	1 (2.63)	27 (71.05)
CET (Weight control)	5 (16.13)	4 (12.90)	–	29.03	1 (3.23)	21 (67.74)
CET (Mood improvement)	5 (16.67)	4 (13.33)	–	30.00	1 (3.33)	20 (66.67)
CET (Lack of enjoyment)	5 (18.52)	4 (14.81)	–	33.33	–	18 (66.67)
CET (Rigidity)	5 (15.15)	4 (12.12)	1 (3.03)	30.30	–	23 (69.70)
EAI	26 (26.80)	9 (9.28)	17 (17.53)	53.61	2 (2.06)	43 (44.33)
EDQ	3 (10.71)	5 (17.86)	8 (28.57)	57.14	–	12 (42.86)
EDQ (Interference)	1 (5.56)	5 (27.78)	5 (27.78)	61.11	–	7 (38.89)
EDQ (Positive reward)	1 (5.88)	5 (29.41)	5 (29.41)	64.71	–	6 (35.29)
EDQ (Withdrawal)	1 (5.56)	5 (27.78)	5 (27.78)	61.11	–	7 (38.89)
EDQ (Weight control)	2 (11.11)	5 (27.78)	5 (27.78)	66.67	–	6 (33.33)
EDQ (Insight into problem)	1 (5.88)	5 (29.41)	5 (29.41)	64.71		6 (35.29)
EDQ (Social reasons)	2 (11.11)	5 (27.78)	5 (27.78)	66.67	–	6 (33.33)
EDQ (Health reasons)	2 (11.11)	5 (27.78)	5 (27.78)	66.67	–	6 (33.33)
EDQ (Stereotyped behaviour)	1 (5.88)	5 (29.41)	5 (29.41)	64.71	–	6 (35.29)
EDS-21	8 (6.30)	15 (11.81)	6 (4.72)	22.83	8 (6.30)	90 (70.87)
EDS-21 (Tolerance)	1 (1.75)	9 (15.79)	–	17.54	4 (7.02)	43 (75.44)
EDS-21 (Withdrawal)	1 (1.79)	9 (16.07)	–	17.86	4 (7.14)	42 (75.00)
EDS-21 (Intention effects)	1 (1.75)	9 (15.79)	–	17.54	4 (7.02)	43 (75.44)
EDS-21 (Lack of control)	1 (1.72)	9 (15.52)	–	17.24	4 (6.90)	44 (75.86)
EDS-21 (Time)	1 (1.75)	9 (15.79)	–	17.54	4 (7.02)	43 (75.44)
EDS-21 (Reduction in other activities)	1 (1.49)	9 (13.43)	–	14.93	4 (5.97)	53 (79.10)
EDS-21 (Continuance)	1 (1.75)	9 (15.79)	–	17.54	4 (7.02)	43 (75.44)
OEQ	7 (10.00)	5 (7.14)	19 (27.14)	44.29	1 (1.43)	38 (54.29)
Total	113 (9.77)	162 (14.00)	98 (8.47)	47.58	43 (3.72)	741 (64.04)

Note. CES-VAS = Commitment Exercise Scale; CET = Compulsive Exercise Test; EAI = Exercise Addiction Inventory; EDS-21 = Exercise Dependence Scale-21; OEQ = Obligatory Exercise Questionnaire; Induced reliability = No reliability values for the data at hand are provided; By omission = No reference to reliability is made; Vague = Some reference to reliability is made, but information concerning the source of such information is missing; Precise report = Reported reliability values correspond to those provided in another studies; Unusable = Reliability values for the data at hand is provided employing indices different to alpha; Usable = Data that were effectively included in the meta-analysis.

are discussed separately below for the scores with particularly high or low alpha values.

The fact that high alpha values were obtained for some of the scores under consideration (i.e., those near to 0.90 and above) may not necessarily indicate that these are highly reliable. Indeed, high alpha values may also be due to redundancy in the content of the items, particularly, the greater the number of items used (Cho & Kim, 2015). This redundancy is nevertheless undesirable since it could compromise coverage of the construct being assessed. Moreover, the greater its theoretical complexity, the more potentially relevant content is excluded (Hoekstra et al., 2019; Panayides, 2013). Such redundancy may also imply leaving a considerable proportion of individuals' estimates outside the items targeting range, which could result in a decreased reliability (Cho & Kim, 2015; Panayides, 2013). Furthermore, it is worth noting that the instruments whose scores were found to have particularly high alpha values do not appear to have been developed with particular attention to their content validity (e.g., almost none of those studies reported that content validity had been evaluated by a panel of experts). Indeed, it was only in the case of a preliminary version of the EDS-21 that the latter was somewhat indicated, although just in terms of "appropriateness" and providing no other further details on the procedure being followed (Hausenblas & Downs, 2002). Additionally, none of the validation studies reported having examined an aspect of content validity, such as comprehensiveness (i.e., no key aspects of the construct are missed), that is particularly relevant in avoiding content redundancy (Mokkink et al., 2010). Consequently, further research is needed that provide evidence on whether the particularly high alpha values obtained in the present study are due to the true high reliability scores or content validity-related shortcomings.

A second important consideration regarding scores that showed the highest levels of alpha concerns the CET, EDS, and EDQ. More specifically, none of these three scales were proposed as being either unidimensional or higher-order instruments (i.e., including a number of first-order factors and one second-order factor). Indeed, evidence exists

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

supporting the multidimensional versus the unidimensional nature of these instruments (Formby et al., 2014; Sicilia & González-Cutre, 2011). It is therefore surprising to find these instrument scores (and their reliability in terms of alpha) have more often been computed on an aggregate basis than a factor-by-factor basis. This is particularly concerning considering that, in instruments with correlated factors, the use of alpha should be limited to such subscale scores, so that in no case should it be used for the overall test score (Cho, 2016; Cho & Kim, 2015). This leads to a suggestion that, should the overall score of any of the instruments under examination be defensible from a theoretical perspective, reliability should be estimated by adopting methodologically sounder alternatives than alpha (see Cho, 2016; Cho & Kim, 2015; Gignac, 2014).

A first point to note with regard to the instruments whose scores showed the lowest alpha estimates concerns the one whose global score showed the lowest alpha estimate among those examined (i.e., the EAI). One explanation for this finding may be that this instrument was developed on six specific theoretical components of behavioural addictions, therefore just one item per component were proposed (Terry et al., 2004). However, the complex nature of some of these components may not be totally represented by a single item without resorting to the use of complex or double-barrelled items (e.g., the item alluding to the conflicts arising between individuals and their "family and/or partner" because of the amount of exercise being engaged in). Such items may be subject to heterogeneous interpretation and, by extension, to contribute to a lesser extent that those more clearly conceptualizing the underlying latent construct (Hayes & Coutts, 2020; Kyriazos & Stalikas, 2018). The latter implies not fulfilling the tau-equivalence assumption for unbiased estimations of alpha, so that this coefficient no longer reflects the true actual reliability of the score but rather its lower bound (Hayes & Coutts, 2020). Consequently, the possibility exists that the EAI's reliability score was above the one calculated by the analysis in the present study. However, the lack of formal testing of the tau-equivalence assumption of the EAI's items detected in the retrieved studies prevents us from providing empirical evidence that support this possibility, the collection of which should be subject of future research.

A second point to be noted is that with regard to the instruments whose scores showed the lowest alpha estimates concerns the three subscale scores of the EDQ showing alpha values below the minimum 0.70 cut-off traditionally employed for discouraging the employment of a given score (i.e., insight into problem, social reasons, and stereotyped behaviour). These findings are not entirely surprising considering the difficulty of achieving high alpha values using only a few items in the subscales (i.e., from two to four) (Greco, O'Boyle, Cockburn, & Yuan, 2018). However, it is worth noting that, despite using a similarly small number of items, the scores on some of the other subscales examined (e.g., those of the EDS-21) showed higher levels of alpha than the three aforementioned EDQ subscales. The explanation for these differences is probably due to the way in which the content of the two instruments were developed. That is, on the basis of the theoretical definition of the seven constructs being assessed (in the EDS-21), or by assigning the statements provided by exercisers concerning their exercise-related feelings and cognitions to the factors emerging from statistical analyses (in the EDQ). Therefore, the fact that the items included in these three subscales of the EDQ with particularly low alpha values did not derive from a predetermined theoretical approach could have meant grouping indicators that do not reflect an unequivocal underlying factor, leading to decreased measurement reliability. This is important because low reliability tends to attenuate the strength of the relationship being examined (Graham & Unterschute, 2015). Consequently, these findings raise the need to review the content and number of items included in these subscales in order to improve their reliability.

Moderators of the reliability scores of self-report instruments of PE

Evidence supported the relationship between some of the characteristics of the studies evaluated and the variability in alpha estimates. For example, higher alpha values were found for the global scores of the CES-VAS and the avoidance and rule-driven behaviour subscale of the CET among clinical populations in terms of eating disorders. These findings are relatively unsurprising given that both instruments include content of particular relevance to individuals with eating disorders such as the negative consequences of being unable to exercise, especially feelings of guilt (Davis et al., 1993; Scharmer et al., 2020; Taranis et al., 2011; Zeeck et al., 2017). It follows that comparing scores derived from these two instruments involving individuals with and without a clinical eating disorder diagnosis may be susceptible to bias.

Findings also suggested that the alpha values of the global scores of the CET and the OEQ may be lower among populations comprising regular exercisers. Moreover, it should be noted that the CET was developed with a particular focus on excessive exercise within the eating disorders domain. Therefore, the possibility exists that some of the content included in the instrument (e.g., exercising due to weight/appearance reasons or to the lack of enjoyment when exercising; Taranis et al., 2011) may not be equally relevant for non-clinical populations in terms of eating disorders (Alcaraz-Ibáñez, Sicilia, Dumitru, Paterna, & Griffiths, 2019). Additionally, the lower alpha values obtained for OEQ scores among regular exercisers may be due to the low potential variability of some of the instrument's items among those featuring very low levels of exercise. Clear examples are items referring to exercise frequency (e.g., exercising on a daily basis) or specific exercise-related habits (e.g., keeping a record of exercise performance) (Pasman & Thompson, 1988). Taken together, these results reinforce the notion that differences in the interpretation of the content of self-report instruments assessing PE may exist among individuals with unequal levels of exercise involvement (Szabo et al., 2015).

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Exercise modality is another exercise-related feature that support the likely relationship in alpha estimate variability (i.e., the global scores of the EDS-21). In particular, results suggested that alpha values were lower in studies reporting very precise exercise modalities compared to those that did not. However, the fact that the instrument scores under consideration were found to be similarly reliable in terms of alpha values suggests that comparisons across modalities could be reasonably made. This is important given that this kind of comparison has been a matter of research interest (Di Lodovico et al., 2019).

Findings also suggested that the alpha estimates of the linguistically adapted versions may be lower than original versions in the case of CES-VAT and EAI global scores, and several EDS-21 subscale scores. These findings suggest the existence of possible weaknesses in the linguistic adaptation processes. However, it should be noted that cross-cultural and cross-linguistic research in this field is scarce (Griffiths et al., 2015). Consequently, further research is needed that examines the extent to which the psychometric properties of the scores of the self-report instruments assessing PE are equivalent across their different linguistic adaptations.

There was no conclusive evidence found linking the proportion of females included in the samples with the alpha estimates of the global scores of the instruments under consideration. This suggests that the reliability of such scores does not greatly differ between males and females. However, this was not the case for some of the subscale scores (i.e., weight control and exercise rigidity subscales of the CET; and tolerance, lack of control, and time subscales of the EDS-21). Indeed, evidence suggested that the higher the number of females in the sample, the higher the reliability alpha estimates of these subscale scores. Therefore, the reliability of these scores may be lower for males than for females. These findings are relevant considering that gender has been proposed as a potential risk factor for several potentially addictive behaviours and, particularly, PE (Bueno-Antequera et al., 2020; Cunningham, Pearman, & Brewerton, 2016). The existence of gender differences in reliability scores may have led to biased estimates in comparisons involving these two population groups.

A last notable group of findings emerging from moderator analyses concerns continuous variables. The fact that no evidence was obtained relating alpha values to mean scores on the scales suggests that the reliability of the scores examined is likely to be similar among individuals with very different levels of self-reported PE. An exception to this general trend was the negative relationship observed between the mean scores and the associated reliability values in the case of the EAI. This is important because it suggests that the reliability of the EAI scores may decrease among individuals scoring high on this instrument. This might be explained by evidence suggesting that individuals with similarly high levels of PE on the EAI may differ markedly on the score for the item reflecting conflict (Chamberlain & Grant, 2020; Sicilia, Alcaraz-Ibáñez, Chiminazzo, & Fernandes, 2020). This may imply a decreased level of intercorrelations among items and, by extension, a decrease in alpha values (Greco et al., 2018).

Finally, it worth noting that the variance of scores under consideration were found to be positively related to alpha estimates in just in three cases (i.e., the avoidance and mood modification subscales of the CET, and the tolerance subscale of the EDS-21). These findings are somewhat unexpected considering that psychometric theory points to score variance as one of the main components of reliability estimation (Nunnally & Bernstein, 1994). From this, it follows that the population characteristics already discussed here may help explain the variability of alpha to a greater extent than the standard deviation of the scores. On balance, findings from the moderator analyses underscore the need to examine reliability in each of the groups involved in cross-groups comparisons on self-reported PE symptoms.

Reliability reporting practices in studies using self-report assessment of problematic exercise

The global induction rate found in the present study (i.e., 47.58%) appears to be slightly higher than the one reported for exercise psychology research more generally (i.e., 41.20%; Wilson, Mack, & Sylvester, 2011). It is worth noting that induction rates above the mean were found for the instruments whose scores showed the lowest values of alpha at the global level (i.e., EAI) and subscale level (i.e., EDQ). This suggests that information concerning reliability in this field may be more likely to be omitted for those scores with lower values of alpha. In the case of the EAI, one explanation for these findings may be that this instrument has been used not only for providing a continuous score representing the construct of interest but also as a screening instrument for the purpose of distinguishing individuals at-risk from those having some or no symptoms of exercise addiction. Therefore, the possibility exists that the focus on classifying individuals on the basis of a fixed cut-off point may have led some authors to overlook the issue of examining the reliability of the instrument's global score.

A particularly worrying issue in view of the highly prevalent use of alpha is the almost non-existent testing of the assumptions required for its unbiased employment. Researchers in this field may opt instead to use the reliability index that is most appropriate to the data (Cho & Kim, 2015). A misconception that may deter researchers from approaching this task is the alleged difficulty of both testing the assumptions of alpha and using the alternative methods required when its assumptions are violated (Cho, 2016; Hayes & Coutts, 2020). However, it should be noted that convenient practical guidelines for addressing these tasks have been provided, with some involving relatively non-complex tools (e.g., spreadsheet-based solutions; Cho, 2016) or software that is familiar to large numbers of researchers (e.g., SPSS; Hayes & Coutts, 2020).

Limitations

Despite the many strengths of the present review, there are a number of limitations. A first group of limitations concerns the limited data available on the population

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

characteristics being examined as potential moderators. For example, the small number of studies reporting reliability estimates in some populations meant that, in many cases, only a small number of primary estimates were available. This prevented providing a higher level of evidence for some of the moderation analyses conducted or even, in some cases, from carrying them out at all. The latter was the case for the EDQ, for which it was impossible to examine the variables that may contribute to the variability of the alpha estimates of its global and subscale scores. Also related to the limited availability of data were the characteristics of the study participants. For example, there were more studies that omitted information on exercise modalities or minimum exercise levels of the participants than those that provided such information. These omissions are particularly relevant in view of the limited amount of variance (i.e., <50%) explained by some of the regression models aimed at exploring the potential sources of variability in the alpha estimates. This is so because these relatively low levels of explained variance point towards the existence of other important moderator variables beyond those considered in the present study. This scarcity of data is also relevant given the results here pointed to some of the variables for which limited data were available (e.g., region or exercise modality) as potential moderators of the alpha estimates under consideration. In view of these limitations, a two suggestions can be made. Firstly, researchers in this field should pay particular attention to reporting the characteristics of study participants. This means providing sociodemographic information that, in view of the findings here, may be of interest due to its likely influence on the reliability levels of the scores in terms of coefficient alpha. Examples of the latter include the type of survey, volume of exercise, and the main exercise modality practised. Moreover, it would be particularly useful to provide specific information for the subgroups identified on the basis of these or other socio-demographic variables, because this would facilitate further meta-analytical research. Secondly, more research is needed that examines the reliability of the scores of self-report instruments assessing PE among populations for which limited evidence is currently available. Depending on the instrument, this would involve regions or linguistic contexts still under represented, as well as clinical populations in terms of eating disorders.

A second important limitation is that the fact that there were virtually no primary studies reporting test-retest reliability. This prevented the providing of summarized evidence on the consistency of instrument scores over time. Therefore, further primary research is needed examining the reliability of the test scores under consideration in terms of temporal stability. Finally, it worth mentioning the lack of testing of the assumptions required for the unbiased function of alpha. This makes it advisable to treat the results presented with caution, particularly in the case of the global scores of instruments with a non-clearly unidimensional character (i.e., EDQ, CET, and EDS-21).

Conclusions and practical implications

First, the alpha estimates of the global and subscale scores of existing self-report instruments assessing PE vary largely not just from one to the other but also across different applications. Indeed, the 95% CI of the summarized alpha estimates obtained in the present study did not contain (in most cases) the alpha values reported in the studies in which the instruments under consideration were originally proposed. Therefore, the possibility exists that the originally-reported alpha values were not the most adequate ones to be compared with those obtained in primary research, nor to correct for measurement-related artefacts in quantitative meta-analytic research. It is therefore suggested that the values provided in the present study should be used for such purposes.

Second, the reliability of test scores of existing self-report instruments assessing PE appears to be particularly sensitive to the characteristics of the study population. Researchers including the self-report PE instruments in their studies are encouraged to report specific reliability estimates for the different population groups of interest. This would provide insight into the potential for cross-group comparisons to be biased by the presence of differences in inter-group reliability. Future research efforts aimed at refining existing instruments or proposing new ones should be conducted including not just one or two convenience samples but, instead, several groups according to the characteristics that were proved to be related with the variability in alpha estimates (e.g., clinical condition in terms of eating disorders, language, and exercise modality). This would allow for examining the extent to which the instrument's scores are acceptable in terms of reliability for a minimum number of target groups of interest, which, if this were not the case, would allow the instrument to be refined at an early stage of development.

Third, existing quantitative research using self-report instruments assessing PE suffers from two main deficiencies in terms of reliability reporting: (i) the frequent omission of reliability estimates for the data at hand; and (ii) the (almost exclusive) employment of alpha without proper testing of the assumptions necessary for its unbiased use or even when the nature of the test to be examined would make its use particularly unsuitable. Researchers, journal editors, and reviewers should be aware of the need to report the reliability of scores derived from instruments assessing PE for the data at hand in all primary research. Therefore, the suitability of reliability index to be used should be justified on the basis of the theoretical nature of the constructs under consideration and the characteristics of the data being examined, for example, in terms of test dimensionality and measurement model.

Funding sources: This research is part of the I+D+I project (grant number PID2019-107674RB-I00), funded by Ministerio de Ciencia e Innovación (MCIN), Agencia Estatal de Investigación (AEI/10.13039/501100011033), Spain. AP

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

(FPU18/01055) is funded by MCIN/AEI/10.13039/501100011033 and Fondo Social Europeo (FSE) "El FSE invierte en tu futuro". MAI (UAL RRA202101) is funded by Ministerio de Universidades (Plan de Recuperación, Transformación y Resiliencia, Next Generation EU).

Authors' contribution: AP and MAI designed the study, performed the systematic search and data extraction, completed all statistical analyses and initial drafts of the manuscript. AS and MDG contributed to the drafting of the manuscript and revisions. All authors assisted with drafting of the final version of the manuscript, including critical revisions for intellectual content.

Conflicts of interest: The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

SUPPLEMENTARY MATERIAL

Supplementary data to this article can be found online at https://doi.org/10.1556/2006.2022.00014.

REFERENCES

Alcaraz-Ibáñez, M., Aguilar-Parra, J. M., & Álvarez-Hernández, J. F. (2018). Exercise addiction: Preliminary evidence on the role of psychological inflexibility. *International Journal of Mental Health and Addiction*, *16*(1), 199–206. https://doi.org/10.1007/s11469-018-9875-y.

Alcaraz-Ibáñez, M., Paterna, A., Sicilia, A., & Griffiths, M. D. (2020). Morbid exercise behaviour and eating disorders: A meta-analysis. *Journal of Behavioral Addictions*, *9*(2), 206–224. https://doi.org/10.1556/2006.2020.00027.

Alcaraz-Ibáñez, M., Paterna, A., Sicilia, A., & Griffiths, M. D. (2021). A systematic review and meta-analysis on the relationship between body dissatisfaction and morbid exercise behaviour. *International Journal of Environmental Research and Public Health*, *18*, 585. https://doi.org/10.3390/ijerph18020585.

Alcaraz-Ibáñez, M., Sicilia, A., Dumitru, D. C., Paterna, A., & Griffiths, M. D. (2019). Examining the relationship between fitness-related self-conscious emotions, disordered eating symptoms, and morbid exercise behavior: An exploratory study. *Journal of Behavioral Addictions*, *8*(3), 603–612. https://doi.org/10.1556/2006.8.2019.43.

Alchieri, J. C., Gouveia, V. V., de oliveira, I. C. V., de medeiros, E. D., Grangeiro, A. S. de M., & da Silva, C. F. de L. S. (2015). Exercise dependence scale: Adaptação e evidências de validade e precisão. *Jornal Brasileiro de Psiquiatria*, *64*(4), 279–287. https://doi.org/10.1590/0047-2085000000090.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders. DSM-IV* (4th ed.). American Psychiatric Association.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335–340. https://doi.org/10.3102/10769986027004335.

Bueno-Antequera, J., Mayolas-Pi, C., Reverter-Masià, J., López-Laval, I., Oviedo-Caro, M. Á., Munguía-Izquierdo, D., … Legaz-Arrese, A. (2020). Exercise addiction and its relationship with health outcomes in indoor cycling practitioners in fitness centers. *International Journal of Environmental Research and Public Health*, *17*, 4159. https://doi.org/10.3390/ijerph17114159.

Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., … Willumsen, J. F. (2020). World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine*, *54*(24), 1451–1462. https://doi.org/10.1136/bjsports-2020-102955.

Chamberlain, S. R., & Grant, J. E. (2020). Is problematic exercise really problematic? A dimensional approach. *CNS Spectrums*, *25*(1), 64–70. https://doi.org/10.1017/S1092852919000762.

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.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessments instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.28.

Cunningham, H. E., Pearman, S., & Brewerton, T. D. (2016). Conceptualizing primary and secondary pathological exercise using available measures of excessive exercise. *International Journal of Eating Disorders*, *49*(8), 778–792. https://doi.org/10.1002/eat.22551.

Davis, C., Brewer, H., & Ratusny, D. (1993). Behavioral frequency and psychological commitment: Necessary concepts in the study of excessive exercising. *Journal of Behavioral Medicine*, *16*(6), 611–628. https://doi.org/10.1007/BF00844722.

Di Lodovico, L., Poulnais, S., & Gorwood, P. (2019). Which sports are more at risk of physical exercise addiction: A systematic review. *Addictive Behaviors*, *93*, 257–262. https://doi.org/10.1016/j.addbeh.2018.12.030.

Ding, D., Lawson, K. D., Kolbe-Alexander, T. L., Finkelstein, E. A., Katzmarzyk, P. T., van Mechelen, W., & Pratt, M. (2016). The economic burden of physical inactivity: A global analysis of major non-communicable diseases. *The Lancet*, *388*(10051), 1311–1324. https://doi.org/10.1016/S0140-6736(16)30383-X.

Downs, D. S., Hausenblas, H. A., & Nigg, C. R. (2004). Factorial validity and psychometric examination of the exercise dependence scale-revised. *Measurement in Physical Education and Exercise Science*, *8*(4), 183–201. https://doi.org/10.1207/s15327841mpee0804.

Egan, S. J., Bodill, K., Watson, H. J., Valentine, E., Shu, C., & Hagger, M. S. (2017). Compulsive exercise as a mediator between clinical perfectionism and eating pathology. *Eating Behaviors*, *24*, 11–16. https://doi.org/10.1016/j.eatbeh.2016.11.001.

Formby, P., Watson, H. J., Hilyard, A., Martin, K., & Egan, S. J. (2014). Psychometric properties of the Compulsive Exercise Test in an adolescent eating disorder population. *Eating Behaviors*, *15*(4), 555–557. https://doi.org/10.1016/j.eatbeh.2014.08.013.

Fu, R., Gartlehner, G., Grant, M., Shamliyan, T., Sedrakyan, A., & Wilt, T. J., (2011). Conducting quantitative synthesis when

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

comparing medical interventions: AHRQ and the effective health care program. *Journal of Clinical Epidemiology*, *64*, 1187–1197. https://doi.org/10.1016/j.jclinepi.2010.08.010.

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, *30*(2), 130–139. https://doi.org/10.1027/1015-5759/a000181.

Graham, J. M., & Unterschute, M. S. (2015). A reliability generalization meta-analysis of self-report measures of adult attachment. *Journal of Personality Assessment*, *97*(1), 31–41. https://doi.org/10.1080/00223891.2014.927768.

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-analysis of coefficient alpha: A reliability generalization study. *Journal of Management Studies*, *55*(4), 583–618. https://doi.org/10.1111/joms.12328.

Griffiths, M. D., Szabo, A., & Terry, A. (2005). The exercise addiction inventory: A quick and easy screening tool for health practitioners. *British Journal of Sports Medicine*, *39*(6), 1–3. https://doi.org/10.1136/bjsm.2004.017020.

Griffiths, M. D., Urbán, R., Demetrovics, Z., Lichtenstein, M. B., de la Vega, R., Kun, B., . . . Szabo, A. (2015). A cross-cultural re-evaluation of the Exercise Addiction Inventory (EAI) in five countries. *Sports Medicine - Open*, *1*(5), 1–7. https://doi.org/10.1186/s40798-014-0005-5.

Hausenblas, H. A., & Downs, D. S. (2002). How much is too much? The development and validation of the exercise dependence scale. *Psychology & Health*, *17*(4), 387–404. https://doi.org/10.1080/0887044022000004894.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But. . .. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses testing for heterogeneity. *BMJ*, *327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557.

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology*, *22*(4), 351–364. https://doi.org/10.1080/13645579.2018.1547523.

Juwono, I. D., & Szabo, A. (2021). 100 cases of exercise addiction: More evidence for a widely researched but rarely identified dysfunction. *International Journal of Mental Health and Addiction*, *19*, 1799–1811. https://doi.org/10.1007/s11469-020-00264-6.

Kern, L. (2007). Validation de l'adaptation française de l'échelle de dépendance à l'exercice physique: l'EDS-R. *Pratiques Psychologiques*, *13*(4), 425–441. https://doi.org/10.1016/j.prps.2007.06.003.

Kern, L., & Baudin, N. (2011). Validation franaise du questionnaire de dépendance de l'exercice physique (Exercise Dependence Questionnaire). *Revue Europeene de Psychologie Appliquee*, *61*(4), 205–211. https://doi.org/10.1016/j.erap.2011.08.001.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482.

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.

Lichtenstein, M. B., & Jensen, T. T. (2016). Exercise addiction in CrossFit: Prevalence and psychometric properties of the exercise addiction inventory. *Addictive Behaviors Reports*, *3*, 33–37. https://doi.org/10.1016/j.abrep.2016.02.002.

Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PloS One*, *13*(9), e0204056. https://doi.org/10.1371/journal.pone.0204056.

Li, M., Nie, J., & Ren, Y. (2016). Verification of Exercise Addiction Inventory for Chinese college students based on SEM model. *International Journal of Simulation: Systems, Science and Technology*, *17*(12), 21.1–21.6. https://doi.org/10.5013/IJSSST.a.17.12.21.

Lipsey, M. W., & Wilson, D. (2001). Practical meta analysis. In *Applied social research methods series*. Sage Publications.

Marques, A., Peralta, M., Sarmento, H., Loureiro, V., Gouveia, É. R., & Gaspar de Matos, M. (2019). Prevalence of risk for exercise dependence: A systematic review. *Sports Medicine*, *49*(2), 319–330. https://doi.org/10.1007/s40279-018-1011-4.

Meule, A., Schrambke, D., Furst Loredo, A., Schlegl, S., Naab, S., & Voderholzer, U. (2021). Inpatient treatment of anorexia nervosa in adolescents: A 1-year follow-up study. *European Eating Disorders Review*, *29*, 165–177. https://doi.org/10.1002/erv.2808.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Plos Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.

Mónok, K., Berczik, K., Urbán, R., Szabo, A., Griffiths, M. D., Farkas, J., . . . Demetrovics, Z. (2012). Psychometric properties and concurrent validity of two exercise addiction measures: A population wide study. *Psychology of Sport and Exercise*, *13*(6), 739–746. https://doi.org/10.1016/j.psychsport.2012.06.003.

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory*. McGraw-Hil.

Ogden, J., Veale, D., & Summers, Z. (1997). The development and validation of the exercise dependence questionnaire. *Addiction Research*, *5*(4), 343–356. https://doi.org/10.3109/16066359709004348.

Page, M. J., Higgins, J. P. T., & Sterne, J. A. C. (2019). Assessing risk of bias due to missing results in a synthesis. In J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions*. Cochrane.

Panayides, P. (2013). Coefficient alpha: Interpret with caution. *Europe's Journal of Psychology*, *9*(4), 687–696. https://doi.org/10.5964/ejop.v9i4.653.

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Parastatidou, I. S., Doganis, G., Theodorakis, Y., & Vlachopoulos, S. P. (2012). Addicted to exercise: Psychometric properties of the exercise dependence scale-revised in a sample of Greek exercise participants. *European Journal of Psychological Assessment*, *28*(1), 3–10. https://doi.org/10.1027/1015-5759/a000084.

Pasman, L., & Thompson, J. K. (1988). Body image and eating disturbance in obligatory runners, obligatory weightlifters, and sedentary individuals. *International Journal of Eating Disorders*, *7*(6), 759–769. https://doi.org/10.1002/1098-108X(198811)7:6<759::AID-EAT2260070605>3.0.CO;2-G.

Pigott, T. D. (2012). *Advances in meta-analysis*. Springer. https://doi.org/10.1007/978-1-4614-2278-5.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*(2), 183–192. https://doi.org/10.1037/0033-2909.118.2.183.

Rubio-Aparicio, M., Badenes-Ribera, L., Sánchez-Meca, J., Fabris, M. A., & Longobardi, C. (2020). A reliability generalization meta-analysis of self-report measures of muscle dysmorphia. *Clinical Psychology: Science and Practice*, *27*(1), 1–24. https://doi.org/10.1111/cpsp.12303.

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 402–425. https://doi.org/10.1111/j.2044-8317.2012.02057.x.

Sauchelli, S., Arcelus, J., Granero, R., Jiménez-Murcia, S., Agüera, Z., Del Pino-Gutiérrez, A., & Fernández-Aranda, F. (2016). Dimensions of compulsive exercise across eating disorder diagnostic subtypes and the validation of the Spanish version of the Compulsive Exercise Test. *Frontiers in Psychology*, *7*, 1852. https://doi.org/10.3389/fpsyg.2016.01852.

Scharmer, C., Gorrell, S., Schaumberg, K., & Anderson, D. A. (2020). Compulsive exercise or exercise dependence? Clarifying conceptualizations of exercise in the context of eating disorder pathology. *Psychology of Sport and Exercise*, *46*, 101586. https://doi.org/10.1016/j.psychsport.2019.101586.

Shin, K., & You, S. (2015). Factorial validity of the Korean version of the exercise dependence scale–revised. *Perceptual and Motor Skills*, *121*(3), 889–899. https://doi.org/10.2466/03.08.PMS.121c27x8.

Sicilia, A., Alcaraz-Ibáñez, M., Chiminazzo, J. G. C., & Fernandes, P. T. (2020). Latent profile analysis of exercise addiction symptoms in Brazilian adolescents: Association with health-related variables. *Journal of Affective Disorders*, *273*, 223–230. https://doi.org/10.1016/j.jad.2020.04.019.

Sicilia, A., Alcaraz-Ibáñez, M., Lirola, M. J., Burgueño, R., & Maher, A. (2018). Exercise motivational regulations and exercise addiction: The mediating role of passion. *Journal of Behavioral Addictions*, *7*(2), 482–492. https://doi.org/10.1556/2006.7.2018.36.

Sicilia, A., Alías-García, A., Ferriz, R., & Moreno-Murcia, J. A. (2013). Spanish adaptation and validation of the exercise addiction inventory (EAI). *Psicothema*, *25*(3), 377–383. https://doi.org/10.7334/psicothema2013.21.

Sicilia, A., Bracht, V., Penha, V., Almeida, U. R., Ferriz, R., & Alcaraz-Ibáñez, M. (2017). Propiedades psicométricas del Exercise Addiction Inventory (EAI) en una muestra de estudiantes brasileños universitarios [Psychometric properties of the Exercise Addiction Inventory (EAI) in a sample of Brazilian university students]. *Universitas Psychologica*, *16*(2), 176–185. https://doi.org/10.11144/Javeriana.upsy16-2.ppea.

Sicilia, A., & González-Cutre, D. (2011). Dependence and physical exercise: Spanish validation of the exercise dependence scale-revised (EDS-R). *The Spanish Journal of Psychology*, *14*(1), 421–431. https://doi.org/10.5209/rev_SJOP.2011.v14.n1.38.

Sicilia, A., Paterna, A., Alcaraz-Ibáñez, M., & Griffiths, M. D. (2021). Theoretical conceptualisations of problematic exercise in psychometric assement instruments: A systematic review. *Journal of Behavloral Addictions*, *10*(1), 4–20. https://doi.org/10.1556/2006.2021.00019.

Slaney, K. (2017). *Validating psychological constructs*. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-38523-9.

Szabo, A., Demetrovics, Z., & Griffiths, M. D. (2018). Morbid exercise behavior: Addiction or psychological escape? In H. Budde & M. Wegner (Eds.), *The exercise effect on mental health: Neurobiological mechanisms* (pp. 277–311). Routledge.

Szabo, A., Griffiths, M. D., de La Vega Marcos, R., Mervó, B., & Demetrovics, Z. (2015). Methodological and conceptual limitations in exercise addiction research. *Yale Journal of Biology and Medicine*, *88*, 303–308.

Szabo, A., Pinto, A., Griffiths, M. D., Kovácsik, R., & Demetrovics, Z. (2019). The psychometric evaluation of the Revised Exercise Addiction Inventory: Improved psychometric properties by changing item response rating. *Journal of Behavioral Addictions*, *8*(1), 157–161. https://doi.org/10.1556/2006.8.2019.06.

Taranis, L., Touyz, S., & Meyer, C. (2011). Disordered eating and exercise: Development and preliminary validation of the compulsive exercise test (CET). *European Eating Disorders Review*, *19*(3), 256–268. https://doi.org/10.1002/erv.1108.

Terry, A., Szabo, A., & Griffiths, M. D. (2004). The exercise addiction inventory: A new brief screening tool. *Addiction Research and Theory*, *12*(5), 489–499. https://doi.org/10.1080/16066350310001637363.

Trott, M., Yang, L., Jackson, S. E., Firth, J., Gillvray, C., Stubbs, B., & Smith, L. (2020). Prevalence and correlates of exercise addiction in the presence vs. absence of indicated eating disorders. *Frontiers in Sports and Active Living*, *2*(84). https://doi.org/10.3389/fspor.2020.00084.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*(4), 562–569. https://doi.org/10.1177/0013164402062004002.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those of test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*(4), 502–522. https://doi.org/10.1177/00131640021970682.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44*(3), 159–168. https://doi.org/10.1177/0748175611409845.

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC

Vicent, M., Rubio-Aparicio, M., Sánchez-Meca, J., & Gonzálvez, C. (2019). A reliability generalization meta-analysis of the child and adolescent perfectionism scale. *Journal of Affective Disorders*, *245*, 533–544. https://doi.org/10.1016/j.jad.2018.11.049.

Wilson, P. M., Mack, D. E., & Sylvester, B. (2011). When a little myth goes a long way: The use (or misuse) of cut-points, interpretations, and discourse with coefficient-alpha in exercise psychology. In A. M. Columbus (Ed.), *Advances in psychology research* (pp. 1–17). Nova Science Publishers.

Zeeck, A., Schlegel, S., Giel, K. E., Junne, F., Kopp, C., Joos, A., … Hartmann, A. (2017). Validation of the German version of the commitment to exercise scale. *Psychopathology*, *50*(2), 146–156. https://doi.org/10.1159/000455929.

Open Access. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (https://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and reproduction in any medium for non-commercial purposes, provided the original author and source are credited, a link to the CC License is provided, and changes – if any – are indicated.

t to you by Library and Information Centre of the Hungarian Academy of Sciences MTA | Unauthenticated | Downloaded 08/30/23 07:30 AM UTC