INTERNATIONAL REVIEW OF SPORT AND EXERCISE PSYCHOLOGY
https://doi.org/10.1080/1750984X.2022.2111664

Psychometric properties of problematic exercise measures: a systematic review

Manuel Alcaraz-Ibáñez [a†], Adrian Paterna [a†], Mark D. Griffiths [b] and Álvaro Sicilia [a]

[a]Health Research Centre and Department of Education, University of Almería, Almería, Spain; [b]Psychology Department, Nottingham Trent University, Nottingham, UK

ABSTRACT

The present study summarized, compared, and critically appraised the methodological quality of the most used self-report measures assessing problematic exercise (PE) (i.e. CES, CET, EAI, EDQ, EDS, OEQ). A pre-registered systematic review was conducted in accordance with the 2018 COSMIN criteria and PRISMA methodology. Six electronic databases were searched for studies developing, validating and/or testing the psychometric properties of the psychometric instruments. Data from 48 studies comprising the six original instruments and their eight modified versions were included. The methodological quality (risk-bias) of the development studies of all 14 instruments was rated as 'inadequate'. Limited evidence base in support of most of the measurement properties under examination was found, with the most relevant being that concerning content validity. Findings call into question (i) the accuracy and usefulness of the body of evidence obtained by employing these instruments, and (ii) the advisability of persisting with its use, at least until the issues identified in the present study have been adequately addressed. Obtaining further evidence on the measurement properties of existing self-report PE instruments as well as providing them in early stages of development for those to be proposed in the future should be a priority for research in this field.

ARTICLE HISTORY
Received 16 November 2021
Accepted 1 August 2022

KEYWORDS
Psychometric evaluation; problematic exercise; morbid exercise; exercise dependence; exercise addiction; COSMIN

Introduction

Regular exercise has been found to provide many valuable health benefits (Thompson et al., 2020). Nevertheless, there is also evidence to suggest that specific patterns of exercise behaviour can become problematic (Juwono & Szabo, 2020). Examples of the latter include exercising to the point where social and/or professional life is impaired or persisting in exercising even in the presence of physical or psychological harm (Szabo et al., 2018). Consequently, a better understanding of these potentially dysfunctional forms of exercise behaviour (which given the multiplicity of terms used in the literature will be referred to hereafter by using the umbrella term 'problematic exercise' [PE]; Sicilia, Paterna, Alcaraz-Ibáñez, et al., 2021) is warranted.

Much of the existing literature on PE derives from the use of quantitative techniques and, more specifically, self-report instruments (Szabo et al., 2018). As far as these instruments are concerned, several important considerations need to be made. Two of these considerations are drawn from the fact that PE has not yet been recognized as a nosographic entity in any psychological or medical diagnostic frameworks (e.g., the Diagnostic and Statistical Manual of Mental Disorders, DSM-5, American Psychiatric Association, 2013; or the International Classification of Diseases, ICD-11, World Health Organization, 2019). Firstly, that these instruments do not serve as clinical categorical diagnostic tools but produce continuous scores reflecting an increased presence of a potentially problematic patterns of exercise behaviour (Szabo et al., 2015). Secondly, that these instruments were not created for the purpose of assessing the very same construct (e.g., by covering a set of previously agreed diagnostic symptoms), but rather different manifestations of the behavioural pattern under study that were considered relevant from the perspectives or theoretical frameworks adopted by their respective scale developers. As a result, PE has been conceptualised (i) only in terms of the occurrence of excessive amounts of practice, (ii) as a likely maladaptive compensatory behaviour within the context of weight loss and body appearance modification, or (iii) as a potential disorder analogous to behavioural or substance dependence problems (Sicilia, Paterna, Alcaraz-Ibáñez, et al., 2021). This implies that the number and the very specific nature of the different manifestations of PE covered by each psychometric instrument vary from one scale to another (Sicilia et al., 2022).

Another noteworthy consideration concerning self-report instruments proposed for assessing PE relates to the fact that no summarized evidence has yet been provided on a key twofold issue concerning their methodological quality (Terwee et al., 2018). Firstly, the extent to which evidence in support of their measurement properties is derived from studies following compliance with standards for study design requirements and preferred statistical methods. Secondly, the quality of the instruments in terms of the availability of evidence in support of their measurement properties. In this vein, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN; Mokkink, de Vet, et al., 2018; Prinsen et al., 2018; Terwee et al., 2018) constitutes a robust and widely employed tool that provides a comprehensive overview of the strengths and weaknesses of psychometric instruments across different populations and application conditions (Cassidy et al., 2018; Saini et al., 2019). Findings emerging from the employment of the COSMIN initiative's guidelines for the purpose of examining self-report measures assessing PE may result in the provision of comprehensive

evidence-based recommendations to researchers and health practitioners concerning the use of these instruments, as well as identifying shortcomings which may open up important avenues for future research.

Therefore, the purpose of the present study was to examine the methodological quality of the evidence on the measurement properties of the most used self-report instruments that assess the risk of PE. To address this goal, a systematic review was carried out focused on the most frequently used self-report instruments assessing PE to summarize, compare, and critically appraise both the methodological quality of the studies evaluating their measurement properties and the available evidence on their measurement properties (Mokkink, de Vet, et al., 2018; Prinsen et al., 2018; Terwee et al., 2018).

Method

This systematic review was conducted in accordance with the checklist from Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA; see Appendix A) (Page et al., 2021) and was registered on PROSPERO (CRD42021237106).

Locating studies

Electronic bibliographic databases MEDLINE, PsycINFO, Web of Science, Current Contents Connect, SciELO, and Dissertations & Theses Global were searched for eligible studies from inception to April 20, 2021. The search terms were chosen taking into account those proposed in previous recent reviews concerning PE (e.g., Alcaraz-Ibáñez et al., 2020, 2021), these being: "problematic exercise", "morbid exercise", "exercise addiction", "Exercise Addiction Inventory", "exercise dependence", "Exercise Dependence Scale", "compulsive exercise", "Compulsive Exercise Test", "compulsive physical activity", "obligatory exercise", "Obligatory Exercise Questionnaire", "commitment to exercise", "Commitment to Exercise Scale", "Exercise Dependence Questionnaire", and "excessive exercise" (see Appendix B for the full search strategy). No geographical or cultural restrictions were applied. Reference lists of retrieved studies were hand-searched for further potentially eligible studies.

The references of the retrieved studies were managed in EndnoteX9. Studies were independently selected by the first author and corresponding author (being respectively a postdoctoral researcher and a doctoral researcher, both with a publication record in the field of PE) in two stages by examining (i) their titles and abstracts, and (ii) their full-texts. Disagreements were discussed and resolved by consensus with the assistance of the third author (a professor with a publication record in the field of PE) if necessary.

Eligibility criteria

The review gathered data from studies either developing, validating and/or testing the psychometric properties of the main instruments (in terms of their number of citations, see Appendix C) proposed for assessing potentially problematic exercise behaviours. Based on the findings from previous reviews on PE (e.g., Alcaraz-Ibáñez et al., 2020, 2021) the six instruments were the Commitment to Exercise Scale (CES) (Davis et al., 1993), Compulsive Exercise Test (CET) (Taranis et al., 2011), Exercise Addiction Inventory (EAI) (Terry et al., 2004), Exercise Dependence Questionnaire (EDQ) (Ogden et al., 1997), the Exercise Dependence Scale (EDS) (Downs et al., 2004), and Obligatory Exercise Questionnaire (OEQ) (Pasman & Thompson, 1988). The decision of including these six instruments was further supported by the results of a *Google Scholar* search performed for all the 17 instruments previously identified within the field (Sicilia et al., 2021).

Inclusion criteria

Studies were considered eligible when the following three criteria were met: (i) they addressed the initial development or further psychometric validation work of any of the self-report PE instruments defined in the eligibility criteria or their different versions; (ii) they were written in English, French, Portuguese, or Spanish (the working languages of the review team); and (iii) they were published in a peer-reviewed journal.

Exclusion criteria

Studies were excluded on the basis of the following criteria: (i) only composite scores comprising two or more instruments assessing PE were provided so that individual scores were not available; (ii) no information concerning the measurement properties proposed by the COSMIN initiative was provided (e.g., Coen & Ogles, 1993); and (iii) the content of the version of the instrument being examined narrows the study of the phenomenon under consideration to a specific exercise or sport modality (e.g., dancing; Maraz et al., 2015). This latter exclusion criterion was due to wanting to focus on the instruments with the greatest potential for use in research and professional practice. Adopting this criterion also allowed the research team to handle a reasonable number of somewhat comparable instruments in terms of their focus, as well as go into greater detail on the features under examination.

Assessing the measurement properties of PE psychometric instruments

The evidence concerning the measurement properties of self-report instruments assessing symptoms of PE included in the review was evaluated following the recommendations of the COSMIN guidelines for systematic reviews of PROM (Patient-Reported Outcome Measures) (Prinsen et al., 2018). This involved conducting four different sets of evaluations for each instrument under review. More specifically, this concerned (i) risk

of bias of the retrieved studies; (ii) content validity; (iii) psychometric evidence (i.e., structural validity, internal consistency, reliability, measurement error, hypotheses testing for construct validity, cross-cultural validity/measurement invariance, criterion validity, and responsiveness); and (iv) quality grading of the evidence provided.

Risk of bias

The methodological quality of the retrieved studies was assessed using the COSMIN Risk of Bias Checklist (Mokkink, de Vet, et al., 2018). This checklist includes 116 items, scored using a four-point scale (i.e., 4 = 'very good', 3 = 'adequate', 2 = 'doubtful', 1 = 'inadequate'), and covering the following three areas: (i) content validity (i.e., PROM development and content validity; e.g., 'Is a clear description provided of the construct to be measured?'); (ii) internal structure (i.e., structural validity, internal consistency, and cross-cultural validity/measurement invariance; e.g., 'Was an internal consistency statistic calculated for each unidimensional scale or subscale separately?''); and (iii) remaining psychometric properties (i.e., reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness; e.g., 'For continuous scores: Was an intraclass correlation coefficient [ICC] calculated?'). The risk of bias assessment was independently conducted by the first author and corresponding author. Inter-coder percentage agreement ranged from 96% to 100% (ReCal software; Freelon, 2013). Disagreements between coders were discussed and resolved by consensus with the assistance of the third author if necessary.

Content validity assessment

Content validity (i.e., the degree to which the content of the instrument reflects the construct to be measured; Mokkink et al., 2010), was assessed according to the criteria proposed in the COSMIN initiative (Terwee et al., 2018). Therefore, evidence of content validity provided in each of the retrieved studies was rated as 'sufficient' (+), 'insufficient' (−), or 'indeterminate' (?) in terms of (i) relevance (i.e., the extent to which items are relevant for the construct to be assessed within a specific population and context), (ii) comprehensiveness (i.e., the extent to which key aspect of the construct to be assessed are not missed), and (iii) comprehensibility (i.e., the extent to which the items are interpreted as intended by the target population). In view of the ratings from each individual study, a 'sufficient' ('+'), 'insufficient' ('–'), or 'inconsistent' ('±') overall rating for content validity was assigned to each instrument. When the retrieved studies did not present sufficient information to assess the content validity of the instruments, the overall ratings were derived from the reviewers' ratings utilizing COSMIN criteria (Terwee et al., 2018). The content validity assessment was independently conducted by the first author and corresponding author. Inter-coder percentage agreement

ranged from 75% to 100% (ReCal software; Freelon, 2013). Disagreements between coders were discussed and resolved on a consensual basis with the assistance of a third author if necessary.

Measurement properties

The measurement properties of the six instruments were evaluated according to the checklist and the updated criteria for good measurement properties proposed within the content of the COSMIN initiative (Prinsen et al., 2018) (see Appendix D). Therefore, 'sufficient' ('+'), 'insufficient' ('–'), or 'indeterminate' ('?') ratings were assigned for those properties for which usable data were available in the retrieved studies (i.e., structural validity, internal consistency, reliability, hypotheses testing for construct validity, and cross-cultural validity/measurement invariance). The lack of usable data concerning measurement error, criterion validity, and responsiveness prevented the evaluation of these properties. The criteria employed for evaluating this set of measurement properties (Prinsen et al., 2018) were supplemented or refined as follows. Firstly, where structural validity was evaluated by Exploratory Factor Analysis (EFA) instead of by Confirmatory Factor Analysis (CFA), a 'sufficient' (+) rating was assigned when total variance explained was at least 50% (Terwee et al., 2012). Secondly, the cut-off point of ≥ .70 employed for assigning 'sufficient' (+) ratings for reliability in terms of temporal stability as expressed by intra-class correlation coefficient (ICC) (Prinsen et al., 2018) was also employed when this property was examined throughout Pearson's correlation coefficient (*r*). Thirdly, a 'sufficient' (+) rating in terms of hypotheses testing for construct validity was given when at least a 75% of the effect sizes of the correlations of interest fell within the following range: (i) >.50 when the scores of interest were derived somewhat similar from a theoretical perspective instruments (e.g., the EAI and the EDS-R); and (ii) between .20 and .49 when the scores of interest were derived from instruments assessing theoretically related but distinctive constructs (e.g., the EAI and the bulimia subscale of the EDI). Concerning the later, an 'indeterminate' (?) rating was assigned in cases where the lack of precision in the reporting of results made it impossible to locate the effect sizes under examination within the ranges described above (e.g., Plateau et al., 2014). Fourthly, an 'indeterminate' (?) rating in terms of cross-cultural validity/measurement invariance in cases where reasonable doubts exist on the consistency of the statistical analyses used with those commonly recommended for this purpose (Bowen & Masa, 2015; Milfont & Fischer, 2010).

The results of each study were summarized by instrument, each of them being assigned an overall 'sufficient' ('+'), 'insufficient' ('–'), 'inconsistent' (±), or 'indeterminate' ('?') rating. Consequently, an overall rating was given to each of the six instruments reviewed. The

evaluation of the measurement properties of the instruments was conducted by the corresponding author. The reliability of ratings was confirmed by the first author of the paper, who independently rated 20% of the cases under consideration.

Quality grading of the evidence

The quality of the evidence concerning the psychometric properties of the instruments under consideration was graded by using the modified Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach (Prinsen et al., 2018). By adopting this methodological approach, both individual and global ratings are respectively subject to the high, moderate, low, or very low levels of grading depending on whether the raters are very, moderately, limitedly, or poorly confident that the estimate of the measurement property is close to the true measurement property. These four levels can be subject to further downgrading in case of concerns with any of the following issues: (i) risk of bias (when the methodological quality of the studies is very doubtful or inadequate); (ii) inconsistent results (when aggregated results from different studies for a given instrument are inconsistent or hardly explainable); (iii) low sample size (when the results are derived from studies with sample sizes below 100); and (iv) indirectness (i.e., studies conducted in populations or contexts of use other than those of interest for the purpose of the systematic review are included). In the presence of contradictory grading ratings resulting from evidence derived from the employment of both CFA and EFA, the latter was ignored.

Recommendations for use

According to the proposal of the COSMIN initiative, the scales and subscales of the instruments under consideration were classified in descending order in terms of their recommendation for use into three categories (Mokkink, Prinsen, et al., 2018). Therefore, Category 'A' encompass those scales/subscales whose results can be trusted (i.e., those featuring both sufficient content validity at any quality evidence level and at least low-quality evidence for sufficient internal consistency). Category 'B' includes those scales/subscales which, although having some potential to be recommend for use, warrant further investigation for the purpose of verifying their quality (i.e., those having categorized not in 'A' or 'C'). Finally, Category 'C' encompass those scales/subscales which are not recommended for use (i.e., those with high quality evidence for an insufficient measurement property).

Results

Selection and description of studies

A total of 4,102 studies were identified from multiple database searches. As a result of

the study selection procedure (see forest plot in Figure 1), 48 papers were included in the systematic review. Modified versions were found for the CET (i.e., CET-A and CET-4F), the EAI (i.e., EAI-R and EAI-Y), the EDS (i.e., EDS-R), and the OEQ (i.e., OEQ-10, OEQ-11 and OEQ-R). The development of original and modified versions of the instruments under consideration were respectively addressed in six (11 studies) and eight (10 studies) of the retrieved papers. The main characteristics of the original measures and their modified versions are shown in Table 1.

Content validity

The detailed results of the risk of bias assessments of the PROM development studies of the PE instruments under review are shown in Figure 2 and Appendix F. These results showed the (i) 'inadequate' total ratings for the PROM design of all the 14 instruments under review, these being due to a lack of a clear definition of both the construct and the target population (with three out of 14 inadequate individual ratings in both cases) and the lack of matching between the target population and the population included in the development studies (with seven out of 14 inadequate individual ratings); (ii) complete absence of cognitive interview studies; and (iii) low use of external consultation when examining content validity (i.e., relevance, comprehensiveness, and comprehensibility), which limited to asking participants and experts about relevance in two and six cases out of the possible 14, respectively, with 'doubtful' ratings being assigned as it was not clear how the assessment was conducted. This latter circumstance meant that no total ratings in terms of risk of bias for the content validity were assigned (Terwee et al., 2018).

A total of 17 studies provided evidence concerning some of the components proposed in the COSMIN initiative (Terwee et al., 2018) for the purpose of assessing content validity. Here, all one-dimensional instruments under consideration (i.e., CES, EAI, EAI-R, EAI-Y and OEQ) were rated as 'indeterminate' (?). Most of the subscales of the multidimensional instruments under review were also rated 'indeterminate' (?), the exception to the above being the Mood Improvement subscale of CET, the Lack of Enjoyment subscale of the CET-4F, the Positive Reward subscale of the EDQ, the Withdrawal subscale of EDS, and the Preoccupation with Exercise subscale of the OEQ-R, whose content validity was rated as 'sufficient' (+).

As a result of the evidence grading process (GRADE approach), the quality of the evidence concerning content validity of the one-dimensional instruments and the sub-scales of the multidimensional ones were rated as 'low' or 'very low'. The exception was the EDS-R, with most of its subscales being rated as 'moderate' (see Table 2).

Measurement properties assessed

Commitment to Exercise Scale

A total of four studies comprising seven samples provided evidence concerning some of the seven measurement properties proposed in the COSMIN initiative (Prinsen et al., 2018) for the CES (see Appendix G). The evidence on the measurement properties of the CES was derived from samples of the general population (n = 2), regular exercisers (n = 3), athletes (n = 1), and individuals diagnosed with eating disorders (n = 1) of both sexes.

The available data allowed ratings to be assigned to the CES according to three of the psychometric properties under consideration (see Table 2). Firstly, in terms of structural validity, for which a 'sufficient' (+) rating with a very low quality of evidence was assigned. Secondly, in terms of internal consistency. In this case, the low individual ratings assigned on structural validity and their associated poor levels of quality of evidence (see Appendix G) meant that, in accordance with the criteria of the COSMIN initiative (Prinsen et al., 2018), an 'indeterminate' (?) rating for internal consistency should be assigned to the CES. Thirdly, in terms of hypothesis testing, for which an 'insufficient' (-) rating with a low quality of evidence was assigned. These findings implied that the CES was placed into the 'B' category in terms of recommendation for use.

Original and modified versions of the Compulsive Exercise Test

A total of 13 studies comprising 14 samples provided evidence concerning some of the seven measurement properties proposed in the COSMIN initiative (Prinsen et al., 2018) for the original (CET; 10 studies, 11 samples) and modified versions (CET-A, two studies, two samples; and CET-4F, one study, one sample) of the CET (see Appendix G). The evidence on the measurement properties of the CET was derived from samples consisting of clinical populations (n = 5) and the general population (n = 6), whose members were predominantly female (72.69%). Only three of the studies retrieved included 'regular exercisers', which were defined in the case with the most relaxed criteria as those who carried out at least one activity per week during the last month (Young et al., 2017). The evidence on the measurement properties of the CET-A was derived from two adult samples of competitive athletes (Plateau et al., 2014) and regular sport participants and exercisers (Limburg et al., 2019). Finally, the evidence on the measurement properties of the CET-4F was derived from a single adolescent sample diagnosed with an eating disorder.

The available data allowed ratings to be assigned to the CET and its modified versions according to three psychometric properties under consideration (see Table 2). Firstly, in terms of structural validity, for which 'insufficient' (-) ratings with a low quality of evidence (CET and the CET-A) and 'sufficient' (+) ratings with a very low quality of evidence (CET-4F) were

assigned. Secondly, in terms of internal consistency. In this case, the low individual ratings assigned on structural validity and their associated poor levels of quality of evidence meant that, in accordance with the criteria of the COSMIN initiative (Prinsen et al., 2018), 'indeterminate' (?) ratings should be assigned to all the full scales and subscales of the different versions of the CET. Thirdly, in terms of hypothesis testing, for which 'sufficient' (+) ratings were assigned to the full scale of the CET (with a low quality of evidence), as well as to the Avoidance subscales of both the CET (with a low quality of evidence) and the CET-A (with a moderate quality of evidence). These findings implied that all versions of the CET were placed into the 'B' category in terms of recommendation for use.

Original and modified versions of the Exercise Addiction Inventory

A total of 13 studies comprising 17 samples provided evidence concerning some of the seven measurement properties proposed in the COSMIN initiative (Prinsen et al., 2018) in the case of the original EAI (11 studies, 15 samples) and the modified versions (EAI-R; one study, one sample, and EAI-Y, one study, one sample) (see Appendix G). The evidence on the measurement properties of the EAI was derived from samples consisting of the general population (n = 1), regular exercisers (n = 8), mixed populations (i.e., those including regular exercisers with individuals who do not necessarily have such a condition; n = 1), and sports practitioners (n = 5). The evidence on the measurement properties of the EAI-R was derived from a single adult sample of regular exercisers (Szabo et al., 2019). Finally, the evidence on the measurement properties of the EAI-Y was derived from a single sample of adolescent regular exercisers (Lichtenstein et al., 2018).

The available data allowed ratings to be assigned to the different versions of the EAI according to five of the seven psychometric properties under consideration (see Table 2). Firstly, in terms of structural validity, for which both 'sufficient' (+) ratings with a low (EAI) and moderate (EAI-R) quality evidence, and 'insufficient' (-) ratings with a low quality of evidence (EAI-Y) were assigned. Secondly, in terms of internal consistency, for which 'sufficient' (+) ratings with a low quality of evidence were assigned both to the EAI and the EAI-R. In the case of the EAI-Y, the low individual ratings assigned on structural validity and their associated poor levels of quality of evidence meant that, in accordance with the criteria of the COSMIN initiative (Prinsen et al., 2018), an 'indeterminate' (?) rating in terms of internal consistency should be assigned to such an instrument. Thirdly, in terms of reliability, for which a 'sufficient' (+) rating with a low quality of evidence was assigned to the EAI. Fourthly, in terms of hypothesis testing, for which 'sufficient' (+) and 'indeterminate' (?) ratings were respectively assigned to the EAI and the EAI-Y, in both cases with a low quality of the evidence.

Fifthly, in terms of cross-cultural validity/measurement invariance, for which an 'insufficient' (-) rating with a low quality of evidence was assigned to the EAI. This poor quality of evidence was due to the fact that adequate techniques for the evaluation of this issue were employed in just one of the three studies providing evidence on the matter (Griffiths et al., 2015). These findings implied that all versions of the EAI were placed into the 'B' category in terms of recommendation for use.

Exercise Dependence Questionnaire

A total of three studies comprising three samples provided evidence concerning some of the measurement properties proposed in the COSMIN initiative (Prinsen et al., 2018) in the case of the EDQ (see Appendix G). The evidence on the measurement properties of the EDQ was derived from samples consisting of sports practitioners (n = 1) and regular exercisers (n = 2).

The available data allowed ratings to be assigned to the EDQ according to four of the seven psychometric properties under consideration (see Table 2). Firstly, in terms of structural validity, for which a 'sufficient' (+) rating with a moderate quality of evidence was assigned. Secondly, in terms of internal consistency, for which 'sufficient' (+) ratings were given to (i) the full scale of the EDQ (with a high quality of evidence), (ii) the Positive Reward and Health Reasons subscales (both with a high quality of evidence), and (iii) the Withdrawal and Weight Control subscales (with a low quality of evidence). Thirdly, in terms of reliability, for which 'sufficient' (+) and 'insufficient' (-) ratings were respectively assigned for the full scale and all the subscales of the EDQ, in both cases with a very low quality of evidence. Fourthly, in terms of hypothesis testing, for which a 'sufficient' (+) rating with a low quality of evidence was assigned to the full scale of the EDQ. None of the retrieved studies presented information that allowed for assigning ratings to the subscales of the EDQ on the basis of hypothesis testing. These findings implied that the Positive Reward subscale of the EDQ and the remaining subscales included in this instrument were respectively placed into the 'A' and 'B' categories in terms of recommendation for use.

Original and modified versions of the Exercise Dependence Scale

A total of 17 studies comprising 22 samples provided evidence concerning some of the measurement properties proposed in the COSMIN initiative (Prinsen et al., 2018) in the case of the original (EDS; four studies, four samples) and modified version (EDS-R; 13 studies, 18 samples) of the EDS (see Appendix G). The evidence on the measurement properties of the EDS was derived from samples consisting of undergraduate students. The evidence on the

measurement properties of the EDS-R was derived from samples consisting of the general population (n = 2), sports practitioners (n = 7), and regular exercisers (n = 9).

The available data allowed ratings to be assigned for the EDS and the EDS-R according to five of the seven psychometric properties under consideration (see Table 2). Firstly, in terms of structural validity, for which both 'indeterminate' (?) rating with a very low quality of evidence (EDS) and 'sufficient' (+) rating with a 'high' quality of evidence (EDS-R) were assigned. Secondly, in terms of internal consistency, for which 'sufficient' (+) ratings with a moderate quality of evidence were assigned to all the subscales of the EDS-R. The only exception to the above was the Time subscale, for which a 'insufficient' (-) rating with a very low quality of evidence was assigned. In the case of the EDS, the low individual ratings assigned on structural validity and their associated poor levels of quality of evidence meant that, in accordance with the criteria of the COSMIN initiative (Prinsen et al., 2018), 'indeterminate' (?) rating should be assigned to such an instrument. Thirdly, in terms of reliability, for which 'sufficient' (+) ratings with a very low and low quality of evidence were respectively assigned both to the full scale of the EDS and all the subscales of de EDS-R. Fourthly, in terms of hypothesis testing, for which 'sufficient' (+) ratings were given to the full scale the EDS (with a low quality of evidence), and the Tolerance, Intention effects, Lack of Control, Time, and Reduction in Other Activities subscales of the EDS-R (with a low quality of evidence). In turn, 'insufficient' (-) ratings with a low quality of evidence were assigned to both the full scale the EDS-R and the Withdrawal and Continuance subscales of the instrument. Fifthly, in terms of cross-cultural validity/measurement invariance, for which a 'sufficient' (+) rating with a moderate quality of evidence was assigned to the EDS-R. These findings implied that both the EDS and the EDS-R were placed into the 'B' category in terms of recommendation for use.

Original and modified versions of the Obligatory Exercise Questionnaire

A total of seven studies comprising seven samples provided evidence concerning some of the measurement properties proposed in the COSMIN initiative (Prinsen et al., 2018) in the case of the original (OEQ, two studies, two samples) and modified versions (OEQ-10, two studies, two samples; OEQ-11 one study, one sample; and OEQ-R, two studies, two samples) of the OEQ (see Appendix G). The evidence on the measurement properties of the different versions of the OEQ was derived from samples consisting of undergraduate students (n = 2), sports practitioners (n = 4), and the general population (n = 1).

The available data allowed ratings to be assigned to the OEQ and its modified versions according to three of the seven psychometric properties under consideration (see Table 2).

Firstly, in terms of structural validity, for which 'indeterminate' (?) ratings with a 'low' quality of evidence (OEQ), 'insufficient' (-) ratings with a 'low' quality of evidence (OEQ-10), and 'sufficient' (+) ratings with a 'very low' (OEQ-11) and 'low' quality of evidence (OEQ-R) were assigned. Secondly, in terms of internal consistency. In this case, the low individual ratings assigned on structural validity and their associated poor levels of quality of evidence meant that, in accordance with the criteria of the COSMIN initiative (Prinsen et al., 2018), 'indeterminate' (?) ratings for internal consistency should be assigned to the full scale and the subscales of the original and modified versions of the OEQ. Thirdly, in terms of hypothesis testing, for which 'insufficient' (-) ratings either with a high (for the three subscales of the OEQ-11), moderate (for the full scale and the Emotional Element of Exercise and Exercise Frequency and Intensity subscales of the OEQ-10), or a low quality of evidence (for the Exercise Preoccupation subscale of the OEQ-10) were assigned. In turn, a 'sufficient' (+) rating with a moderate quality of evidence was assigned to the full scale of the OEQ-11. These findings implied that three subscales of the OEQ-11 were placed into the 'C' category in terms of recommendation for use (i.e., the Exercise Fixation, Exercise Frequency, and Exercise Commitment, while the remaining ones corresponding to the different versions of the OEQ were placed into the 'B' category.

Discussion

The present systematic review examined the evidence on the methodological quality and the measurement properties of the six most widely used self-report instruments proposed for assessing the symptoms of PE. Data from 48 studies concerning six original instruments and their eight modified versions were included and subsequently evaluated according to the criteria proposed by the COSMIN initiative. The results obtained allow clear recommendations for use to be made only for a small number of the scales and sub-scales examined, which were favourable in one (i.e., the Positive Reward subscale of the EDQ) and unfavourable in three (i.e., the Exercise Fixation, Exercise Frequency, and Exercise Commitment subscales of the OEQ-11) of the cases. This limited nature of these recommendations is due to two main issues emerging in the process of summarizing and appraising the evidence on the measurement properties of the instruments under review. Firstly, the existence of numerous methodological shortcomings in the studies addressing the development of the instruments; and secondly, the limited evidence base in support of the majority of their measurement properties, with the most relevant being that concerning content validity. This implies that comprehensive recommendations concerning the use of most of the scales and subscales included in the instruments under review remains pending further investigation. Based on the shortcomings

identified, a number of suggestions are set out below that could be useful in addressing such research and, by extension, in facilitating progress in the field of PE research.

A first potential avenue for research to be proposed from the findings of the present study derives from the methodological shortcomings in the development studies of the instruments under consideration. This is particularly evident in light of the limited effort made in these studies to examine a key issue such as content validity in the early stages of the development process. This is an important flaw within the context of the appraisal methodology employed in the present study since the practical relevance of the remaining properties is largely dependent on the fact that there is prior evidence of content validity (i.e., that the content of the instrument has proven to be relevant, comprehensive, and comprehensible with respect to the construct of interest and the study population; Mokkink, Prinsen, et al., 2018; Terwee et al., 2018). This limitation is even more of a concern in view of the results of the content analysis that, as a result of the lack of evidence in this respect in the retrieved studies, had to be conducted by the team of reviewers of the present study (Terwee et al., 2018).

In particular, as the findings emerging from this analysis raise serious doubts concerning the extent to which the instruments under consideration are content-valid, particularly, in terms of comprehensiveness. Consequently, further studies specifically designed for the purpose of confirming or dispelling such doubts are warranted. The arguably questionable methodological soundness of the development studies of the instruments under review is also evident in view of the overall lack of matching between the populations included in these studies and the target population. It is worth noting that samples consisting of undergraduate students (CET, EAI, EDS, EDS-R, OEQ-11) or female (or mostly female) participants (CET, EAI-R, OEQ-11) have been employed for the purpose of developing instruments initially aimed to assess PE in broader populations in terms of demographic characteristics such as age, educational level, gender, practiced exercise modality, or exercise involvement. Admittedly, the existence of the methodological shortcomings concerning the lack of examination of content validity and matching between target and study populations identified above does not appear to be restricted to the development processes of instruments under examination. Indeed, this same two kinds of flaws have to some extent been reported by previous reviews that, aimed at examining other instruments, have also been conducted under the COSMIN initiative (e.g., Cassidy et al., 2018; Wittkowski et al., 2020).

However, this does not prevent the present authors from encouraging researchers proposing either adaptions of existing measures or hypothetical new ones in this field to consider study populations similar to the target populations for the purpose of examining

content validity before moving on to explore the remaining measurement properties. This process should include pilot assessments and cognitive interviews that, involving both participants from the target population and professionals, may provide evidence on the relevance, comprehensiveness, and comprehensibility of the instrument and, by extension, on its quality and potential practice utility (Mokkink, Prinsen, et al., 2018; Terwee et al., 2018).

Another major research avenue in the field under consideration stems from the limited evidence available in support of most of the measurement properties being examined (e.g., with regard to structural validity and cross cultural/measurement invariance). Here, the available evidence on these two measurement properties is noteworthy not only for being on the whole scarce but also for having been in quite a few cases obtained by employing somewhat questionable methodological approaches. Examples of the latter are the seemingly prevalent use of exploratory vs. confirmatory factor analysis techniques for the purpose of examining structural validity (e.g., in the case of the OEQ and its different versions) or the employment of statistical procedures of doubtful appropriateness when testing measurement invariance (mostly in this latter case, within the context of proposing translations of the instruments under consideration; e.g., Sauchelli et al., 2016; Sicilia et al., 2013, 2017; Sicilia & González-Cutre, 2011). These shortcomings imply that reasonable doubts still exist with regard to (i) the degree to which the scores of the instruments being reviewed are an adequate reflection of the dimensionality of the construct to be measured and (ii) whether the differences observed in those same scores across populations groups are due to measurement deficiencies (Bowen & Masa, 2015; Milfont & Fischer, 2010; Mokkink, Prinsen, et al., 2018). This is a pending significant issue in view of the ongoing debate on the dimensionality of some of the instruments under review (Chamberlain & Grant, 2020; Sicilia & González-Cutre, 2011). Additionally, in view of the research interest in examining the differences in the risk levels of PE across population groups, for instance, according to gender (Cook et al., 2013; Costa et al., 2013), eating disorder status (Trott et al., 2020), or exercise modality (Marques et al., 2019).

In sum, these shortcomings point to an urgent need for further examination of both structural validity and cross cultural/measurement invariance across population groups of interest of the scores derived from the instruments under consideration, a recommendation that is also applicable to the development of future instruments on the field. This recommendation is even more pertinent in the case of measurement invariance given the likely varying interpretations that different population groups (e.g., highly committed athletes vs. recreational exercisers) might make of the content of the instruments under review (Szabo et al., 2015).

The scarcity of evidence on the measurement properties on which the proposals here are grounded for future avenues of research is also noteworthy in the case of criterion validity, measurement error, and responsiveness. Admittedly, the fact that the instruments under review have not been developed for the purpose of assessing a common and clearly defined outcome (as would be the case for a nosographic entity delimited according to precise diagnostic criteria) makes it difficult to adopt an unequivocal 'gold standard' for the purpose of providing evidence of criterion validity. In the event that a consensus was reached on specific criteria that unequivocally qualify a particular pattern of exercise behaviour as problematic, these could be included in a clinical interview to be used as a gold standard. A complementary approach to that above could be proposed on the basis of an inherent feature of the many different expressions of problematic behaviours: the existence of harm/distress of a functionally impairing nature (Kardefelt-Winther et al., 2017). Moreover, objective indices (e.g., injury data) could be also used for the purpose of providing additional evidence on the validity of the self-report instruments assessing PE.

As far as measurement error and responsiveness are concerned, the fact that these properties remain totally unexplored to date implies that the possibility that changes over time in the scores derived from self-report instruments of PE are due to measurement deficiencies cannot be ruled out (Mokkink, Prinsen, et al., 2018). This is a worrying prospect in view of the research interest in examining these changes, for instance, within the context of testing the efficacy of interventions aimed at reducing the symptoms of PE (Outar et al., 2018). Consequently, further research providing evidence on measurement error and responsiveness of the scores of the self-report instruments of PE are warranted.

Strengths and limitations

A key strength of the present review is its comprehensiveness, which is evidenced by considering several relevant databases and publishing languages, so that near to 4,000 records were screened. Also worth noting is the fact that the review was pre-registered, which adds to the transparency of the research. Another key strength is the employment of the latest COSMIN criteria and standards for the purpose of conducting a rigorous and methodologically sound evaluation of the instruments of interest (Mokkink, Prinsen, et al., 2018; Terwee et al., 2018), which were supplemented where necessary in the light of the specificities of the studies retrieved. This translated, for example, into setting standards that allowed for rating structural validity when evidence from EFA instead of CFA was available or into rating the measurement properties reported in all the available studies irrespective of their methodological quality. The latter allowed – as in the case in previous reviews conducted under the COSMIN initiative

(Jewell et al., 2019; Wittkowski et al., 2020) – that a comprehensive overview of the instruments under consideration was gathered.

As regards the limitations of the present study, a first one arises from the specificities of the methodology employed. On the one hand, in view of both the very close nature of the cut-off points leading to assigning 'adequate' or 'inadequate' ratings and the 'worst case counts' rule penalizing single flaws when assigning ratings which have led some authors to suggest the likely underestimated nature of the results emerging from applying the COSMIN tool (Jewell et al., 2019; Wittkowski et al., 2020). On the other hand, the implementation of the COSMIN tool is not without some degree of subjectivity (Cassidy et al., 2018). However, it does not seem to be the case that these two issues had a major impact on the conclusions of the present study. In the first case, because most of the low ratings given seemed to be caused more by the absence of formal testing of several of the properties under examination than by the negative assessment of one of them. In the second case, in view of the high level of inter-coder agreement in the present study (75% to 100%), which is similar to those reported in previous reviews conducted utilizing the COSMIN initiative (e.g., Cassidy et al., 2018; Wittkowski et al., 2020). A second limitation derives from the focus on 'generic' instruments of PE (i.e., those not specifically proposed for a given exercise context or sport modality) adopted in the present study. Therefore, the possibility cannot be excluded that some of the instruments proposed within specific exercise contexts or sport modalities in the literature (e.g., Carmack & Martens, 1979; Smith & Hale, 2004) could have been proved to be more robust than those examined in the present study either in terms of the methodological quality of its development studies or according to the evidence in support of its measurement properties.

Conclusions

The present study is the first to conduct a structured and rigorous methodological evaluation of the evidence concerning the methodological quality and the measurement properties of the six main self-report instruments assessing the risk of PE using a robust research tool (COSMIN). The findings showed (i) a general lack of methodological quality in the development of the instruments under consideration, and (ii) a rather limited evidence base for their robustness in terms of validity and reliability. Indeed, from the 14 instruments and nearly 48 subscales examined, only one of these latter (i.e., the Positive Reward subscale of the EDQ) could be clearly recommended for use according to the proposal of the COSMIN initiative.

These results are relevant as they call into question the (i) accuracy and usefulness of the body of evidence obtained by employing these instruments, and (ii) advisability of persisting with its use, at least, until the issues identified in the present study have been adequately

addressed. Admittedly, significant progress in the study of the aetiology, consequences, and treatment of PE from a quantitative perspective will hardly be made in the absence of valid and reliable instruments. Even more so in the case of a complex phenomenon such as PE, whose comprehensive assessment may well require instruments specifically focused on the evaluation of its different manifestations. Consequently, obtaining further evidence concerning the measurement properties of currently available self-report instruments of PE as well as providing them in early stages of development for those to be proposed in the future (particularly in relation to content validity) should be a priority for research in the field.

References

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

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders. DSM-5* (5th ed.). American Psychiatric Association.

Bowen, N. K., & Masa, R. D. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research*, *6*(2), 229–249. https://doi.org/10.1086/681607

Carmack, M. A., & Martens, R. (1979). Measuring commitment to running: A survey of runners' attitudes and mental states. *Journal of Sport Psychology*, *1*(1), 25–42.

Cassidy, S. A., Bradley, L., Bowen, E., Wigham, S., & Rodgers, J. (2018). Measurement properties of tools used to assess suicidality in autistic and general population adults: A systematic review. *Clinical Psychology Review*, *62*, 56–70. https://doi.org/10.1016/j.cpr.2018.05.002

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

Coen, S. P., & Ogles, B. M. (1993). Psychological characteristics of the obligatory runner: A critical examination of the anorexia analogue hypothesis. *Journal of Sport & Exercise Psychology*, *15*, 338–354. https://doi.org/10.1123/jsep.15.3.338

Cook, B., Hausenblas, H. A., & Rossi, J. (2013). The moderating effect of gender on ideal-weight goals and exercise dependence symptoms. *Journal of Behavioral Addictions*, *2*(1), 50–55. https://doi.org/10.1556/JBA.1.2012.010

Costa, S., Hausenblas, H. A., Oliva, P., Cuzzocrea, F., & Larcan, R. (2013). The role of age, gender, mood states and exercise frequency on exercise dependence. *Journal of Behavioral Addictions*, *2*(4), 216–223. https://doi.org/10.1556/JBA.2.2013.014

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

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

Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, *8*(1), 10–16.

Griffiths, M. D., Urbán, R., Demetrovics, Z., Lichtenstein, M. B., de la Vega, R., Kun, B., Ruiz-Barquín, R., Youngman, J., & 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

Jewell, T., Gardner, T., Susi, K., Watchorn, K., Coopey, E., Simic, M., Fonagy, P., & Eisler, I. (2019). Attachment measures in middle childhood and adolescence: A systematic review of measurement properties. *Clinical Psychology Review*, *68*, 71–82. https://doi.org/10.1016/j.cpr.2018.12.004

Juwono, I. D., & Szabo, A. (2020). 100 cases of exercise addiction: More evidence for a widely researched but rarely identified dysfunction. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00264-6

Kardefelt-Winther, D., Heeren, A., Schimmenti, A., van Rooij, A., Maurage, P., Carras, M., Edman, J., Blaszczynski, A., Khazaal, Y., & Billieux, J. (2017). How can we conceptualize behavioural addiction without pathologizing common behaviours? *Addiction*, *112*(10), 1709–1715. https://doi.org/10.1111/add.13763

Lichtenstein, M. B., Griffiths, M. D., Hemmingsen, S. D., & Støving, R. K. (2018). Exercise addiction in adolescents and emerging adults - Validation of a youth version of the Exercise Addiction Inventory. *Journal of Behavioral Addictions*, *7*(1), 117–125. https://doi.org/10.1556/2006.7.2018.01

Maraz, A., Urbán, R., Griffiths, M. D., & Demetrovics, Z. (2015). An empirical investigation of dance addiction. *PloS One*, *10*(5), 1–13. https://doi.org/10.1371/journal.pone.0125988

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

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111–121. https://doi.org/10.21500/20112084.857

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M.,

& Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs) – user manual*. https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Strat- ford, 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

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

Outar, L., Turner, M. J., Wood, A. G., & Lowry, R. (2018). "I need to go to the gym": Exploring the use of rational emotive behaviour therapy upon exercise addiction, irrational and rational beliefs. *Performance Enhancement and Health*, *6*(2), 82–93. https://doi.org/10.1016/j.peh.2018.05.001

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Pasman, L. N., & 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.

Plateau, C. R., Shanmugam, V., Duckham, R. L., Goodwin, H., Jowett, S., Brooke-Wavell, K. S. F., Laybourne, A., Arcelus, J., & Meyer, C. (2014). Use of the Compulsive Exercise Test with athletes: Norms and links with eating psychopathology. *Journal of Applied Sport Psychology*, *26*(3), 287–301. https://doi.org/10.1080/10413200.2013.867911

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported

outcome measures. *Quality of Life Research*, *27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Saini, S. M., Hoffmann, C. R., Pantelis, C., Everall, I. P., & Bousman, C. A. (2019). Systematic review and critical appraisal of child abuse measurement instruments. *Psychiatry Research*, *272*, 106–113. https://doi.org/10.1016/j.psychres.2018.12.068

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

Sicilia, A., Alcaraz-Ibáñez, M., Paterna, A., & Griffiths, M. D. (2022). A review of the components of problematic exercise in psychometric assessment instruments. *Frontiers in Public Health*. https://doi.org/10.3389/fpubh.2022.839902

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), 176–165. 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

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 Behavioral Addictions*, *10*(1), 4–20. https://doi.org/10.1556/2006.2021.00019

Smith, D., & Hale, B. (2004). Validity and factor structure of the Bodybuilding Dependence Scale. *British Journal of Sports Medicine*, *38*(2), 177–181. https://doi.org/10.1136/bjsm.2002.003269

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, R., Mervó, B., & Demetrovics, Z. (2015). Methodological and conceptual limitations in exercise addiction research. *Yale Journal of Biology and Medicine*, *88*(3), 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

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & De Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research*, *21*(4), 651–657. https://doi.org/10.1007/s11136-011-9960-1

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research*, *27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Thompson, W. R., Sallis, R., Joy, E., Jaworski, C. A., Stuhr, R. M., & Trilk, J. L. (2020). Exercise is medicine. *American Journal of Lifestyle Medicine*, *14*(5), 511–523. https://doi.org/10.1177/1559827620912192

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

Wittkowski, A., Vatter, S., Muhinyi, A., Garrett, C., & Henderson, M. (2020). Measuring bonding or attachment in the parent-infant-relationship: A systematic review of parent-report assessment measures, their psychometric properties and clinical utility. *Clinical*

Psychology Review, *82*, 101906. https://doi.org/10.1016/j.cpr.2020.101906

World Health Organization (2019). *International classification of diseases: ICD-11 for mortality and morbidity statistics*. https://icd.who.int/dev11/l-m/en

Young, S., Touyz, S., Meyer, C., Arcelus, J., Rhodes, P., Madden, S., Pike, K., Attia, E., Crosby, R. D., Wales, J., & Hay, P. (2017). Validity of exercise measures in adults with anorexia nervosa: The EDE, Compulsive Exercise Test and other self-report scales. *International Journal of Eating Disorders*, *50*(5), 533–541. https://doi.org/10.1002/eat.22633

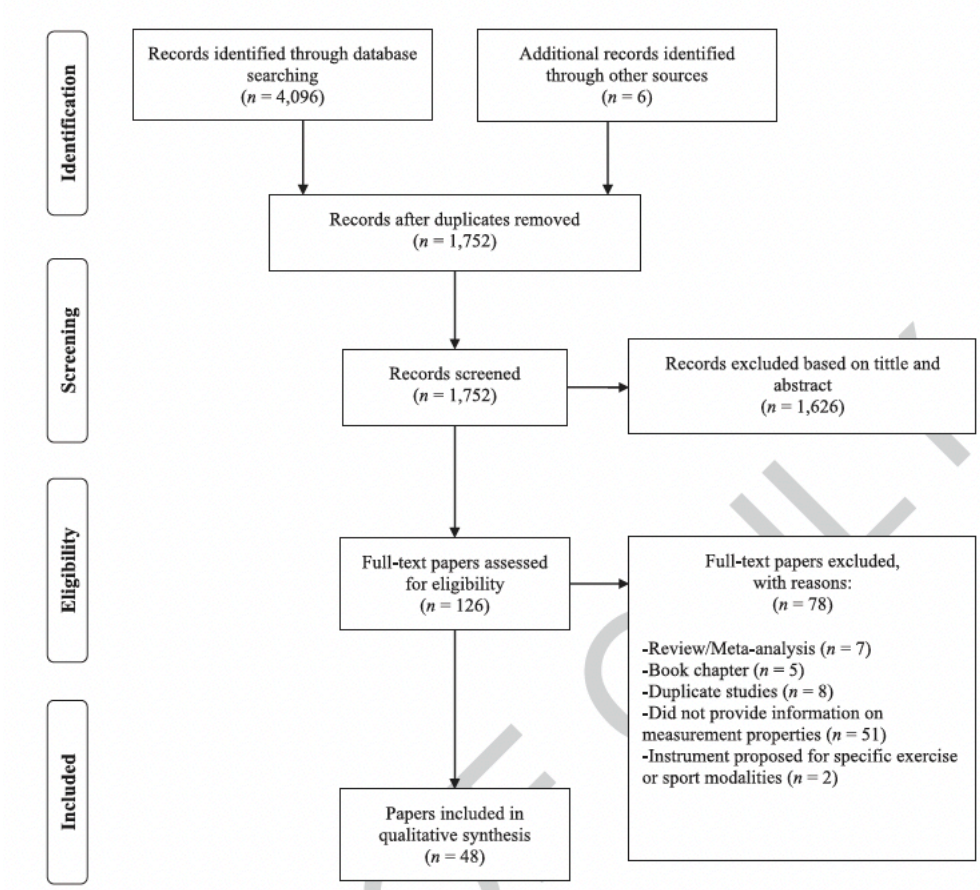

Figure 1. PRISMA flow diagram of study.

Table 1. Main characteristics of the instruments of problematic exercise included in the review.

Measure (Authors)	Modified from	Focus of the measure (Aim of tool)	Study population	Number of items: Total	Number of items: Subscales	Response options	Recall period	Scoring
CES (Davis et al., 1993)	None	Core features believed to characterize excessive exercisers	Male/female Canadian exercisers recruited from recreational facilities at universities, health and fitness clubs/associations	8	None, one-dimensional[1]	VAS (155 mm)	None (trait level)	Continuous: Higher scores mean greater risk of PE[2]
CET (Taranis et al., 2011)	None	Primary factors operating in the maintenance of excessive exercise within the eating disorders domain	Young women from UK and Australian universities engaged in regular exercise or sport over the last 4 weeks	24	Avoidance and rule-driven behaviour (8); Weight control exercise (5); Mood improvement (5); Lack of exercise enjoyment (3); Exercise rigidity (3)	6-point Likert scale	None (trait level)	Continuous: Higher scores mean greater risk of PE[3]
CET-A (Plateau et al., 2014)	CET	Similar to the CET	Male/female competitive athletes recruited from sports clubs and teams at UK universities	15	Avoidance of Negative Affect (6); Weight Control Exercise (4); Mood Improvement (5)	6-point Likert scale	None (trait level)	Similar to the CET
CET-4F (Swenne 2016)	CET	Similar to the CET	Both sedentary and physically active male/female Swedish adolescents diagnosed with an eating disorder	21	Avoidance and rule-driven behaviour (8); Weight control exercise (5); Mood improvement (5); Lack of exercise enjoyment (3)	6-point Likert scale	None (trait level)	Similar to the CET
EAI (Terry et al., 2004)	None	Theoretical components of behavioural addictions (Griffiths 1996)	Male/female UK university students reporting regular participation in exercise	6	None, one-dimensional	5-point Likert scale	None (trait level)	(a) Continuous: Higher scores mean greater risk of PE[4] (b) Screening: At-risk (24-30), Symptomatic (13-20), Asymptomatic (0-12)
EAI-R (Szabo et al., 2019)	EAI	Similar to the EAI	Mainly male (88%) adults that reported exercising at least three times per week for at least 30 min each time	6	None, one-dimensional	6-point Likert scale	None (trait level)	Similar to the EAI
EAI-Y (Lichtenstein et al., 2018)	EAI	Similar to the EAI	Male/female adolescents and young adults expected to perform regular exercise including sport school students, attendees at fitness centres, and eating disorder patients	6	None, one-dimensional	5-point Likert scale	None (trait level)	Similar to the EAI
EDQ (Ogden et al., 1997)	None	Feelings and cognitions about exercise behaviour reported by self-declared as "exercise addicts"	Male/female young adults that reported exercising more than 4 hours/week, recruited from sports clubs, leisure centres, and ads in magazines	29	Interference with life (5); Positive reward (4); Withdrawal (4); Exercise for weight control (4); Insight into problem (4); Exercise for social reasons (3); Exercise for health reason (3); Stereotyped behaviour (2)	7-point Likert scale	Past month	Continuous: Higher scores mean greater risk of PE[5]
EDS (Hausenblas & Downs 2002)	None	Criteria for substance dependence (APA, 1994) adapted to the context of exercise	Male/female U.S. university students	30	Tolerance (4); Withdrawal (12); Intention effects (2); Lack of control (3); Time (2); Reduction in other activities (4); Continuance (3)	6-point Likert scale	Past 3-month	(a) Continuous: Higher scores mean greater risk of PE[5] (b) Screening: At-risk (scores of 5-6 on $\geq$ 3 criteria); Symptomatic (scores $\geq$ 3 on $\geq$ 3 criteria without meeting the at-risk condition); Asymptomatic (scores of 1-2 on $\geq$ 3 criteria)
EDS-R (Downs et al., 2004)	EDS	Similar to the EDS	Male/female U.S. university students enrolled in fitness classes that exercised at least thrice a week for approximately 1 hr per session	27	Tolerance (3); Withdrawal (3); Intention effects (3); Lack of control (3); Time (3); Reduction in other activities (3); Continuance (3)	6-point Likert scale	Past 3-month	Similar to the EDS
	None			20	One-dimensional			

Table 1. Continued.

Measure (Authors)	Modified from	Focus of the measure (Aim of tool)	Study population	Number of items: Total	Number of items: Subscales	Response options	Recall period	Scoring
OEQ (Pasman & Thompson 1988)		Subjective need to engage in repetitive exercise behaviours	Male/female runners, weightlifters and sedentary controls			4-point Likert scale	None (trait level)	Continuous: Higher scores mean greater risk of PE[6]
OEQ-10 (Steffen & Brehm 1999)	OEQ	Similar to the OEQ	Male/female U.S. high school students	10	Emotional element of exercise (4); Exercise frequency and intensity (4); Exercise preoccupation (2)	4-point Likert scale	None (trait level)	Similar to the OEQ
OEQ-11 (Ackard et al., 2002)	OEQ	Similar to the OEQ	Female U.S. university students	11	Exercise Fixation (5); Exercise Frequency (3); Exercise Commitment (3)	4-point Likert scale	None (trait level)	Similar to the OEQ
OEQ-R (Duncan et al., 2012)	OEQ	Similar to the OEQ	Male/female Canadian regular exercisers	10	Preoccupation with exercise (4); Exercise behaviour (3); Exercise emotionality (3)	4-point Likert scale	None (trait level)	Similar to the OEQ

Note. CES = Commitment to Exercise Scale; CET = Compulsive Exercise Test; CET-A = Compulsive Exercise Test-Athletes; CET-4F = Compulsive Exercise Test (4-factors); EAI = Exercise Addiction Inventory; EAI-R = Exercise Addiction Inventory-Revised; EAI-Y = Exercise Addiction Inventory-Youth; EDQ = Exercise Dependence Questionnaire; EDS = Exercise Dependence Scale; EDS-R = Exercise Dependence Scale-Revised; OEQ = Obligatory Exercise Questionnaire; OEQ-10 = Obligatory Exercise Questionnaire-10-items; OEQ-11 = Obligatory Exercise Questionnaire-11-items; OEQ-R = Obligatory Exercise Questionnaire-Revised; PE = Problematic exercise.

[1] In the absence of evidence that would allow a clear delimitation of the items belonging to each of the two factors drawn for the CES (i.e., Obligatory exercise and Pathological exercise; Davis et al., 1993), this instrument was treated as one-dimensional for the purposes of the present study.

[2] Termed in this instrument as "excessive exercise".

[3] Termed in this instrument as "compulsive exercise".

[4] Termed in this instrument as "exercise addiction".

[5] Termed in this instrument as "exercise dependence".

[6] Termed in this instrument as "obligatory exercise".

Measure	Design: General design requirements: Clear construct	Clear origin of the construct	Clear target population	Clear context of use	Developed in a sample representing the target population	Design: Concept elicitation	Design: Total PROM design	Cognitive interview (CI) study: General design requirements: CI study performed in a sample representative of the target population	Comprehensibility	Comprehensiveness	Total CI study	Content validity: Asking participants: Relevance	Comprehensiveness	Comprehensibility	Asking experts: Relevance	Comprehensiveness
CES	V	D	I	D	D	-	I	-	-	-	-	-	-	-	D	-
CET (Taranis et al. 2011)	V	D	V	D	I	-	I	-	-	-	-	D	-	-	D	-
CET-A (Plateau et al. 2014)	V	D	V	D	D	I	I	-	-	-	-	-	-	-	-	-
CET-4F (Limburg et al. 2019)	V	D	V	D	A	I	I	-	-	-	-	-	-	-	-	-
EAI (Terry et al. 2004)	V	V	V	V	D	I	I	-	-	-	-	-	-	-	D	-
EAI-R (Szabo et al. 2019)	V	V	V	D	I	-	I	-	-	-	-	-	-	-	-	-
EAI-Y (Lichtenstein et al. 2018)	V	V	V	D	D	I	I	-	-	-	-	-	-	-	D	-
EDQ (Ogden et al. 1997)	I	D	I	D	I	-	I	-	-	-	-	-	-	-	-	-
EDS (Hausenblas & Downs 2002)	V	V	V	D	D	I	I	-	-	-	-	D	-	-	D	-
EDS-R (Symons-Downs et al. 2004)	V	V	V	V	I	-	I	-	-	-	-	-	-	-	D	-
OEQ (Pasman & Thompson 1988)	I	D	I	D	I	-	I	-	-	-	-	-	-	-	-	-
OEQ-10 (Steffen & Brehm 1999)	V	D	V	D	I	-	I	-	-	-	-	-	-	-	-	-
OEQ-11 (Ackard et al. 2002)	V	D	V	D	I	-	I	-	-	-	-	-	-	-	-	-
OEQ-R (Duncan et al. 2012)	I	D	V	D	D	-	I	-	-	-	-	-	-	-	-	-

Note. V = Very good; A = Adequate; D = Doubtful; I = Inadequate; - = Not addressed/reported in the PROM development studies.

Figure 2. Results of the risk of bias of the PROM development studies of problematic exercise.

Table 2. Result of the ratings of the psychometric properties and the quality of the evidence of the instruments of problematic exercise.

		Content validity														Cross-cultural validity/ measurement invariance	
		Relevance		Comprehensiveness		Comprehensibility		Structural validity		Internal consistency		Reliability		Hypothesis testing			
Measure	Full scale/ Subscale	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence
CES	Full scale	[±]	Very low	[-]	Low	[±]	Low	[+]	Very low	[?]	High	Not evaluated[3]		[-]	Low	Not evaluated[3]	
CET	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[-]	Low	[?]	High	Not evaluated[3]		[+]	Low	Not evaluated[3]	
CET	Avoidance	[+]	Very low	[+]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[+]	Low	Not evaluated[3]	
CET	Weight control	[+]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[-]	Low	Not evaluated[3]	
CET	Mood improvement	[+]	Very low	[+]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
CET	Lack of enjoyment	[+]	Very low	[+]	Low	[+]	Low	Not applicable[2]		[?]	High	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
CET	Rigidity	[+]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
CET-A	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[-]	Low	[?]	High	Not evaluated[3]		[?]	Very low	Not evaluated[3]	
CET-A	Avoidance	[+]	Low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[+]	Moderate	Not evaluated[3]	
CET-A	Weight control	[+]	Low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
CET-A	Mood improvement	[+]	Low	[+]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
CET-4F	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[+]	Very low	Not evaluated[2]		Not evaluated[3]		Not evaluated[2]		Not evaluated[3]	
CET-4F	Avoidance	[+]	Low	[+]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[?]	Low	Not evaluated[3]	
CET-4F	Weight control	[+]	Low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[?]	Low	Not evaluated[3]	
CET-4F	Mood improvement	[+]	Low	[+]	Low	[±]	Very low	Not applicable[2]		[?]	High	Not evaluated[3]		[?]	Low	Not evaluated[3]	
CET-4F	Lack of enjoyment	[+]	Low	[+]	Low	[+]	Low	Not applicable[2]		[?]	High	Not evaluated[3]		[?]	Low	Not evaluated[3]	
EAI	Full scale	[+]	Very low	[-]	Low	[±]	Very low	[+]	Low	[+]	Low	[+]	Low	[+]	Low	[-]	Low
EAI-R	Full scale	[±]	Very low	[-]	Low	[±]	Very low	[+]	Moderate	[+]	Low	Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EAI-Y	Full scale	[±]	Very low	[-]	Low	[±]	Very low	[-]	Low	[?]	Low	Not evaluated[3]		[?]	Low	Not evaluated[3]	
EDQ	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[+]	Moderate	[+]	High	[+]	Very low	[+]	Very low	Not evaluated[3]	
EDQ	Interference	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[-]	Very low	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDQ	Positive reward	[+]	Low	[+]	Low	[+]	Low	Not applicable[2]		[+]	High	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDQ	Withdrawal	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[+]	Low	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDQ	Weight control	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[+]	Low	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDQ	Insight into problem	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[-]	Very low	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDQ	Social reasons	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[-]	Very low	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDQ	Health reasons	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[+]	High	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDQ	Stereotyped behaviour	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[-]	High	[-]	Very low	Not evaluated[3]		Not evaluated[3]	
EDS	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[?]	Very low	[?]	Low	[+]	Very low	[+]	Low	Not evaluated[3]	
EDS	Tolerance	[+]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EDS	Withdrawal	[+]	Very low	[+]	Low	[+]	Low	Not applicable[2]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EDS	Intention effects	[+]	Very low	[+]	Low	[±]	Very low	Not applicable[2]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EDS	Lack of control	[+]	Very low	[-]	Low	[+]	Low	Not applicable[2]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EDS	Time	[±]	Very low	[-]	Low	[+]	Low	Not applicable[2]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EDS	Reduction in other activities	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EDS	Continuance	[+]	Very low	[-]	Low	[+]	Low	Not applicable[2]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
EDS-R	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[+]	High	Not evaluated[3]		Not evaluated[3]		[-]	Moderate	[+]	Moderate
EDS-R	Tolerance	[+]	Low	[-]	Moderate	[±]	Low	Not applicable[2]		[+]	Moderate	[+]	Low	[+]	Low	Not evaluated[3]	

Table 2. Continued.

		Content validity														Cross-cultural validity/ measurement invariance	
		Relevance		Comprehensiveness		Comprehensibility		Structural validity		Internal consistency		Reliability		Hypothesis testing			
Measure	Full scale/ Subscale	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence
EDS-R	Withdrawal	[+]	Low	[-]	Moderate	[+]	Moderate	Not applicable[2]		[+]	Moderate	[+]	Low	[-]	Moderate	Not evaluated[3]	
EDS-R	Intention effects	[+]	Low	[-]	Moderate	[+]	Moderate	Not applicable[2]		[+]	Moderate	[+]	Low	[+]	Low	Not evaluated[3]	
EDS-R	Lack of control	[+]	Low	[-]	Moderate	[+]	Moderate	Not applicable[2]		[+]	Moderate	[+]	Low	[+]	Low	Not evaluated[3]	
EDS-R	Time	[+]	Low	[-]	Moderate	[+]	Moderate	Not applicable[2]		[-]	Very low	[+]	Low	[+]	Low	Not evaluated[3]	
EDS-R	Reduction in other activities	[±]	Very low	[-]	Moderate	[±]	Low	Not applicable[2]		[+]	Moderate	[+]	Very low	[+]	Low	Not evaluated[3]	
EDS-R	Continuance	[+]	Low	[-]	Moderate	[+]	Moderate	Not applicable[2]		[+]	Moderate	[+]	Very low	[-]	Moderate	Not evaluated[3]	
OEQ	Full scale	[±]	Very low	[-]	Low	[-]	Low	[?]	Low	[?]	Moderate	Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
OEQ-10	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[-]	Low	[?]	Moderate	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
OEQ-10	Emotional element of exercise	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	Moderate	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
OEQ-10	Exercise frequency and intensity	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	Moderate	Not evaluated[3]		[-]	Moderate	Not evaluated[3]	
OEQ-10	Exercise preoccupation	[±]	Very low	[-]	Low	[+]	Low	Not applicable[2]		[?]	Moderate	Not evaluated[3]		[-]	Low	Not evaluated[3]	
OEQ-11	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[+]	Very low	[?]	Low	Not evaluated[3]		[+]	Moderate	Not evaluated[3]	
OEQ-11	Exercise fixation	[±]	Very low	[-]	Low	[-]	Low	Not applicable[2]		[?]	Low	Not evaluated[3]		[-]	High	Not evaluated[3]	
OEQ-11	Exercise frequency	[+]	Low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	Low	Not evaluated[3]		[-]	High	Not evaluated[3]	
OEQ-11	Exercise commitment	[+]	Low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	Low	Not evaluated[3]		[-]	High	Not evaluated[3]	
OEQ-R	Full scale	Not applicable[1]		Not applicable[1]		Not applicable[1]		[+]	Low	Not evaluated[3]		Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
OEQ-R	Preoccupation with exercise	[+]	Low	[+]	Low	[+]	Low	Not applicable[2]		[?]	Low	Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
OEQ-R	Exercise behaviour	[+]	Low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	Low	Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	
OEQ-R	Exercise emotionality	[±]	Very low	[-]	Low	[±]	Very low	Not applicable[2]		[?]	Low	Not evaluated[3]		Not evaluated[3]		Not evaluated[3]	

Note. [+] = sufficient; [-] = insufficient; [?] = indeterminate; [±] = inconsistent; CES = Commitment to Exercise Scale; CET = Compulsive Exercise Test; CET-A = Compulsive Exercise Test-Athletes; CET-4F = Compulsive Exercise Test (4-factors); EAI = Exercise Addiction Inventory; EAI-R = Exercise Addiction Inventory-Revised; EAI-Y = Exercise Addiction Inventory-Youth; EDQ = Exercise Dependence Questionnaire; EDS = Exercise Dependence Scale; EDS-R = Exercise Dependence Scale-Revised; OEQ = Obligatory Exercise Questionnaire; OEQ-10 = Obligatory Exercise Questionnaire-10-items; OEQ-11 = Obligatory Exercise Questionnaire-11-items; OEQ-R = Obligatory Exercise Questionnaire-Revised.

[1] In multidimensional instruments, evaluations concerning content validity are applied at the subscale level.

[2] Structural validity is evaluated at the full-scale level.

[3] Due to data unavailability (i.e., the psychometric property under consideration has not been examined in any of the available studies).

Table 3. Results of the ratings of content validity by study.

			Relevance			Comprehensiveness			Comprehensibility		
Study	Measure	Full scale/Subscale	PROM development study	Content validity study	Reviewers' ratings	PROM Development study	Content validity study	Reviewers ratings	PROM Development study	Content validity study	Reviewers ratings
Davis et al. (1993)	CES	Full scale	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Teixeira et al. (2011)	CES	Full scale	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[+]	Not applicable[1]
Taranis et al. (2011)	CET	Avoidance	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[±]
Taranis et al. (2011)	CET	Weight control	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Taranis et al. (2011)	CET	Mood improvement	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[±]
Taranis et al. (2011)	CET	Lack of enjoyment	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]
Taranis et al. (2011)	CET	Rigidity	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Plateau et al. (2014)	CET-A	Avoidance	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Plateau et al. (2014)	CET-A	Weight control	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Plateau et al. (2014)	CET-A	Mood improvement	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Swenne (2016)	CET-4F	Avoidance	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[±]
Swenne (2016)	CET-4F	Weight control	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Swenne (2016)	CET-4F	Mood improvement	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[±]
Swenne (2016)	CET-4F	Lack of enjoyment	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]
Terry et al. (2004)	EAI	Full scale	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Szabo et al. (2019)	EAI-R	Full scale	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Lichtenstein et al. (2018)	EAI-Y	Full scale	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ogden et al. (1997)	EDQ	Interference	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ogden et al. (1997)	EDQ	Positive reward	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]
Ogden et al. (1997)	EDQ	Withdrawal	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ogden et al. (1997)	EDQ	Weight control	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ogden et al. (1997)	EDQ	Insight into problem	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ogden et al. (1997)	EDQ	Social reasons	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ogden et al. (1997)	EDQ	Health reasons	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ogden et al. (1997)	EDQ	Stereotyped behaviour	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Hausenblas & Downs (2002)	EDS	Tolerance	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Hausenblas & Downs (2002)	EDS	Withdrawal	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]
Hausenblas & Downs (2002)	EDS	Intention effects	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[±]
Hausenblas & Downs (2002)	EDS	Lack of control	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Hausenblas & Downs (2002)	EDS	Time	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Hausenblas & Downs (2002)	EDS	Reduction in other activities	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Hausenblas & Downs (2002)	EDS	Continuance	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[+]
Downs et al. (2004)	EDS-R	Tolerance	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[+]
Downs et al. (2004)	EDS-R	Withdrawal	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Downs et al. (2004)	EDS-R	Intention effects	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[+]
Downs et al. (2004)	EDS-R	Lack of control	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[+]
Downs et al. (2004)	EDS-R	Time	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[+]

Table 3. Continued.

			Relevance			Comprehensiveness			Comprehensibility		
Study	Measure	Full scale/Subscale	PROM development study	Content validity study	Reviewers' ratings	PROM Development study	Content validity study	Reviewers ratings	PROM Development study	Content validity study	Reviewers ratings
Downs et al. (2004)	EDS-R	Reduction in other activities	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Downs et al. (2004)	EDS-R	Continuance	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[+]
Pujals et al. (2018)	EDS-R	Tolerance	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[+]	Not applicable[1]
Pujals et al. (2018)	EDS-R	Withdrawal	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[+]	Not applicable[1]
Pujals et al. (2018)	EDS-R	Intention effects	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[+]	Not applicable[1]
Pujals et al. (2018)	EDS-R	Lack of control	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[+]	Not applicable[1]
Pujals et al. (2018)	EDS-R	Continuance	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[?]	Not applicable[1]	Not applicable[1]	[+]	Not applicable[1]
Pasman & Thompson (1988)	OEQ	Full scale	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Steffen & Brehm (1999)	OEQ-10	Emotional element of exercise	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Steffen & Brehm (1999)	OEQ-10	Exercise frequency and intensity	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Steffen & Brehm (1999)	OEQ-10	Exercise preoccupation	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[+]
Ackard et al. (2002)	OEQ-11	Exercise fixation	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[-]
Ackard et al. (2002)	OEQ-11	Exercise frequency	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Ackard et al. (2002)	OEQ-11	Exercise commitment	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Duncan et al. (2012)	OEQ-R	Preoccupation with exercise	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[+]
Duncan et al. (2012)	OEQ-R	Exercise behaviour	[?]	Not applicable[1]	[+]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]
Duncan et al. (2012)	OEQ-R	Exercise emotionality	[?]	Not applicable[1]	[±]	[?]	Not applicable[1]	[-]	[?]	Not applicable[1]	[±]

Note. [+] = sufficient; [-] = insufficient; [?] = indeterminate; [±] = inconsistent; CES = Commitment to Exercise Scale; CET = Compulsive Exercise Test; CET-A = Compulsive Exercise Test-Athletes; CET-4F = Compulsive Exercise Test (4-factors); EAI = Exercise Addiction Inventory; EAI-R = Exercise Addiction Inventory-Revised; EAI-Y = Exercise Addiction Inventory-Youth; EDQ = Exercise Dependence Questionnaire; EDS = Exercise Dependence Scale; EDS-R = Exercise Dependence Scale-Revised; OEQ = Obligatory Exercise Questionnaire; OEQ-10 = Obligatory Exercise Questionnaire-10-items; OEQ-11 = Obligatory Exercise Questionnaire-11-items; OEQ-R = Obligatory Exercise Questionnaire-Revised.

[1] Studies are specifically assessed in terms of relevance, comprehensiveness, and comprehensibility depending on whether they address the PROM development or the subsequent content validity of the instrument (e.g., reviewers' ratings are not needed in content validity studies) (e.g., reviewers' ratings are not needed in content validity studies; Terwee et al., 2018).

Table 4. Result of the psychometric properties by study.

				Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
Study	Measure	Full scale/ Subscale	Language	*n*	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Davis et al. (1993)	CES	Full scale	English	185	EFA	[?]	Inadequate	185	[?]	Very good	Not addressed/reported			185	[?]	Doubtful	Not addressed/reported		
Teixeira et al. (2011)	CES	Full scale	Brazilian	116	None	[?]	Inadequate	76	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Zeeck et al. (2017)	CES	Full scale	German	571	CFA	[+]	Very good	571	[?]	Very good	Not addressed/reported			307	[-]	Doubtful	Not addressed/reported		
Taranis et al. (2011)	CET	Full scale	English	367	EFA	[+]	Adequate	101	[?]	Very good	Not addressed/reported			101; 97	[-]	Adequate	Not addressed/reported		
Taranis et al. (2011)	CET	Avoidance	English	Not applicable (subscale)				101	[?]	Very good	Not addressed/reported			101; 97	[+]	Adequate	Not addressed/reported		
Taranis et al. (2011)	CET	Weight control	English	Not applicable (subscale)				101	[?]	Very good	Not addressed/reported			101; 97	[-]	Adequate	Not addressed/reported		
Taranis et al. (2011)	CET	Mood improvement	English	Not applicable (subscale)				101	[?]	Very good	Not addressed/reported			101; 97	[-]	Adequate	Not addressed/reported		
Taranis et al. (2011)	CET	Lack of enjoyment	English	Not applicable (subscale)				101	[?]	Very good	Not addressed/reported			101; 97	[-]	Adequate	Not addressed/reported		
Taranis et al. (2011)	CET	Rigidity	English	Not applicable (subscale)				101	[?]	Very good	Not addressed/reported			101; 97	[-]	Adequate	Not addressed/reported		
Formby et al. (2014)	CET	Full scale	English	104	CFA	[+]	Very good	104	[?]	Very good	Not addressed/reported			104	[-]	Doubtful	Not addressed/reported		
Formby et al. (2014)	CET	Avoidance	English	Not applicable (subscale)				104	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Formby et al. (2014)	CET	Weight control	English	Not applicable (subscale)				104	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Formby et al. (2014)	CET	Mood improvement	English	Not applicable (subscale)				104	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Formby et al. (2014)	CET	Lack of enjoyment	English	Not applicable (subscale)				104	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Formby et al. (2014)	CET	Rigidity	English	Not applicable (subscale)				104	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Goodwin et al. (2011)	CET	Full scale	English	1012	EFA	[+]	Adequate	1012	[?]	Very good	Not addressed/reported			1012	[+]	Doubtful	Not addressed/reported		
Goodwin et al. (2011)	CET	Avoidance	English	Not applicable (subscale)				1012	[?]	Very good	Not addressed/reported			1012	[-]	Doubtful	Not addressed/reported		
Goodwin et al. (2011)	CET	Weight control	English	Not applicable (subscale)				1012	[?]	Very good	Not addressed/reported			1012	[-]	Doubtful	Not addressed/reported		
Goodwin et al. (2011)	CET	Mood improvement	English	Not applicable (subscale)				1012	[?]	Very good	Not addressed/reported			1012	[-]	Doubtful	Not addressed/reported		
Goodwin et al. (2011)	CET	Lack of enjoyment	English	Not applicable (subscale)				1012	[?]	Very good	Not addressed/reported			1012	[-]	Doubtful	Not addressed/reported		
Goodwin et al. (2011)	CET	Rigidity	English	Not applicable (subscale)				1012	[?]	Very good	Not addressed/reported			1012	[-]	Doubtful	Not addressed/reported		
Meyer et al. (2016)	CET	Full scale	English	354	CFA	[-]	Very good	354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Meyer et al. (2016)	CET	Avoidance	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

				Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
Study	Measure	Full scale/ Subscale	Language	*n*	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Meyer et al. (2016)	CET	Weight control	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Meyer et al. (2016)	CET	Mood improvement	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Meyer et al. (2016)	CET	Lack of enjoyment	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Meyer et al. (2016)	CET	Rigidity	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Sauchelli et al. (2016)	CET	Full scale	Spanish	285	CFA	[-]	Very good	285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	285	[?]	Doubtful
Sauchelli et al. (2016)	CET	Avoidance	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Weight control	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Mood improvement	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Lack of enjoyment	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Rigidity	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Young et al. (2016)	CET	Full scale	English	78	CFA	[-]	Very good	78	[?]	Very good	Not addressed/reported			78	[+]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Avoidance	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[+]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Weight control	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Mood improvement	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Lack of enjoyment	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Rigidity	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Vrabel et al. (2019)	CET	Full scale	Norwegian	166	CFA	[-]	Very good	166	[?]	Very good	Not addressed/reported			166	[+]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Avoidance	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[+]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Weight control	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[+]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Mood improvement	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[-]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Lack of enjoyment	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[-]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Rigidity	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[-]	Doubtful	Not addressed/reported		
	CET-A	Full scale	English	689	CFA	[-]	Very good	689	[?]		Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Lichtenstein et al. (2016)	EAI	Full scale	Danish	603	EFA	[-]	Adequate	603	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Sicilia et al. (2013)	EAI	Full scale	Spanish	584	CFA	[+]	Very good	584	[-]	Very good	42	[+]	Doubtful	Not addressed/reported			584	[?]	Doubtful
Sicilia et al. (2017)	EAI	Full scale	Portuguese	251	CFA	[+]	Very good	251	[-]	Very good	56	[+]	Doubtful	Not addressed/reported			251	[?]	Doubtful
Mónok et al. (2012)	EAI	Full scale	Hungarian	458	CFA	[+]	Very good	458	[+]	Very good	Not addressed/reported			466	[-]	Very good	Not addressed/reported		
Szabo et al. (2013)	EAI	Full scale	Spanish	242	EFA	[-]	Adequate	242	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Szabo et al. (2019)	EAI-R	Full scale	English	227	CFA	[+]	Very good	227	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lichtenstein et al. (2018)	EAI-Y	Full scale	English	471	CFA	[-]	Very good	471	[?]	Very good	Not addressed/reported			471	[?]	Inadequate	Not addressed/reported		
Ogden et al. (1997)	EDQ	Full scale	English	449	EFA	[?]	Adequate	449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Interference	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Positive reward	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Withdrawal	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Weight control	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Insight into problem	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Social reasons	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Health reasons	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Stereotyped behaviour	English	Not applicable (subscale)				449	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Full scale	Italian	259	EFA	[+]	Adequate	259	[+]	Very good	Not addressed/reported			259	[-]	Doubtful	Not addressed/reported		
Grandi et al. (2013)	EDQ	Interference	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Positive reward	Italian	Not applicable (subscale)				259	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Withdrawal	Italian	Not applicable (subscale)				259	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Weight control	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Insight into problem	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Social reasons	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
	EDQ	Health reasons	Italian	Not applicable (subscale)				259	[+]		Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Plateau et al. (2014)										Very good									
Plateau et al. (2014)	CET-A	Full scale	English	689	EFA	[+]	Adequate	689	[?]	Very good	Not addressed/reported			689	[?]	Doubtful	Not addressed/reported		
Plateau et al. (2014)	CET-A	Avoidance	English	Not applicable (subscale)				689	[?]	Very good	Not addressed/reported			689	[?]	Doubtful	Not addressed/reported		
Plateau et al. (2014)	CET-A	Weight control	English	Not applicable (subscale)				689	[?]	Very good	Not addressed/reported			689	[?]	Doubtful	Not addressed/reported		
Plateau et al. (2014)	CET-A	Mood improvement	English	Not applicable (subscale)				689	[?]	Very good	Not addressed/reported			689	[-]	Doubtful	Not addressed/reported		
Limburg et al. (2021)	CET-A	Full scale	English	313	CFA	[-]	Very good	313	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Limburg et al. (2021)	CET-A	Avoidance	English	Not applicable (subscale)				313	[?]	Very good	Not addressed/reported			313	[+]	Doubtful	Not addressed/reported		
Limburg et al. (2021)	CET-A	Weight control	English	Not applicable (subscale)				313	[?]	Very good	Not addressed/reported			313	[-]	Doubtful	Not addressed/reported		
Limburg et al. (2021)	CET-A	Mood improvement	English	Not applicable (subscale)				313	[?]	Very good	Not addressed/reported			313	[-]	Doubtful	Not addressed/reported		
Swenne (2016)	CET-4F	Full scale	Swedish	210	EFA	[+]	Adequate	210	[?]	Very good	Not addressed/reported			-	-	-	Not addressed/reported		
Swenne (2016)	CET-4F	Avoidance	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Swenne (2016)	CET-4F	Weight control	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Swenne (2016)	CET-4F	Mood improvement	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Swenne (2016)	CET-4F	Lack of enjoyment	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Terry et al. (2004)	EAI	Full scale	English	200	EFA	[?]	Adequate	200	[+]	Very good	Not addressed/reported			200	[+]	Adequate	Not addressed/reported		
Griffiths et al. (2005)	EAI	Full scale	English	200	EFA	[+]	Adequate	200	[+]	Very good	79	[+]	Doubtful	200	[+]	Adequate	Not addressed/reported		
Griffiths et al. (2015; Country)	EAI	Full scale	English; Hungarian; Danish; Spanish	Not applicable (secondary data with a focus on measurement invariance)				Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			6031	[-]	Doubtful
Griffiths et al. (2015; Gender)	EAI	Full scale	English; Hungarian; Danish; Spanish	Not applicable (secondary data with a focus on measurement invariance)				Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			6031	[-]	Doubtful
Li et al. (2016)	EAI	Full scale	Chinese	1601	CFA	[+]	Very good	1601	[+]	Very good	50	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Lichtenstein et al. (2014)	EAI	Full scale	Danish	590	EFA	[-]	Adequate	590	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lichtenstein et al. (2014; Fitness)	EAI	Full scale	Danish	176	EFA	[-]	Adequate	176	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Grandi et al. (2013)										Very good									
Grandi et al. (2013)	EDQ	Stereotyped behaviour	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Full scale	French	160	CFA	[+]	Very good	160	[+]	Very good	160	[+]	Doubtful	160	[+]	Adequate	Not addressed/reported		
Kern & Baudin (2011)	EDQ	Interference	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Positive reward	French	Not applicable (subscale)				160	[+]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Withdrawal	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Weight control	French	Not applicable (subscale)				160	[+]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Insight into problem	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Social reasons	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Health reasons	French	Not applicable (subscale)				160	[+]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Stereotyped behaviour	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Full scale	English	266	None	[?]	Inadequate	553	[?]	Doubtful	46	[+]	Doubtful	366	[+]	Adequate	Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Tolerance	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Withdrawal	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Intention effects	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Lack of control	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Time	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Reduction in other activities	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Continuance	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
	EDS-R	Full scale	English	855	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Meyer et al. (2016)	CET	Weight control	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Meyer et al. (2016)	CET	Mood improvement	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Meyer et al. (2016)	CET	Lack of enjoyment	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Meyer et al. (2016)	CET	Rigidity	English	Not applicable (subscale)				354	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Sauchelli et al. (2016)	CET	Full scale	Spanish	285	CFA	[-]	Very good	285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	285	[?]	Doubtful
Sauchelli et al. (2016)	CET	Avoidance	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Weight control	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Mood improvement	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Lack of enjoyment	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Sauchelli et al. (2016)	CET	Rigidity	Spanish	Not applicable (subscale)				285	[?]	Very good	Not addressed/reported			158	[-]	Doubtful	Not addressed/reported		
Young et al. (2016)	CET	Full scale	English	78	CFA	[-]	Very good	78	[?]	Very good	Not addressed/reported			78	[+]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Avoidance	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[+]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Weight control	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Mood improvement	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Lack of enjoyment	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Young et al. (2016)	CET	Rigidity	English	Not applicable (subscale)				78	[?]	Very good	Not addressed/reported			78	[-]	Adequate	Not addressed/reported		
Vrabel et al. (2019)	CET	Full scale	Norwegian	166	CFA	[-]	Very good	166	[?]	Very good	Not addressed/reported			166	[+]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Avoidance	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[+]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Weight control	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[+]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Mood improvement	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[-]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Lack of enjoyment	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[-]	Doubtful	Not addressed/reported		
Vrabel et al. (2019)	CET	Rigidity	Norwegian	Not applicable (subscale)				166	[?]	Very good	Not addressed/reported			166	[-]	Doubtful	Not addressed/reported		
	CET-A	Full scale	English	689	CFA	[-]	Very good	689	[?]		Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Plateau et al. (2014)										Very good									
Plateau et al. (2014)	CET-A	Full scale	English	689	EFA	[+]	Adequate	689	[?]	Very good	Not addressed/reported			689	[?]	Doubtful	Not addressed/reported		
Plateau et al. (2014)	CET-A	Avoidance	English	Not applicable (subscale)				689	[?]	Very good	Not addressed/reported			689	[?]	Doubtful	Not addressed/reported		
Plateau et al. (2014)	CET-A	Weight control	English	Not applicable (subscale)				689	[?]	Very good	Not addressed/reported			689	[?]	Doubtful	Not addressed/reported		
Plateau et al. (2014)	CET-A	Mood improvement	English	Not applicable (subscale)				689	[?]	Very good	Not addressed/reported			689	[-]	Doubtful	Not addressed/reported		
Limburg et al. (2021)	CET-A	Full scale	English	313	CFA	[-]	Very good	313	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Limburg et al. (2021)	CET-A	Avoidance	English	Not applicable (subscale)				313	[?]	Very good	Not addressed/reported			313	[+]	Doubtful	Not addressed/reported		
Limburg et al. (2021)	CET-A	Weight control	English	Not applicable (subscale)				313	[?]	Very good	Not addressed/reported			313	[-]	Doubtful	Not addressed/reported		
Limburg et al. (2021)	CET-A	Mood improvement	English	Not applicable (subscale)				313	[?]	Very good	Not addressed/reported			313	[-]	Doubtful	Not addressed/reported		
Swenne (2016)	CET-4F	Full scale	Swedish	210	EFA	[+]	Adequate	210	[?]	Very good	Not addressed/reported			-	-	-	Not addressed/reported		
Swenne (2016)	CET-4F	Avoidance	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Swenne (2016)	CET-4F	Weight control	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Swenne (2016)	CET-4F	Mood improvement	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Swenne (2016)	CET-4F	Lack of enjoyment	Swedish	Not applicable (subscale)				210	[?]	Very good	Not addressed/reported			210	[?]	Inadequate	Not addressed/reported		
Terry et al. (2004)	EAI	Full scale	English	200	EFA	[?]	Adequate	200	[+]	Very good	Not addressed/reported			200	[+]	Adequate	Not addressed/reported		
Griffiths et al. (2005)	EAI	Full scale	English	200	EFA	[+]	Adequate	200	[+]	Very good	79	[+]	Doubtful	200	[+]	Adequate	Not addressed/reported		
Griffiths et al. (2015; Country)	EAI	Full scale	English; Hungarian; Danish; Spanish	Not applicable (secondary data with a focus on measurement invariance)				Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			6031	[-]	Doubtful
Griffiths et al. (2015; Gender)	EAI	Full scale	English; Hungarian; Danish; Spanish	Not applicable (secondary data with a focus on measurement invariance)				Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			Not applicable (secondary data with a focus on measurement invariance)			6031	[-]	Doubtful
Li et al. (2016)	EAI	Full scale	Chinese	1601	CFA	[+]	Very good	1601	[+]	Very good	50	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Lichtenstein et al. (2014)	EAI	Full scale	Danish	590	EFA	[-]	Adequate	590	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lichtenstein et al. (2014; Fitness)	EAI	Full scale	Danish	176	EFA	[-]	Adequate	176	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Lichtenstein et al. (2016)	EAI	Full scale	Danish	603	EFA	[-]	Adequate	603	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Sicilia et al. (2013)	EAI	Full scale	Spanish	584	CFA	[+]	Very good	584	[-]	Very good	42	[+]	Doubtful	Not addressed/reported			584	[?]	Doubtful
Sicilia et al. (2017)	EAI	Full scale	Portuguese	251	CFA	[+]	Very good	251	[-]	Very good	56	[+]	Doubtful	Not addressed/reported			251	[?]	Doubtful
Mónok et al. (2012)	EAI	Full scale	Hungarian	458	CFA	[+]	Very good	458	[+]	Very good	Not addressed/reported			466	[-]	Very good	Not addressed/reported		
Szabo et al. (2013)	EAI	Full scale	Spanish	242	EFA	[-]	Adequate	242	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Szabo et al. (2019)	EAI-R	Full scale	English	227	CFA	[+]	Very good	227	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lichtenstein et al. (2018)	EAI-Y	Full scale	English	471	CFA	[-]	Very good	471	[?]	Very good	Not addressed/reported			471	[?]	Inadequate	Not addressed/reported		
Ogden et al. (1997)	EDQ	Full scale	English	449	EFA	[?]	Adequate	449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Interference	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Positive reward	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Withdrawal	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Weight control	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Insight into problem	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Social reasons	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Health reasons	English	Not applicable (subscale)				449	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Ogden et al. (1997)	EDQ	Stereotyped behaviour	English	Not applicable (subscale)				449	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Full scale	Italian	259	EFA	[+]	Adequate	259	[+]	Very good	Not addressed/reported			259	[-]	Doubtful	Not addressed/reported		
Grandi et al. (2013)	EDQ	Interference	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Positive reward	Italian	Not applicable (subscale)				259	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Withdrawal	Italian	Not applicable (subscale)				259	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Weight control	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Insight into problem	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Grandi et al. (2013)	EDQ	Social reasons	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
	EDQ	Health reasons	Italian	Not applicable (subscale)				259	[+]		Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Grandi et al. (2013)										Very good									
Grandi et al. (2013)	EDQ	Stereotyped behaviour	Italian	Not applicable (subscale)				259	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Full scale	French	160	CFA	[+]	Very good	160	[+]	Very good	160	[+]	Doubtful	160	[+]	Adequate	Not addressed/reported		
Kern & Baudin (2011)	EDQ	Interference	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Positive reward	French	Not applicable (subscale)				160	[+]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Withdrawal	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Weight control	French	Not applicable (subscale)				160	[+]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Insight into problem	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Social reasons	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Health reasons	French	Not applicable (subscale)				160	[+]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern & Baudin (2011)	EDQ	Stereotyped behaviour	French	Not applicable (subscale)				160	[-]	Very good	160	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Full scale	English	266	None	[?]	Inadequate	553	[?]	Doubtful	46	[+]	Doubtful	366	[+]	Adequate	Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Tolerance	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Withdrawal	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Intention effects	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Lack of control	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Time	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Reduction in other activities	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Hausenblas & Downs (2002)	EDS	Continuance	English	Not applicable (subscale)				Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
	EDS-R	Full scale	English	855	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Downs et al. (2004)																			
Downs et al. (2004)	EDS-R	Tolerance	English	Not applicable (subscale)				408; 855	[+] [+]	Very good	30	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Downs et al. (2004)	EDS-R	Withdrawal	English	Not applicable (subscale)				408; 855	[+] [+]	Very good	30	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Downs et al. (2004)	EDS-R	Intention effects	English	Not applicable (subscale)				408; 855	[+] [+]	Very good	30	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Downs et al. (2004)	EDS-R	Lack of control	English	Not applicable (subscale)				408; 855	[+] [+]	Very good	30	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Downs et al. (2004)	EDS-R	Time	English	Not applicable (subscale)				408; 855	[+] [+]	Very good	30	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Downs et al. (2004)	EDS-R	Reduction in other activities	English	Not applicable (subscale)				408; 855	[-] [+]	Very good	30	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Downs et al. (2004)	EDS-R	Continuance	English	Not applicable (subscale)				408; 855	[+] [+]	Very good	30	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Full scale	Brazilian	376	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Tolerance	Brazilian	Not applicable (subscale)				376; 333	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Withdrawal	Brazilian	Not applicable (subscale)				376; 333	[-] [-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Intention effects	Brazilian	Not applicable (subscale)				376; 333	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Lack of control	Brazilian	Not applicable (subscale)				376; 333	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Time	Brazilian	Not applicable (subscale)				376; 333	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Reduction in other activities	Brazilian	Not applicable (subscale)				376; 333	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Alchieri et al. (2015)	EDS-R	Continuance	Brazilian	Not applicable (subscale)				376; 333	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Allegre & Therme (2008)	EDS-R	Full scale	French	516	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Allegre & Therme (2008)	EDS-R	Tolerance	French	Not applicable (subscale)				516	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Allegre & Therme (2008)	EDS-R	Withdrawal	French	Not applicable (subscale)				516	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Allegre & Therme (2008)	EDS-R	Intention effects	French	Not applicable (subscale)				516	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
	EDS-R	Lack of control	French	Not applicable (subscale)				516	[+]		Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Allegre & Therme (2008)										Very good									
Allegre & Therme (2008)	EDS-R	Time	French	Not applicable (subscale)				516	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Allegre & Therme (2008)	EDS-R	Reduction in other activities	French	Not applicable (subscale)				516	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Allegre & Therme (2008)	EDS-R	Continuance	French	Not applicable (subscale)				516	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Full scale	Italian	523	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Tolerance	Italian	Not applicable (subscale)				523	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Withdrawal	Italian	Not applicable (subscale)				523	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Intention effects	Italian	Not applicable (subscale)				523	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Lack of control	Italian	Not applicable (subscale)				523	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Time	Italian	Not applicable (subscale)				523	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Reduction in other activities	Italian	Not applicable (subscale)				523	[-]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Costa et al. (2012)	EDS-R	Continuance	Italian	Not applicable (subscale)				523	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Full scale	French	811	CFA	[+]	Very good				Not addressed/reported			Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Tolerance	French	Not applicable (subscale)				73; 79	[+] [+]	Very good	79	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Withdrawal	French	Not applicable (subscale)				73; 79	[+] [+]	Very good	79	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Intention effects	French	Not applicable (subscale)				73; 79	[+] [+]	Very good	79	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Lack of control	French	Not applicable (subscale)				73; 79	[+] [+]	Very good	79	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Time	French	Not applicable (subscale)				73; 79	[+] [+]	Very good	79	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Reduction in other activities	French	Not applicable (subscale)				73; 79	[-] [-]	Very good	79	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Kern (2007)	EDS-R	Continuance	French	Not applicable (subscale)				73; 79	[+] [+]	Very good	79	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
	EDS-R	Full scale	Swedish Portuguese	162; 269	CFA	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			431	[+]	Doubtful

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Lindwall & Palmeira (2009)																			
Lindwall & Palmeira (2009)	EDS-R	Tolerance	Swedish Portuguese	Not applicable (subscale)				162; 269	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lindwall & Palmeira (2009)	EDS-R	Withdrawal	Swedish Portuguese	Not applicable (subscale)				162; 269	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lindwall & Palmeira (2009)	EDS-R	Intention effects	Swedish Portuguese	Not applicable (subscale)				162; 269	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lindwall & Palmeira (2009)	EDS-R	Lack of control	Swedish Portuguese	Not applicable (subscale)				162; 269	[-] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lindwall & Palmeira (2009)	EDS-R	Time	Swedish Portuguese	Not applicable (subscale)				162; 269	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lindwall & Palmeira (2009)	EDS-R	Reduction in other activities	Swedish Portuguese	Not applicable (subscale)				162; 269	[-] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Lindwall & Palmeira (2009)	EDS-R	Continuance	Swedish Portuguese	Not applicable (subscale)				162; 269	[+] [+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Mónok et al. (2012)	EDS-R	Full scale	Hungarian	465	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Mónok et al. (2012)	EDS-R	Tolerance	Hungarian	Not applicable (subscale)				465	[+]	Very good	Not addressed/reported			466	[+]	Very good	Not addressed/reported		
Mónok et al. (2012)	EDS-R	Withdrawal	Hungarian	Not applicable (subscale)				465	[+]	Very good	Not addressed/reported			466	[-]	Very good	Not addressed/reported		
Mónok et al. (2012)	EDS-R	Intention effects	Hungarian	Not applicable (subscale)				465	[+]	Very good	Not addressed/reported			466	[+]	Very good	Not addressed/reported		
Mónok et al. (2012)	EDS-R	Lack of control	Hungarian	Not applicable (subscale)				465	[-]	Very good	Not addressed/reported			466	[+]	Very good	Not addressed/reported		
Mónok et al. (2012)	EDS-R	Time	Hungarian	Not applicable (subscale)				465	[+]	Very good	Not addressed/reported			466	[+]	Very good	Not addressed/reported		
Mónok et al. (2012)	EDS-R	Reduction in other activities	Hungarian	Not applicable (subscale)				465	[-]	Very good	Not addressed/reported			466	[+]	Very good	Not addressed/reported		
Mónok et al. (2012)	EDS-R	Continuance	Hungarian	Not applicable (subscale)				465	[-]	Very good	Not addressed/reported			466	[-]	Very good	Not addressed/reported		
Müller et al. (2013)	EDS-R	Full scale	German	1611	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			1611	[-]	Doubtful	Not addressed/reported		
Müller et al. (2013)	EDS-R	Tolerance	German	Not applicable (subscale)				1611	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Müller et al. (2013)	EDS-R	Withdrawal	German	Not applicable (subscale)				1611	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Müller et al. (2013)	EDS-R	Intention effects	German	Not applicable (subscale)				1611	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Müller et al. (2013)	EDS-R	Lack of control	German	Not applicable (subscale)				1611	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Müller et al. (2013)	EDS-R	Time	German	Not applicable (subscale)				1611	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Müller et al. (2013)	EDS-R	Reduction in other activities	German	Not applicable (subscale)				1611	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Müller et al. (2013)	EDS-R	Continuance	German	Not applicable (subscale)				1611	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Full scale	Greek	581	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Tolerance	Greek	Not applicable (subscale)				581	[+]	Very good	[+]	Doubtful		581	[-]	Adequate	Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Withdrawal	Greek	Not applicable (subscale)				581	[+]	Very good	[+]	Doubtful		581	[-]	Adequate	Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Intention effects	Greek	Not applicable (subscale)				581	[+]	Very good	[+]	Doubtful		581	[-]	Adequate	Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Lack of control	Greek	Not applicable (subscale)				581	[+]	Very good	[+]	Doubtful		581	[-]	Adequate	Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Time	Greek	Not applicable (subscale)				581	[+]	Very good	[+]	Doubtful		581	[-]	Adequate	Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Reduction in other activities	Greek	Not applicable (subscale)				581	[-]	Very good	[+]	Doubtful		581	[-]	Adequate	Not addressed/reported		
Parastatidou et al. (2011)	EDS-R	Continuance	Greek	Not applicable (subscale)				581	[+]	Very good	[+]	Doubtful		581	[-]	Adequate	Not addressed/reported		
Shin & You (2015)	EDS-R	Full scale	Korean	402	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			402	[+]	Very good
Shin & You (2015	EDS-R	Tolerance	Korean	Not applicable (subscale)				402	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Shin & You (2015)	EDS-R	Withdrawal	Korean	Not applicable (subscale)				402	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Shin & You (2015)	EDS-R	Intention effects	Korean	Not applicable (subscale)				402	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Shin & You (2015)	EDS-R	Lack of control	Korean	Not applicable (subscale)				402	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Shin & You (2015)	EDS-R	Time	Korean	Not applicable (subscale)				402	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Shin & You (2015)	EDS-R	Reduction in other activities	Korean	Not applicable (subscale)				402	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Shin & You (2015)	EDS-R	Continuance	Korean	Not applicable (subscale)				402	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
	EDS-R	Full scale	Spanish	531	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			531	[?]	Doubtful

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Sicilia & González-Cutre (2011)																			
Sicilia & González-Cutre (2011)	EDS-R	Tolerance	Spanish	Not applicable (subscale)				531	[+]	Very good	81	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Sicilia & González-Cutre (2011)	EDS-R	Withdrawal	Spanish	Not applicable (subscale)				531	[+]	Very good	81	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Sicilia & González-Cutre (2011)	EDS-R	Intention effects	Spanish	Not applicable (subscale)				531	[+]	Very good	81	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Sicilia & González-Cutre (2011)	EDS-R	Lack of control	Spanish	Not applicable (subscale)				531	[+]	Very good	81	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Sicilia & González-Cutre (2011)	EDS-R	Time	Spanish	Not applicable (subscale)				531	[+]	Very good	81	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Sicilia & González-Cutre (2011)	EDS-R	Reduction in other activities	Spanish	Not applicable (subscale)				531	[-]	Very good	81	[-]	Doubtful	Not addressed/reported			Not addressed/reported		
Sicilia & González-Cutre (2011)	EDS-R	Continuance	Spanish	Not applicable (subscale)				531	[+]	Very good	81	[+]	Doubtful	Not addressed/reported			Not addressed/reported		
Pasman & Thompson (1988)	OEQ	Full scale	English	137	None	[?]	Inadequate	137	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Brehm & Steffen (2013)	OEQ	Full scale	English	499	EFA	[+]	Adequate	499	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Steffen & Brehm (1999)	OEQ-10	Full scale	English	255	EFA	[+]	Adequate	255	[?]	Very good	Not addressed/reported			255	[-]	Doubtful	Not addressed/reported		
Steffen & Brehm (1999)	OEQ-10	Emotional element of exercise	English	Not applicable (subscale)				255	[?]	Very good	Not addressed/reported			255	[-]	Doubtful	Not addressed/reported		
Steffen & Brehm (1999)	OEQ-10	Exercise frequency and intensity	English	Not applicable (subscale)				255	[?]	Very good	Not addressed/reported			255	[-]	Doubtful	Not addressed/reported		
Steffen & Brehm (1999)	OEQ-10	Exercise preoccupation	English	Not applicable (subscale)				255	[?]	Very good	Not addressed/reported			255	[+]	Doubtful	Not addressed/reported		
Parastatidou et al. (2011)	OEQ-10	Full scale	English	581	CFA	[-]	Very good	581	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Parastatidou et al. (2011)	OEQ-10	Emotional element of exercise	English	Not applicable (subscale)				581	[?]	Very good	Not addressed/reported			581	[-]	Adequate	Not addressed/reported		

Table 4. Continued.

Study	Measure	Full scale/ Subscale	Language	Structural validity				Internal consistency			Reliability			Hypothesis testing			Cross-cultural validity/ measurement invariance		
				n	Type of analysis	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias	*n*	Rating of result	Rating of risk bias
Parastatidou et al. (2011)	OEQ-10	Exercise frequency and intensity	English	Not applicable (subscale)				581	[?]	Very good	Not addressed/reported			581	[-]	Adequate	Not addressed/reported		
Parastatidou et al. (2011)	OEQ-10	Exercise preoccupation	English	Not applicable (subscale)				581	[?]	Very good	Not addressed/reported			581	[-]	Adequate	Not addressed/reported		
Ackard et al. (2002)	OEQ-11	Full scale	English	586	EFA	[+]	Adequate	586	[?]	Very good	Not addressed/reported			586	[+]	Doubtful	Not addressed/reported		
Ackard et al. (2002)	OEQ-11	Exercise fixation	English	Not applicable (subscale)				586	[?]	Very good	Not addressed/reported			586	[-]	Doubtful	Not addressed/reported		
Ackard et al. (2002)	OEQ-11	Exercise frequency	English	Not applicable (subscale)				586	[?]	Very good	Not addressed/reported			586	[-]	Doubtful	Not addressed/reported		
Ackard et al. (2002)	OEQ-11	Exercise commitment	English	Not applicable (subscale)				586	[?]	Very good	Not addressed/reported			586	[-]	Doubtful	Not addressed/reported		
Duncan et al. (2012)	OEQ-R	Full scale	English	637	CFA	[+]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported			Not addressed/reported		
Duncan et al. (2012)	OEQ-R	Preoccupation with exercise	English	Not applicable (subscale)				241	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Duncan et al. (2012)	OEQ-R	Exercise behaviour	English	Not applicable (subscale)				241	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		
Duncan et al. (2012)	OEQ-R	Exercise emotionality	English	Not applicable (subscale)				241	[?]	Very good	Not addressed/reported			Not addressed/reported			Not addressed/reported		

Note. [+] = sufficient; [-] = insufficient; [?] = indeterminate; CES = Commitment to Exercise Scale; CET = Compulsive Exercise Test; CET-A = Compulsive Exercise Test-Athletes; CET-4F = Compulsive Exercise Test (4-factors); EAI = Exercise Addiction Inventory; EAI-R = Exercise Addiction Inventory-Revised; EAI-Y = Exercise Addiction Inventory-Youth; EDQ = Exercise Dependence Questionnaire; EDS = Exercise Dependence Scale; EDS-R = Exercise Dependence Scale-Revised; OEQ = Obligatory Exercise Questionnaire; OEQ-10 = Obligatory Exercise Questionnaire-10-items; OEQ-11 = Obligatory Exercise Questionnaire-11-items; OEQ-R = Obligatory Exercise Questionnaire-Revised; EFA = Exploratory factor analysis; CFA = Confirmatory factor analysis.