Abstract

This systematic review synthesizes and critically appraises measurement properties of influential body image measures. Eight measures that met the definition of an assessment of body image (i.e., an individual’s cognitive or affective evaluation of their body or appearance with a positive or negative valence), and scored high on systematic expert priority ranking, were included. These measures were: the Body Appreciation Scale (original BAS and BAS-2), the Body Esteem Scale for Adolescents and Adults, the Body Shape Questionnaire, the Centre for Appearance Research Valence Scale, the Drive for Muscularity Scale, two subscales of the Eating Disorders Examination Questionnaire, one subscale of the Eating Disorder Inventory 3, and two subscales of the Multidimensional Body Relations Questionnaire. Articles assessing these scales’ psychometric properties (*N* = 136) were evaluated for their methodological quality using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist, and a best evidence synthesis was performed. The results supported the majority of measures in terms of reliability and validity; however, suitability varied across populations, and some measurement properties were insufficiently evaluated. The measures are discussed in detail, including recommendations for their future use in research and clinical practice.

1. Introduction

Body image research has significantly expanded over the past decades and as a result, a plethora of instruments have now been designed to assess body image (Thompson, Burke, & Krawczyk, 2012). This great diversity makes it challenging for researchers and clinicians to determine which instruments to use, with calls to establish a consensus on measurement choices in order to advance research in the field (Krawczyk, Menzel, & Thompson, 2012). Moreover, there is a lack of systematic investigations into the reliability and validity of body image instruments, despite the imperative of producing empirically sound work. To improve the cohesiveness of research in this field and to increase the comparability of findings, there is a clear need for the systematization of existing measures and for recommendations for use based on their psychometric properties. When conducting the searches for this review (see Method section), we found that more than 150 different body images measures had been used in recent years. The present review offers a compilation and evaluation of the most theoretically important and/or commonly used of these measures, based on our definition of body image provided below, with the aim of bringing body image researchers together and facilitating comparisons across future studies. Hence, the aim of the present systematic review was to rigorously synthesize and evaluate body image measures to move the body image research field forward.

1.1. Body Image Definition

One potential explanation for the great diversity of body image instruments is that body image is multidimensional, and numerous measures exist to assess various components of this construct (Thompson et al., 2012). The present review builds on Thomas F. Cash's definition of body image as a multidimensional construct encompassing self-perceptions and attitudes regarding one’s physical appearance (e.g., Cash, Fleming, Alindogan, Steadman, & Whitehead, 2002). Consistent with Cash’s definition, attitudinal body image consists of at least two dimensions: (1) evaluation/affect, which includes body-image appraisals and satisfaction/dissatisfaction, and (2) investment, such as the salience, centrality, or extent of cognitive-behavioral emphasis on one's appearance (Cash, 1994). In general, among researchers and clinicians as well as the public, the evaluative component is the aspect most commonly considered to represent body image. For instance, Cash (2011) stated that: "Researchers who want to measure 'body image' must give careful thought to what they really mean by this term. Most often, they mean something like 'how people feel about their body'. So perhaps they want a measure of body image satisfaction-dissatisfaction." (pp. 129-130). In line with Cash's notion, Krawczyk et al. (2012) concluded that the most commonly used body image measures are those assessing a person's evaluation of their physical appearance. While acknowledging that body image theoretically consists of evaluative, perceptual, and behavioral components (Cash et al., 2002), the present review adopts a definition of body image that focuses on the evaluative component. This is based on the observation that body (dis)satisfaction is very often referred to as body image more broadly, and the large number of instruments purporting to measure this construct. Hence, throughout this review, body image is defined as an individual making some kind of cognitive or affective evaluation of their body or appearance with a positive or negative valence.

Although body image research often takes a pathologizing lens, focusing on body dissatisfaction, increasing recognition of the importance of considering body appreciation and positive components of body image has emerged in recent years (Tylka & Wood-Barcalow, 2015a). Including research on adaptive or healthy body image is essential to the future of the field (Smolak & Cash, 2011). Thus, the present systematic review includes both negative and positive aspects of the evaluative component of body image, also acknowledging a conceptual distinction between body satisfaction and positive body image (see Tylka & Wood-Barcalow, 2015a).

1.2. Previous Reviews of Body Image Measures

Previous reviews of body image measures are available (e.g., Cafri & Thompson, 2004; Gardner & Brown, 2010; Kashubeck-West, Mintz, & Saunders, 2001; Menzel, Krawczyk, & Thompson, 2011; Skrzypek, Wehmeier, & Remschmidt, 2001; Thompson et al., 2012; Thompson, Penner, & Altabe, 1990; Túry, Güleç, & Kohls, 2010; Webb, Wood-Barcalow, & Tylka, 2015). However, the majority of these reviews have focused on measures of specific body image-related constructs or measures suitable for specific populations. For instance, Gardner and Brown (2010) systematically reviewed figural drawing scales designed to assess body image disturbance, and Webb et al. (2015) reviewed measures of positive body image. Cafri and Thompson (2004) reviewed methods for measuring male body image, while Skrzypek and colleagues (2001) reviewed body image assessment methods among patients with anorexia nervosa. The reviews by Kashubeck-West et al. (2001) and Túry et al. (2010) both focused on eating disorder measures; however, these reviews also included some measures of body image. In addition to these reviews, Thompson and colleagues (i.e., Menzel et al., 2011; Thompson et al., 1990, 2012) have authored several book chapters on currently used body image measures with reported psychometric properties. These articles and chapters make important contributions to efforts aiming to summarize and organize the available body image assessments. Nevertheless, existing reviews have been mainly limited to narrative reviews and to date, a rigorous and comprehensive review of available assessment instruments of body image with standardized quality assessment that can serve to unify and guide the field forward is lacking.

1.3. Contributions of the Current Systematic Review

The present review adds to the research field of body image in several ways. First, it addresses a recent call for more systematic reviews within the field of body image as an area in need of attention (Tylka, 2018), particularly in the area of body image measurement (e.g., Thompson et al., 2012). Second, the review adds to the literature by identifying which body image measures are currently being used in research, but also by identifying which measures are psychometrically sound, and in which populations. Although psychometrically sound measurement is not a guarantee for the accuracy of research findings, poor measures will certainly undermine research conclusions (Cash, 2011), as well as the quality of the research conducted in the field of body image.

Producing empirically sound research in the body image research field is dependent on close attention to issues related to validity and reliability of the measurement of this construct (Thompson et al., 2012). In addition, however, it is important to consider the quality of the studies reporting on the psychometric properties of the measures (Terwee et al., 2012). No previous systematic review of body image measures has assessed the measurement properties of relevant scales *as well as* evaluating the quality of the studies reporting on these psychometric properties. In fact, no previous review has used such a comprehensive methodology as the one employed in the present study, namely the Consensus-Based Standards for the Selection of Health Status Measurement Instruments (COSMIN) method. The COSMIN is increasingly accepted as the gold standard for evidence synthesis of the performance of patient-reported outcome measures (Mokkink et al., 2010a, 2010b, 2010c, 2010d; Terwee et al., 2012).

The current review is also highly relevant to clinical practice since many body image measures are used in clinical settings (Cash, 2011; Rumsey & Harcourt, 2012), for instance with patients affected by eating disorders (e.g., Kashubeck-West et al., 2001; Skrzypek et al., 2001; Túry et al., 2010), cancer (e.g., Lewis-Smith, Diedrichs, Rumsey & Harcourt, 2018), and conditions that affect appearance (e.g., cleft lip and/or palate; Stock, Billaud Feragen, Rumsey, 2018). Also, there has been an increase in published studies focusing on the use of body image measures as patient-reported outcomes (PROs; Cash, 2011; e.g., in burns care; Griffiths et al., 2017).

1.4. Aim

The aim of this systematic review was to rigorously synthesize and appraise the methodological quality of evidence on the measurement properties of influential self-report body image measures and to provide recommendations about instruments most useful and psychometrically sound for research and clinical practice.

2. Method

2.1. Research Team

The review was conducted by an international research team, with expertise across a wide range of body image areas (e.g., disfigurement, appearance dissatisfaction and concerns, eating behaviors, obesity, chronic pain, body image interventions, social and cultural influences on body image, weight and appearance stigmatization), as well as with previous experience in conducting systematic reviews evaluating outcome measures. Specifically the team consisted of: two professors from the Centre for Appearance Research at the University of the West of England, UK; one professor and one PhD researcher from University College Dublin, Ireland; a professor and doctoral student from University of Gothenburg, Sweden; two professors from University of Aveiro, Portugal; a PhD researcher from Radboud University Nijmegen, Netherlands/McGill University, and Lady Davis Institute for Medical Research, Canada; and a PhD researcher from Northeastern University, USA/ Centre Hospitalier Universitaire de Montpellier, France. In addition to the research team, a number of research assistants, as well as three university librarians, assisted with elements of the project (see Acknowledgements). The project was initiated as a part of the European

Cooperation in Science and Technology (COST) Action IS1210 Appearance Matters, to which all research team members belong.

2.2. Search Strategy and Selection of Measures

2.2.1. Search Step 1 – Identifying measures. The aim of the first step of the search process was to identify body image measures that had been used in the three years prior to the commencement of the review [August 2011 to August 2014]. To ensure accuracy in the literature searching, the search strategy was developed by a Senior Research Librarian at the University of the West of England, UK, with expertise in systematic reviews and previous experience with body image research. The widely used databases MEDLINE, PsycINFO, and CINAHL Plus were searched, using EBSCO as the database platform, and the search was limited to the abstract field. To ensure comprehensiveness, an initial version of the MEDLINE search strategy was tested against already-identified publications from a preliminary list created by the research team. The search strategy was then adapted for PsycINFO and CINAHL Plus. See online Supplementary material for the included search terms. The initial search resulted in 2,439 hits. After limiting the “body image” term to the title field only and de-duplication, the initial search identified 404 studies. The names of the measures were extracted from full-text articles, along with basic characteristics such as authors and journals, and entered into an Excel file for review. The 404 studies had used, in aggregate, 151 different body image measures. In teams of two independent raters, all 151 body image measures were reviewed in order to determine if they met the predefined definition of body image. The independent ratings were compared and if needed, a third researcher was consulted. After this process, 58 body image measures remained (see online Supplementary material).

2.2.2. Definition of body image and measure criteria. The research team agreed to include measures that assessed an evaluative component of body image, according to the following definition: "*Contains a measure that generates a total score or subscale scores that assess an evaluative component of body image, defined as an individual making some kind of cognitive or affective evaluation of their body or appearance with a positive or negative valence. Scales or subscales that include one or more items reflective of this construct, but where the total measure or subscale score is clearly not reflective of the construct, will not be included.*" Specifically, in accordance with this definition, measures were included if a clear majority of the items were considered to reflect cognitive or affective dimensions. Body image silhouette measures were excluded as it was concluded that they did not meet the definition, as a rating of one's ideal body, or ideal-actual discrepancy, was judged to be conceptually different from an evaluation of one's own body (e.g., an ideal-actual discrepancy does not automatically indicate a dissatisfaction with the current body size, nor does it speak to the extent of any dissatisfaction). Moreover, despite the fact that body image experiences vary over time and situational contexts, lending themselves to state and trait appraisals (Cash et al., 2002; Tiggemann, 2001), only trait measures were included in this review. In addition, measures specifically developed to assess body image in children were excluded to further refine the focus of the review.

2.2.3. Priority ranking. The research team ranked the remaining 58 measures to prioritize those to include in the review. Specifically, the research team rated the relative priorities of the measures using an adapted version of the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) guidelines for deciding on important outcomes (Guyatt et al., 2011). In accordance with the GRADE guidelines, each team member independently rated each identified body image measure numerically on a scale ranging from 1-9 (1-3, of limited importance; 4-6, important; 7-9, critical) based on the perceived extent to which the measure was used in research, program evaluation and clinical work, the theoretical importance of the measure, and the likely impact on the field of reviewing its psychometric properties.

In a face-to-face meeting of the research team, the results of the ratings were discussed, and eight measures were selected for this systematic review. These measures received markedly higher priority ratings compared to all other measures and were deemed by the research team to be key measures of body image in the field. The eight measures included in the systematic review were: the original and revised Body Appreciation Scale (BAS/BAS-2; Avalos, Tylka, & Wood-Barcalow, 2005; Tylka & Wood-Barcalow, 2015b); the Body Esteem Scale for Adolescents and Adults (BESAA; Mendelson, Mendelson & White, 2001); the Body Shape Questionnaire (BSQ; Cooper, Taylor, Cooper, & Fairburn, 1987*);* the Centre for Appearance Research Valence Scale (CARVAL; Moss & Rosser, 2012); the Drive for Muscularity Scale (DMS; McCreary & Sasse, 2000); the Weight and Shape Concerns subscales of the Eating Disorders Examination Questionnaire (EDE-Q; Fairburn & Beglin, 1994); the Body Dissatisfaction subscale of the Eating Disorder Inventory-3 (EDI-3; Garner, 2004), and the Appearance Evaluation subscale and Body Areas Satisfaction Scale of the Multidimensional Body Relations Questionnaire (MBSRQ; Brown, Cash, & Mikulka, 1990). The measures included in the review are described in Table 1. However, although the priority ranking was systematized following established guidelines (Guyatt et al., 2011), it is important to emphazise that the inclusion of the measures is based on a consensus reached by the 10 authors about the importance of each measure (see Discussion). In addition to the priority ranking, advanced search on Google scholar was performed for all 58 measures that met the definition of body image in order to provide an estimate of the prevalence of each of the included eight measures (see Supplemental material). All included measures had high numbers of citations, except for the CARVAL (Moss & Rosser, 2012). This measure was still included based on high ratings in the criteria concerning likely future impact on the body image research field, especially within the field of visible differences, as the CARVAL is one of very few evaluative body image measures designed with people with visible differences in mind.

2.3. Search Strategy and Selection of Studies

2.3.1. Search Step 2 – Identifying studies. The aim of the second step of the search process was to identify literature focusing on the nine priority body image measures. Again, the search strategy was developed by the Senior Research Librarian at the University of the West of England, UK. Searches were conducted in March and April 2016 by the use of the following databases: CINAHL Plus, EMBASE, ERIC, PsycINFO, Scopus, and Web of Science. As the aim of Step 2 search was to identify studies with primary data on the measurement properties of the nine measures, each search was limited to articles that included the name of the measure or its commonly used abbreviation(s) in the title or abstract. In July 2018, an updated search was conducted in order to identify the most recent literature on the nine measures. This search was conducted by two university librarians at the University of Gothenburg, Sweden, and replicated the 2016 search process.

The flowchart presented in Figure 1 summarizes the search process and number of articles obtained and excluded in each step of the process. Further details regarding the searches (including search terms for each measure) are available as online Supplementary material. The citation management database RefWorks (RefWorks-COS, Bethesda, MD, USA) and the systematic review program Rayyan QCRI (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016) were used in the review process.

2.3.2. Study inclusion and exclusion criteria. For each measure, studies were included in the review if they reported measurement properties (e.g., reliability, validity, factor structure) for any of the nine measures in any population. Studies were excluded if they did not report original data (e.g., excluded if they were a letter, editorial, systematic review or meta-analysis). Studies reported only in conference paper, poster, or symposium abstracts were excluded, but authors of the abstracts were contacted to seek full study reports, published or unpublished, which provided sufficient information to extract results of analyses on measurement properties. Studies written in a language other than English were excluded. Studies that used the measures but did not focus on the instrument's development or the evaluation of one or more of its measurement properties were excluded. Examples of this process include studies that (1) used the measure in the validation process of another instrument, (2) as a correlate or outcome measure, without specifically studying measurement properties, (3) mentioned one aspect of measurement, such as Cronbach's alpha, but did not focus on measurement, and (4) no section of the article specifically dedicated to measurement. Moreover, we did not include specific children's versions of the measures, but we did include samples of all ages (including children and adolescents) where the study had used the original measure. If it was not possible to retrieve the full-text (after all team members had searched databases and the authors had been e-mailed) the article was excluded.

Regarding the EDI, we decided to only include articles evaluating its most recent version, the EDI-3, given that the third version is the most used in recent years, and addresses some important psychometric issues of the EDI-2 (Cumella, 2006). Regarding the BAS, both the BAS and the BAS-2 were included, since BAS-2 was only recently developed and both versions have been used in parallel in recent years.

2.3.3. Evaluation of eligibility. In sub-teams of two researchers, all articles were independently reviewed for eligibility using a standardized Excel sheet. The process for evaluating eligibility started with the following number of studies for each measure after duplicates had been removed: BAS, N = 195; BESAA, N = 419; BSQ, N = 756; CARVAL, N = 4; DMS, N = 311; EDE-Q, N = 1072; EDI-3, N = 2912; MBSRQ, N = 357. Any study deemed potentially eligible by either reviewer at the title/abstract level proceeded to full-text review. Disagreements after full-text review were, when necessary, resolved by consensus in consultation with a third researcher. Details of the studies obtained and excluded in each step of the process are described in the flowchart (Figure 1). The most common reason for exclusion, both after title/abstract review and full-text review, was that the study had used another measure with a similar name, another version of the measure, or did not include the subscales of interest (see Figure 1). For instance, specifically concerning the BESAA (Mendelson et al., 2001) and the EDI-3 (Garner, 2004), a large number of studies were excluded since they referred to other body esteem scales or previous versions of the EDI respectively. After full-text review, the following number of studies for each measure was included in the review: BAS, *N* = 23; BESAA, *N* = 6; BSQ, *N* = 23; CARVAL, *N* = 2; DMS, *N* = 16; EDE-Q, *N* = 44; EDI-3, *N* = 11; MBSRQ, *N* = 15. The total number of included studies was 136, as three studies (Franko et al., 2012; Kashubeck-West et al., 2013; Reilly, Anderson, Schaumberg, & Anderson, 2014) used more than one of the measures.

2.4. Data Synthesis and Quality Assessment

2.4.1. Quality of the articles. Evaluation of the methodological quality of the included articles was carried out using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist (Mokkink et al., 2010a, 2010b, 2010c, 2010d; Terwee et al., 2012). The COSMIN method and checklist involves assessing the methodological quality for each study across nine domains ("COSMIN boxes"). For each of the nine domains of measurement properties, the COSMIN checklist includes 5 to 18 items assessing methodological quality: internal consistency (11 items); reliability (including test-retest; 14 items); measurement error (11 items); content validity (5 items); structural validity (7 items); hypothesis testing (including convergent and divergent validity; 10 items); cross-cultural validity (15 items); criterion validity (7 items); and responsiveness (18 items; Mokkink et al., 2010b). In our review, no studies evaluating measurement error were found, and therefore this domain was excluded. In certain implementations of the COSMIN checklist, a tenth domain directed at studies using Item Response Theory (IRT) methods is used; however, in the present review none of the included studies used IRT methods and therefore this domain was also excluded. In addition, we modified the cross-cultural validity box of the COSMIN list. This box evaluates two different aspects: (1) translation of the measure, and (2) the cross-cultural validation analysis between two linguistically different groups. In the present review, part 2 of this box was excluded as no studies in our review conducted multi-group confirmatory factor analysis for different language groups, and thus including this could have led to the studies being automatically rated as poor. To avoid confusion, we therefore refer to the cross-cultural validity dimension as *translation validity* throughout this review. Moreover, studies claiming to address criterion validity but only examining correlations with comparison instruments were not considered to address criterion validity but were evaluated under “hypothesis testing.”

Each item of each domain was scored on a 4-point rating scale (i.e., “poor,” “fair,” “good,” or “excellent”) based on the COSMIN coding manual (Terwee et al., 2012). In subteams of two, the research team members and research assistants independently selected the measurement properties evaluated in the study and scored the relevant items via the above-mentioned ordinal scoring system. In accordance with the COSMIN guidelines (Terwee et al., 2012), methodological quality scores for a study were assigned for each measurement property domain separately by taking the lowest rating of any item in the domain. For instance, if internal consistency was evaluated in a sample and eight of the questions in that domain were ranked as “fair,” two as “excellent,” and one as “poor,” the overall rating for internal consistency in this sample was “poor.” In addition to the nine COSMIN domains, data on interpretability (e.g., the actual psychometric properties) and generalizability (e.g., sample characteristics) were extracted. Two data extraction sheets were designed and used for each of the nine included measures to extract relevant interpretability and generalizability information from the full text of eligible studies. The first data extraction sheet was designed to describe the general characteristics of the study and included: (1) the measure used, (2) the country the study took place in, (3) the language of the measure, (4) the setting the study took place in, (5) the number of participants for each sample, (6) the mean age of participants, and (7) other sample characteristics (e.g., participants' medical or psychiatric diagnosis). The second data extraction sheet included information on the psychometrics of the measure including: (1) instrument version, (2) internal consistency, (3) reliability/test-retest, (4) structural validity, (5) hypothesis testing, and (6) additional information about the psychometric properties of the measure (e.g., content validity and criterion validity).

2.4.2. Rating process. Assessment of methodological quality, extraction of generalizability, and interpretability data were performed by sub-teams of two independent research team members using standardized forms in Excel. Prior to the independent rating, to ensure that all researchers scored the papers in accordance to the guidelines, the research team met to discuss the COSMIN manual and its terminology and ratings. Any discrepancies in ratings were resolved via consensus by the two reviewers, with a third reviewer involved when necessary. The assessments of methodological quality and data extraction were conducted for each sample in each paper (and not each article) to provide as rigorous data as possible.

2.4.3. Quality of the measures. In addition to the methodological quality of the studies, the usefulness of the nine included measures was also evaluated. This was done by combining results on the measurement properties of the different samples, adjusted for their methodological quality in the COSMIN ratings for each measure. As recommended by the Cochrane Back Review Group (Furlan et al., 2009), a best evidence synthesis was performed using categories of 'strong,' 'moderate,' 'limited,' 'conflicting,' or 'unknown' (see Table 2; see online Supplementary material for the specific quality criteria for each measurement property). One research team member rated all measures based on the Cochrane Back Review method and the ratings were double-checked by another research team member, and subsequently discussed with members of the research team. The best evidence synthesis was not performed for the translation validity domain, since this domain is a modified version of the COSMIN's cross-cultural validity box.

3. Results

The quality assessment of the included studies and their samples are described in Appendices 1-8. Sample characteristics and psychometric properties by sample for each measure are reported in Appendices 9-16. The overall evidence rating for each measure is described in Table 3. Below, the results for each measure are summarized.

3.1. The original and revised Body Appreciation Scale

Appendix 9 provides an overview of the 23 studies (including 50 samples) that assessed measurement properties of the BAS and the BAS-2 (Avalos et al., 2005; Tylka & Wood-Barcalow, 2015b). The majority of these samples were university/school samples. Based on the COSMIN guidelines (Terwee et al., 2012) and the Cohrane Back Review method (Furlan et al., 2009), moderate evidence was found for good internal consistency of the BAS, while strong evidence emerged for the BAS-2. Nearly all studies reported Cronbach's alpha ≥ .70. Moderate support was found for good test-retest reliability of both the BAS and BAS-2 with the ICC and Pearson's $r \geq .80$ in all studies examining this property. Conflicting evidence was found for structural validity of the BAS, while strong support was found for the BAS-2. Most studies examining the validity of these measures supported a one-dimensional factor structure using exploratory factor analysis (EFA) or confirmatory factor analysis (CFA). A smaller number of studies supported a two-factor structure for the original BAS, and most of these studies identified a two-factor structure (1) General Body Appreciation and (2) Body Image Investment and were conducted mainly among non-Western samples in China, Indonesia, and Malaysia, with the exception of one study conducted in Poland. Several studies reported that the BAS and BAS-2 were invariant across gender, weight status, ethnic groups, university and community samples, and partly invariant across countries (including Danish, Portuguese, and Swedish samples). Strong support was found for good content validity of the BAS, but only one poor quality study evaluated content validity for the BAS-2. Regarding hypothesis testing, moderate evidence emerged for convergent and discriminant validity of the BAS, and strong evidence for the BAS-2, with studies reporting significant correlations between the BAS/BAS-2 and other body image and well-being measures (e.g., self-esteem). Importantly, the incremental validity of the BAS measures was also supported. Moderate support also emerged for the translation validity of both the BAS and BAS-2. The BAS was translated into different languages including Greek, Brazilian Portuguese, Spanish, Malay, Indonesia, Turkish, Polish, and German, while the BAS-2 was translated into Brazilian Portuguese, Dutch, Persian, French, Danish, European Portuguese, Swedish, Polish, Cantonese, Standard Chinese, and Romanian. Limited support was present for a negative rating of criterion validity of the BAS-2 (correlations with “gold standard” were < .70 and not adequate according to quality criteria), with only one study of fair quality conducted (Tylka et al., 2015).

3.2. The Body Esteem Scale for Adolescents and Adults

Appendix 10 provides an overview of the six studies (including seven samples) that assessed measurement properties of the BESAA (Mendelson et al., 2001). Most studies used American adolescent school samples, and all of the samples were limited to children/adolescents. Strong evidence was found for good internal consistency of the BESAA, and all studies reported Cronbach’s alpha ≥ .70 for the different subscales

(Appearance Esteem, Weight Esteem, and Attribution). Moderate evidence emerged for good test-retest reliability, including Pearson's *r* > .80 for each subscale. Strong evidence was present for good structural validity of the BESAA, with four out of five studies supporting a 3-factor structure. Regarding hypothesis testing, moderate support was found for convergent validity. The three subscales of the BESAA correlated with self-esteem, and other measures of body satisfaction such as the BSQ (Cooper et al., 1987), and eating disorders (e.g., the EDE-Q; Fairburn & Beglin, 1994). Discriminant validity was supported in one study of fair methodological quality (Mendelson et al., 2001). Another study of fair methodological quality supported the incremental validity of the BESAA in the prediction of depression, independent of other components of self-esteem (Jónsdóttir et al., 2008). Evidence for content validity and translation validity was unknown as all sections of the studies assessing these properties were of poor methodological quality.

3.3. The Body Shape Questionnaire

Appendix 11 provides an overview of the 23 studies (including 38 samples) that assessed the measurement properties of the BSQ, including the full 34-item version by Cooper et al. (1987) and shortened versions by Evans and Dolan (1993; 8- and 16-item versions), Dowson and Henderson (2001; 14-item version), and Mazzeo (1999; 10-item version). Most of the studies were conducted in the USA, with the other studies conducted in European countries. Studies included mostly university/school samples. Strong evidence was found for good internal consistency. All studies, with the exception of one, showed a Cronbach's alpha ≥ .70 for the 34-item and shortened versions of the BSQ. Moderate support was found for good test-retest reliability with all studies that examined reliability being of fair or poor quality and reporting Pearson's *r* ≥ .80 or ICC ≥ .70. Regarding validity, strong evidence emerged for good structural validity. For both the short and full forms of the BSQ, most studies supported a one-factor structure. The three studies that did not support a one-factor structure for the BSQ translated the measure into languages other than English (Turkish, Korean, and French). Strong support was also found for content validity of the BSQ. Content validity for the 34-item BSQ was partly supported, with Items 26, 32 (Silva et al., 2014; Silva et al., 2016), and 27 (Silva et al., 2014) deemed inadequate. Results for translation validity could not be evaluated due to poor quality of the translations sections of the studies. Evidence for criterion validity of the BSQ was absent. Regarding hypothesis testing, moderate evidence for convergent validity of the BSQ (full 34-item, 16-item, 14-item and 8-item) emerged, with 17 studies ranging from poor to good methodological quality reporting significant correlations between the BSQ and other body image measures. Moderate evidence for the discriminant validity of the BSQ (full 34-item and 14-item versions) was found, with two studies of fair quality. No study examined the convergent/discriminant validity of the 10-item version of the BSQ. Limited evidence to support good responsiveness of the BSQ emerged in one study of fair methodological quality (Pook et al., 2008).

3.4. The Centre for Appearance Research Valence Scale

Two studies (including three samples) examined the psychometric properties of the CARVAL (Moss & Rosser, 2012; see Appendix 12). Strong evidence was found for good internal consistency of the CARVAL. Both studies reported a Cronbach’s alpha > .80. Moderate negative evidence (i.e., below the threshold for adequate reliability) was found for test-retest reliability, with Pearson’s $r \geq .69$ in both studies. Moderate evidence was found for good structural validity, as a one-factor structure for the CARVAL was supported in one study of good methodological quality. The CARVAL demonstrated strong evidence for good convergent validity. Independent relationships with valid measures of appearance-related psychosocial distress, social anxiety and avoidance in relation to appearance, depression, and anxiety were observed. The support for content validity of the CARVAL is unknown, given that there was only one study that had poor methodological quality for that section. Limited evidence was found for good criterion validity of the measure, with one study of fair methodological quality.

3.5. The Drive for Muscularity Scale

Appendix 13 provides an overview of the 16 studies (including 22 samples) that assessed the psychometric properties of the DMS (McCreary & Sasse, 2000). Strong evidence was found for good internal consistency, with all studies (multiple of excellent quality) examining the DMS reporting a Cronbach's alpha > .70. Moderate evidence for good test-retest reliability emerged, with three studies reporting a Pearson's *r* > .80. Strong evidence was also found for good structural validity. The majority of studies examining the structural validity of the DMS confirmed a two-factor structure: (1) Attitudes and (2) Behaviors. Most of the samples in which structural validity was examined included males only, and the two-factor structure was not supported in female samples. Regarding hypothesis testing, strong support was found for good convergent validity of the DMS. Numerous studies showed that the DMS correlated significantly with other measures of body image (i.e., negatively with the BAS-2; Tylka & Wood-Barcalow, 2015b), and self-esteem. Moderate evidence emerged for good discriminant validity. Support for poor content validity of the DMS was also found, with one study of good quality (Campana et al., 2013), rating some of the items as not relevant to the concept of drive for muscularity and body ideals among Brazilian men (Campana et al., 2013). Moderate support for good translation validity of the DMS was observed. The DMS was shown to be suitable in several populations including among Mexican men (Escoto et al. 2013), Brazilian men (Campana et al., 2013), among French-speaking male athletes (Chaba et al., 2013), Italian men (Nerini et al., 2016), adolescent males in Spain (Sepulveda et al., 2016), Malay men (Swami et al, 2016), and among male university students in Romania (Swami et al., 2018). The evidence was unknown for the criterion validity of the DMS, since the one study reporting criterion validity (Cafri & Thompson, 2004) did not meet the evidence synthesis standards (Furlan et al., 2009).

3.6. The Weight and Shape Concerns Subscales of the Eating Disorders Examination Questionnaire

Appendix 14 provides an overview of the 44 studies (including 63 samples) that assessed psychometric properties of the Weight Concerns (WC) and Shape Concerns (SC) subscales of the EDE-Q (Fairburn & Beglin, 1994). Studies were conducted internationally examining the psychometric properties of the EDE-Q. Strong evidence was found for good internal consistency for the WC and SC subscales of the EDE-Q. In the majority of studies rated as being of excellent quality, alphas were ≥ .70 for both WC and SC subscales. Moderate evidence was found for good test-retest reliability of the WC and SC subscale; the majority of studies examining test-retest reliability reported that the ICC/weighted Kappa ≥ .70 or Pearson's *r* ≥ .80 for these subscales. Strong evidence was found for inadequate structural validity of the WC and SC subscales. The majority of studies examining the factor structure of the EDE-Q identified a combined weight/shape concerns factor, rather than two separate factors for these constructs. The original factor structure of the WC and SC subscales was not supported in five studies of excellent methodological quality; however, two other excellent quality studies conducted in Mexico and Spain (Unikel Santoncini et al., 2018; Villarroel et al., 2011) did confirm WC and SC as two separate factors. Regarding hypothesis testing, strong evidence was found for good convergent validity of the WC and SC subscales of the EDE-Q, while moderate evidence was found for good discriminant validity of these subscales. Moderate evidence was found for good translation validity of the EDE-Q. In 13 studies, good reliability and/or validity of the EDE-Q was shown in Italian (Calugi et al., 2016), French (Carrard et al., 2015), Fijian (Becker et al., 2010), Spanish (Elder & Grilo, 2007), Greek (Giovazolias et al., 2013), German (Hilbert et al., 2012), Finnish (Isomaa et al. ,2016), European Portuguese (Machado et al., 2014), Persian (Mahmoodi et al., 2016), Japanese (Mitsui et al., 2017), Norwegian (Ro et al., 2010), Mexican Spanish (Unikel Santoncini et al., 2018), and Turkish (Yucel el al., 2011) versions of the measure. Moderate evidence was found for good criterion validity of the EDE-Q. The EDE interview (Cooper & Fairburn, 1987) was the criterion measure selected in many studies to assess the criterion validity of the EDE-Q. The WC and SC subscales of the EDE-Q and the EDE interview showed strong correspondence in diverse samples such as adolescent eating disorder samples (Binford et al., 2005), Black US patients with binge eating disorder (Lydecker et al., 2016), Spanish-speaking Latino women (Elder & Grilo, 2007), and a community sample of women (Mond et al., 2004b). Results for content validity of the EDE-Q could not be evaluated due to the poor quality of this section of the only study examining this measurement property (Gideon et al., 2016).

3.7 The Body Dissatisfaction subscale of the Eating Disorder Inventory-3

Appendix 15 provides information on the 11 studies (including 16 samples) that assessed psychometric properties of the Body Dissatisfaction (BD) subscale of the EDI-3 (Garner, 2004). Studies examining the measurement properties of the BD subscale of the EDI-3 were conducted in the USA, Denmark, Netherlands, Sweden, Spain, and Iran. These studies were conducted with university samples, the general population, or in medical settings. Moderate evidence was found for good internal consistency for the BD subscale of the EDI-3. All studies reported a Cronbach’s alpha ≥ .70, with the exception of one study that reported an alpha for the BD subscale of .60 for males (.80 was reported for females) (Dadgostar et al., 2017). Limited evidence emerged for good test-restest reliabilty of the EDI-3 BD subscale, with one study of fair methodological quality (Elosua & Lopez, 2012) reporting Pearson's *r* ≥ .70. Strong support was found for good content validity of the EDI-3 BD subscale. Regarding hypothesis testing, moderate support for good convergent and discriminant validity of the EDI-3 BD subscale emerged. Moderate evidence was found for its criterion validity. The structural validity of the EDI-3 BD subscale was not supported in the studies included in this review given that there were inconsistent findings regarding the factor structure of this measure. Five studies confirmed the unidimensionality of this subscale (Belon et al., 2015; Clausen et al., 2011; Cordero et al., 2013; Lehmann et al., 2013; Rothstein et al., 2017, sample a), while different two-factor structures were reported in four studies (Elosua & Hermosilla, 2013; Kashubeck-West et al., 2013; Rothstein et al., 2017, sample b; Stein et al., 2015). Limited evidence for translation validity of the BD subscale of the EDI-3 emerged. Three studies translated the EDI-3 into different languages, including a Danish version (Clausen et al., 2011), a Spanish version (Elosua & López-Jáuregui, 2012), and an Iranian version (Dadgostar et al., 2017).

3.8. The Appearance Evaluation subscale and Body Areas Satisfaction Scale of the Multidimensional Body Relations Questionnaire

Appendix 16 summarizes the 15 studies (including 19 samples) that assessed the psychometric properties of the MBSRQ (Brown et al., 1990; Cash, 2000) subscales of Appearance Evaluation (AE) and Body Areas Satisfaction Scale (BASS). Most studies were conducted in university/school settings, with studies conducted in ten countries. Strong evidence was found for good internal consistency of the AE and BASS subscales. All studies examining this property reported Cronbach's alphas of ≥ .70 for both these subscales, with the exception of Untas et al. (2009), who examined a French adaptation of the MBSRQ and reported an alpha of .66 for the BASS. Conflicting evidence was found for test-retest reliability and structural validity of both subscales. Regarding hypothesis testing, moderate evidence was found for good convergent and discriminant validity of the AE and BASS subscales. The evidence for criterion validity was unknown. Limited evidence was found to support the responsiveness of the measure. In a prospective study examining patients waiting for breast reduction mammoplasty, Thoma et al. (2005) found evidence to support high responsiveness of the shorter form of the MBSRQ-Appearance Scales (which include the AE and the BASS). Moderate evidence was found to support good translation validity of the MBSRQ. Evidence in these studies support the psychometric properties of the Greek (Argyrides & Kkeli, 2013), Urdu (Naqvi & Kamal, 2017), Spanish (Roncero et al., 2015), French (Untas et al., 2009) and German (Vossbeck-Elsebusch et al., 2014) versions of the MBSRQ, including the AE and BASS subscales.

4. Discussion

The aim of the present systematic review was to rigorously synthesize and critically appraise the psychometric properties of the most influential currently used self-report measures of evaluative aspects of body image. The results revealed that many of these measures have documented evidence of reliability and validity; however, the results were not consistent for all psychometric properties, nor across all populations. Below, the results are further discussed, and recommendations for future research and clinical practice are detailed.

4.1. Recommendations and Considerations by Measure

The original and revised Body Appreciation Scale (BAS and BAS-2; Avalos et al., 2005; Tylka & Wood-Barcalow, 2015b) generally displayed good psychometric properties across samples of different age and gender. However, the evidence for structural validity for the original BAS was conflicting with a two-factor structure reported in several non-Western countries. Given that the evidence for structural validity was excellent for the BAS-2, and since the evidence for internal consistency, as well as convergent and divergent validity, was stronger for the BAS-2 than the original BAS, the BAS-2 is recommended for use in future studies. However, it must also be noted that findings for the BAS-2 were inconclusive regarding content validity, and negative regarding criterion validity, and future studies examining the BAS-2 should consider evaluating those properties more thoroughly. The importance of assessing body appreciation in clinical practice has previously been mentioned (Tylka & Wood-Barcalow, 2015a); however, to date, no study has evaluated the BAS or BAS-2 in clinical populations, which presents another focus for future research.

The Body Esteem Scale for Adolescents and Adults (BESAA; Mendelson, Mendelson & White, 2001) generally displayed good internal consistency, test-retest reliability, structural validity, and convergent validity among both female and male children and adolescents. However, this measure is also widely used among adults (Thompson et al., 2012), and future studies are encouraged to evaluate psychometric properties in adult samples. Results for the BESAA's content validity and translational validity were inconclusive, and future studies using non-English versions of the measure should consider adopting a more thorough cross-cultural validation process to ensure the validity of the measure. Since the BESAA has not been evaluated in clinical settings, and its criterion validity and responsiveness have not been evaluated, this measure is not currently recommended for assessing body image in clinical groups, and future studies are encouraged to evaluate the BESAA in clinical settings.

Evidence supported scores on the Body Shape Questionnaire (BSQ; Cooper et al., 1987) as reliable and valid within a wide range of clinical and non-clinical, mainly female and White Western, populations. Interestingly, this was also true for all short versions of the measure. Since short versions of a measure have a number of advantages over longer versions in terms of participant burden and interpretation, the use of short versions is recommended for future studies and in clinical practice. However, it must also be taken into consideration that the one-factor structure of the measure was not supported in some of the translated versions of the BSQ and that the translation processes of the instrument generally were of poor quality. Hence, the cross-cultural validity of the BSQ can be questioned, and future studies are suggested to adopt more thorough cross-cultural validation processes before using translated versions of the instrument.

As for the Centre for Appearance Research Valence Scale (CARVAL; Moss & Rosser, 2012), the inclusion of this measure was justified by its importance to and future potential within visible difference research. The CARVAL is one of very few evaluative body image measures designed with people with visible differences in mind and has also been previously used in one of the largest samples of adults (*N* = 1,265) with visible differences, which demonstrated that this measure was psychometrically sound for this population (Moss et al., 2014). However, the evidence for reliability and validity was limited to two UK based studies. Although the results from those two studies (Moss et al., 2012, 2014) were promising, more research evaluating this measure is needed.

The Drive for Muscularity Scale (DMS; McCreary & Sasse, 2000) was considered to be an important body image measure as it concerns a significant area of male body dissatisfaction – muscularity. Consequently, this measure was mainly evaluated in various male populations of varying age, sexual orientation, athletes/non-athletes) with good results in terms of validity and reliability. Generally, a two-factor structure was supported among males: Attitudes and Behaviors. The Behaviors subscale does not meet our definition of evaluative body image, but both subscales were included in the present review since the developers did not originally make this division (McCreary & Sasse, 2000), and the two-factor structure was based on subsequent factor analyses rather than a theoretical distinction. Future studies assessing attitudinal aspects of male body dissatisfaction could consider using only the attitudinal subscale. The content validity of the DMS was not fully supported cross-culturally, which is important to take into consideration when using the measure in non-Western contexts. Moreover, the DMS has not been evaluated psychometrically in clinical samples which is an important focus of future evaluations of the measure.

Scores on the Weight and Shape Concerns (WC, SC) subscales of the EDE-Q (Fairburn & Beglin, 1994) were generally considered valid and reliable, in a wide range of mainly female populations. However, the structural validity of these subscales was not supported due to the tendency of the subscales to load onto the same factor. Hence, future studies need to consider that the WC and SC subscales might not measure two distinct aspects of body image, but rather be expressions of the same construct. Since the WC and SC subscales assess clinically significant body dissatisfaction (Krawczyk et al., 2012), and have been evaluated with good results in clinical settings (mainly among patients with eating disorders), the use of the WC and SC subscales are recommended for assessing evaluative body image in clinical groups. Good criterion validity in terms of high correspondence between the EDE-Q and the EDE interview further supports the WC and SC subscales to assess evaluative body image among patients with eating disorders.

As for the Body Dissatisfaction subscale (BD) of the EDI-3 (Garner, 2004), good evidence for the reliability and validity was found in various, mainly female, populations. Evidence for structural validity was conflicting since some studies (with African-American, Mexican-American, and Spanish participants) reported different two-factor structures for the subscale. Therefore, future studies assessing evaluative body image using the BD subscale of the EDI-3 should bear in mind that the cross-culturally validity of the subscale might be limited. As with the WC and SC subscales of the EDE-Q, the BD subscale of EDI-3 displayed good criterion validity and was evaluated in clinical settings (mainly among patients with eating disorders), and can therefore be considered to assess evaluative body image in such settings.

The Appearance Evaluation subscale (AE) and Body Areas Satisfaction Scale (BASS) of the Multidimensional Body Relations Questionnaire (MBSRQ; Brown et al., 1990) displayed adequate psychometric properties in terms of internal consistency, structural validity, convergent validity, translation validity, and responsiveness, in different (Western and mainly female) samples. Thus, the AE and the BASS are recommended for use in non-clinical samples. Although the AE subscale and the BASS have been evaluated in clinical settings (patients with eating disorders and patients waiting for reduction mammoplasty), these studies are scarce and more evaluations in clinical settings are needed.

4.2. Overall Recommendations Concerning Body Image Measurements

The initial search for measures to include in this review revealed 58 recently used measures that met our definition of evaluative body image, and over 150 measuring body image more broadly. Many of these measures had been developed for, and used in, only one specific study. We strongly recommend that researchers think twice before developing new body image measures to assess evaluative body image, since scale development is a demanding and onerous process (see Krawczyk et al., 2012). Importantly, the results from the present systematic review indicate that sufficiently well-established and psychometrically sound measures exist to assess evaluative body image in various populations. Future studies should primarily focus on the further evaluation of already existing measures to move the body image research field forward. Yet, if a construct is revealed that cannot be tapped by existing body image measures, researchers may want to create a measure to assess it. Within their psychometric investigation, there should be examinations of incremental validity; that is, the developed measure predicts some criterion above and beyond existing body image measures. For example, the Functionality Appreciation Scale (Alleva, Tylka, & Kroon Van Diest, 2017) has been shown to predict unique variance in well-being above and beyond body appreciation and other measures of body image.

A number of recommendations for future studies evaluating psychometric properties of body image measures can be made. First, regarding all measures included in the present review, studies of measurement error (i.e., the systematic and random error of a score that is not attributed to true changes in the construct to be measured; Mokkink et al., 2012a) were completely lacking. Future studies are recommended to investigate measurement error in the evaluation process of a measure, preferably using standard error of measurement (Mokkink et al., 2012a). Internal consistency was primarily evaluated using Cronbach’s alpha. However, it is important to acknowledge that while Cronbach’s alpha provides an estimate of internal consistency, an adequate Cronbach’s alpha value does not necessarily mean that a group of scale scores are internally consistent (i.e., a large pool of items with low inter-item correlations may have a high Cronbach’s alpha coefficient). Moreover, many studies evaluating test-retest reliability received poor quality ratings because they only reported the correlation of the measure between the time points. According to the COSMIN guidelines (Mokkink et al., 2012a; Terwee et al., 2012), the use of merely the Pearson’s and Spearman’s correlation coefficients is considered inadequate as it fails to take systematic error into account. Future studies assessing test-retest are recommended to provide evidence that no systematic error occurs between time points, for instance using the intraclass correlation coefficient (ICC; for continuous scores).

Cross-cultural validity could not be fully determined for any of the included measures, due to a lack of multi-group confirmatory factor analyses (MGCFA) and differential item functioning (DIF) with different language groups. Moreover, many studies received poor ratings regarding their translational processes. Studies tended to use a single translation-back-translation methodology, which does not meet the COSMIN guidelines standards (Terwee et al., 2012). To establish good translation validity, the measure should undergo multiple forward and multiple backward translations and the final translated version of the measure should be pilot-tested. Moreover, translators should work independently, report how inconsistencies were resolved, and preferably the translation would be reviewed by a committee (involving people other than the translators, e.g., the original scale developers). Future studies should consider a more rigorous translation processes, as well as the use of MGCFAs and DIF, to ensure both translation validity and cross-cultural validity in the measures.

Regarding criterion validity, many studies were considered inconclusive because they did not sufficiently justify the choice of a gold standard criterion. According to the COSMIN guidelines, authors are frequently overly generous in their choice of a gold standard, for example by assuming that instruments would qualify as such on the basis of being widely used (Mokkink et al., 2012a). For the WC and SC subscales of the EDE-Q, criterion validity could be confirmed as studies tended to use the EDE interview as criteria. Other strong examples of gold standard criteria according to COSMIN (Mokkink et al., 2012a), are longer versions (with established criterion validity) of the same measure that is being evaluated.

The findings of the present review also strongly highlight the importance of using gender-appropriate measures. As concluded by Krawczyk et al. (2012) and Cash (2011), measures of body image must often be adapted, modified, or created separately for different gender groups due to differences in appearance ideals. Consistent with this, assessment tools have been developed to target gendered appearance concerns. The present review led to a number of important recommendations regarding assessment for different gender groups. In multiple studies, the BAS and BAS-2 were shown to be invariant across gender, supporting their use among both among individuals who identify as male and female, and their usefulness for examining gender differences among these gender identities. Based on the studies evaluating the BESAA, the two studies evaluating the CARVAL, and the studies evaluating the AE and the BASS of the MBSRQ, these measures accrued evidence of validity and reliablity among female and male samples. However, evidence is limited and comparisons across genders must be made with caution given support for their invariance across gender is lacking. The BSQ, WC and SC subscales of the EDE-Q, and the BD subscale of the EDI-3 focus on dissatisfaction with body fat typically associated with female body image (Krawzcyk et al., 2012), and have subsequently been mostly evaluated in female samples, which limits our understanding of the usefulness of these measures among males. Regarding the EDE-Q, it was found that the WC and SC subscales of the EDE-Q were invariant across gender among Mexican adolescents (Penelo et al., 2013). Nevertheless, future studies should investigate whether this result is generalizable to other populations. Furthermore, the focus on body fat suggests that these measures may miss important aspects of male body image. In contrast, the DMS is tailored to measure drive for muscularity typically associated with male body image and in the present review, the great majority of evidence was derived from all-male samples. The few exceptions among female samples (Cafri & Thompson, 2004, sample b; McCreary et al., 2004, sample b; Wojtowicz & von Ranson, 2006, sample a), failed to confirm the two-factor structure described among males. However, the psychometric properties for a one-factor structure were adequate, suggesting that this measure may be useful among females as well. Future studies assessing whether the DMS is invariant across gender, as well as extending its evaluation beyond male samples, are warranted. Moreover, future studies should include evalutations of content validity of the DMS in female samples since all items may not be applicable to women in general (e.g., "I think that I would look better if I gained 10 pounds in bulk"). In addition, studies extending the evidence for the usefulness of measures of muscularity concerns designed specifically for use among females (e.g., Rodgers et al., 2018) would be useful.

In relation to questions about generalizability more broadly, most studies included in this review tended to use female-only samples and additionally relied on data derived from White, Western, heterosexual, and school- or university participants. Overall, more studies are needed to evaluate psychometric properties of body image measures across genders, cultural contexts, clinical conditions, as well as sexual orientations and other dimensions of identity. Notably, not a single study was found that had evaluated any of the body image measures in a sample of transgender participants, which is an important area for future studies given recent findings of high rates of body dissatisfaction in this population (e.g., Jones, Haycraft, Murjan, & Arcelus, 2016).

4.3. Limitations and Strengths

The results of the present systematic review should be interpreted in the light of its limitations. The first limitation concerns our definition of body image. Body image is a multifaceted concept (Cash, 1994), and the exclusive focus on evaluative body image, may have led to many widely used behavioral and cognitive instruments being excluded. However, this decision was based on the notion that measures of body (dis)satisfaction probably are the most commonly used (Krawcyk et al., 2012), and that evaluative body image is the facet most commonly referred to as body image (Cash, 2011). Additional rigorous systematic reviews focusing on other aspects of body image would make valuable contributions to the literature. For instance, two important future foci would be to conduct systematic reviews evaluating state body image measures as well as body image measures for children. In addition, body image silhouette measures (e.g., figural drawing scales), were excluded due to their failing to provide an explicit assessment of the evaluative component that was the focus of the current review despite being frequently used to assess body dissatisfaction (Gardner & Brown, 2010). Readers interested in an overview of silhouette measures are referred to the review by Gardner and Brown (2010).

Another limitation concerns the inclusion of measures in the present review. Although the priority ranking was systematized following established guidelines (Guyatt et al., 2011), this method is inherently subjective, and the included measures are a reflection of the 10 authors’ collective perspections of influential measures. Hence, another team of body image researchers might have chosen other body image measures to include in the review.

However, since no established method exists to objectively choose instruments to include in a systematic review of measures, our approach to follow the GRADE guidelines (Guyatt et al., 2011), in conjuction with a search for number of citations on Google scholar, was considered most appropriate. Further, as concerns the screening and extraction processes, it would have been favourable to estimate the inter-rater reliability in order to support the reliability of these processes. However, the chosen approach was deemed comprehensive, as it followed the recommendations by COSMIN which includes to complete the checklist by at least two independent raters, and to reach concensus on one final rating using a third rater when necessary.

Of the initial number of studies identified in the Step 2 search, a large number were excluded, which could be viewed as a limiting factor, e.g., in that studies reporting one aspect of psychometric properties not were included. However, as described in the Method section, the main reason for exclusion (both after title/abstract review and full-text review) was that the study had not used the measure/subscales of interest. Moreover, regarding the number of excluded studies, our exclusion rates are well in line with previous similar systematic reviews of measures using the COSMIN (see Balzer, van der Linden, Mercer, van Hedel, 2017; Evans, Spiby, & Morell, 2015; Matarese, Lommi, & De Marinis, 2017; Speyer et al., 2018, Weldam, Schurmans, Liu, & Lammers, 2013). Although it might have been informative to include more studies, the decision to include only papers with dedicated sections on psychometrics is in line with the COSMIN methodology, and standard procedure in systematic reviews evaluating both the quality of the included studies and the actual measures. Moreover, the research team concluded that only including studies dedicated to psychometric evaluation of the measures of interest was the only feasible approach, as otherwise all studies that used one of the measures and reported correlations with other measures or Cronbach’s alpha for their sample (which almost all studies do) would have been included. Moreover, including all studies reporting internal consistency statistics would not have added much information, since all measures had moderate or strong positive evidence for internal consistency. Furthermore, Cronbach's alpha alone is not sufficient evidence to determine the psychometric properties of a measure.

Regarding measure inclusion, one limitation concerns specifically the EDI-3. We decided to only include articles evaluating the most recent version of the EDI, i.e., the EDI-3, given that the third version is the most used in recent years, and addresses some important psychometric issues of the EDI-2 (e.g., problematic factor structure; Cumella, 2006). However, the BD subscale of EDI-3 is similar to previous versions of the EDI and including studies on previous versions of this measure would probably have added more studies to the review. Another limitation concerns other studies that were excluded from our review. For instance, conference and symposium abstracts were excluded, although the authors of the abstracts were contacted to determine if a full study report was available for inclusion. Non-English language studies were also excluded despite potentially containing psychometric information. In total, 21 articles written in Japanese, German, Persian, Dutch, Greek, Chinese, Spanish, French, and Hungarian were affected by this decision. These non-English articles were excluded as our research team only had fluency in half of these languages. For the procedure used we also needed to have at least two researchers that had knowledge in every language which was often not the case. Although English articles of researchers from these countries were included in the present review, this may have limited the findings regarding the usefulness of the included measures across cultures.

Limitations also exist concerning the rating of the level of evidence for each measurement property for each measure. For instance, in accordance with the COSMIN guidelines (Terwee et al., 2012), methodological quality scores for a study were assigned for each measurement property domain separately by taking the lowest rating of any item in the domain (i.e., “worst score counts”). This method might seem overly restrictive, resulting in below average ratings of quality. However, as all items assessed for each measurement property are inter-related, this method was deemed the most appropriate. For instance, if the ratings for sample size is poor, this is likely to affect other aspects of the same measurement property even if those aspects are reported by approved means.

Important strengths with the present systematic review include the use of the COSMIN methodology to provide a structured way of assessing all measures in a consistent way, as well as the assessment of level of evidence for quality of measurement properties. No previous body image review has adopted this approach. In sum, our review is a comprehensive assessment of evaluative body image measures that no other study has completed to date and it directly addresses a gap in the literature.

4.4. Conclusion

The present systematic review synthesized and critically evaluated currently used influential self-report measures of evaluative body image. The results revealed support for the majority of the measures in terms of adequate reliability and validity, although suitability varied across populations, and some measurement properties were insufficiently evaluated. Future studies should primarily focus on extending the available evidence for already existing measures rather than developing new measures of evaluative body image. Overall, more studies examining the psychometric properties of body image measures across different populations, focusing on cross-cultural validity, are warranted. Additional systematic reviews of body image measures are also needed in order to continue to build towards a cohesive core group of measures that will promote the comparability of findings across studies and support the growth of our field.

1

Acknowledgements. The authors thank Chris Jarrett for developing the search strategies and performing the initial searches, and Alejandro Crespo and David Kjellin for assisting with the updated searches. We thank Diana Harcourt and Brett Thombs for their guidance and support. We also thank Rachel Berry, Nadia Craddock, Russel DuBois, Kirsty Garbett, Emma Gaskin, Lisa Jewett, Caroline Järdmo, Hanna Larsson, Alice Lowy, and Kelcey O Brien for assisting the project. Finally, we thank the European Cooperation in Science and Technology (*COST) Action IS1210* Appearance Matters for enabling our collaboration.

References

Akdemir, A., Inandi, T., Akbas, D., Kahilogullari, A. K., Eren, M., & Canpolat, B. I. (2012). Validity and reliability of a Turkish version of the Body Shape Questionnaire among female high school students: Preliminary examination. *European Eating Disorders Review*, *20*, 2011–2012. http://doi.org/10.1002/erv.1106

Alcaraz-Ibáñez, M., Cren Chiminazzo, J. G., Sicilia, Á., & Teixeira Fernandes, P. (2017). Examining the psychometric properties of the Body Appreciation Scale-2 in Brazilian adolescents. *Psychology, Society and Education, 9*, 505–515. http://doi.org/10.25115/psye.v9i3.1101

Alexias, G., Togas, C., & Mellon, R. (2016). Psychometric properties of the Greek version of the Body Appreciation Scale. *Hellenic Journal of Psychology*, *13*, 73–92.

Allen, K. L., Byrne, S. M., Lampard, A., Watson, H., & Fursland, A. (2011). Confirmatory factor analysis of the Eating Disorder Examination-Questionnaire (EDE-Q). *Eating Behaviors*, *12*, 143–151. http://doi.org/10.1016/j.eatbeh.2011.01.005

Alleva, J. M., Martijn, C., Veldhuis, J., & Tylka, T. L. (2016). A Dutch translation and validation of the Body Appreciation Scale-2: An investigation with female university students in the Netherlands. *Body Image, 19*, 44-48. https://doi.org/10.1016/j.bodyim.2016.08.008

Alleva, J. M., Tylka, T. L., & Kroon Van Diest, A. M. (2017). The Functionality Appreciation Scale (FAS): Development and psychometric properties in U.S. community women and men. *Body Image*, *23*, 28-44. doi:10.1016/j.bodyim.2017.07.008

Argyrides, M., & Kkeli, N. (2013). Multidimensional Body-Self Relations Questionnaire-Appearance Scales: Psychometric properties of the Greek version. *Psychological Reports*, *113*, 885–897. http://doi.org/10.2466/03.07.PR0.113x29z6

Atari, M. (2016). Factor structure and psychometric properties of the Body Appreciation Scale-2 in Iran. *Body Image*, *18*, 1–4. http://doi.org/10.1016/j.bodyim.2016.04.006

Avalos, L., Tylka, T. L., & Wood-barcalow, N. (2005). The Body Appreciation Scale: Development and psychometric evaluation, *2*, 285–297. http://doi.org/10.1016/j.bodyim.2005.06.002

Balzer, J., van der Linden, M. L., Mercer, T. H., & van Hedel, H. J. A. (2017). Selective voluntary motor control measures of the lower extremity in children with upper motor neuron lesions: A systematic review. *Developmental Medicine & Child Neurology, 59,* 699-705. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1111/dmcn.13417

Bardone-Cone, A. M., & Boyd, C. A. (2007). Psychometric properties of eating disorder instruments in Black and White young women: Internal consistency, temporal stability, and validity. *Psychological Assessment*, *19*, 356–362. http://doi.org/10.1037/1040-3590.19.3.356

Barnes, J., Prescott, T., & Muncer, S. (2012). Confirmatory factor analysis for the Eating Disorder Examination Questionnaire: Evidence supporting a three-factor model. *Eating Behaviors*, *13*, 379–381. http://doi.org/10.1016/j.eatbeh.2012.05.001

Becker, A. E., Thomas, J. J., Bainivualiku, A., Richards, L., Navara, K., Roberts, A. L., … Striegel-Moore, R. H. (2010). Validity and reliability of a Fijian translation and adaptation of the Eating Disorder Examination Questionnaire. *International Journal of Eating Disorders*, *43*, 171–178. http://doi.org/10.1002/eat.20675

Belon, K. E., McLaughlin, E. A., Smith, J. E., Bryan, A. D., Witkiewitz, K., Lash, D. N., & Winn, J. L. (2015). Testing the measurement invariance of the Eating Disorder Inventory in nonclinical samples of Hispanic and Caucasian Women. *International Journal of Eating Disorders*, *48*, 262–270. http://doi.org/10.1002/eat.22286

Binford, R. B., Le Grange, D., & Jellar, C. C. (2005). Eating Disorders Examination versus Eating Disorders Examination-Questionnaire in adolescents with full and partial-syndrome bulimia nervosa and anorexia nervosa. *International Journal of Eating Disorders*, *37,* 44–49. http://doi.org/10.1002/eat.20062

Brown, T. A., Cash, T. F., & Mikulka, P. J. (1990). Attitudinal body-image assessment: Factor analysis of the body-self relations questionnaire. *Journal of Personality Assessment, 55*, 135-144. https://doi.org/10.1207/s15327752jpa5501&2_13

Brytek-Matera, A., & Rogoza, R. (2015). Validation of the Polish version of the Multidimensional Body-Self Relations Questionnaire among women. *Eating and Weight Disorders*, *20*, 109–117. http://doi.org/10.1007/s40519-014-0156-x

Cafri, G., & Thompson, J. K. (2004). Evaluating the convergence of muscle appearance attitude measures. *Assessment*, *11*, 224–229. http://doi.org/10.1177/1073191104267652

Calugi, S., Milanese, C., Sartirana, M., El Ghoch, M., Sartori, F., Geccherle, E., … Dalle Grave, R. (2017). The Eating Disorder Examination Questionnaire: Reliability and validity of the Italian version. *Eating and Weight Disorders*, *22*, 509–514. http://doi.org/10.1007/s40519-016-0276-6

Campana, A. N. N. B., Tavares, M. da C. G. C. F., Swami, V., & da Silva, D. (2013). An examination of the psychometric properties of Brazilian Portuguese translations of the Drive for Muscularity Scale, the Swansea Muscularity Attitudes Questionnaire, and the Masculine Body Ideal Distress Scale. *Psychology of Men and Masculinity*, *14,* 376–388. http://doi.org/10.1037/a0030087

Carrard, I., Lien Rebetez, M. M., Mobbs, O., & Van der Linden, M. (2015). Factor structure of a French version of the Eating Disorder Examination-Questionnaire among women with and without binge eating disorder symptoms. *Eating and Weight Disorders*, *20*, 137–144. http://doi.org/10.1007/s40519-014-0148-x

Cash, T. F. (1994). Body-image attitudes: Evaluation, investment, and affect. *Perceptual and Motor Skills, 78*, 1168-1170. https://doi.org/10.2466/pms.1994.78.3c.1168

Cash, T. F. (2000). *User's manual for the Multidimensional Body Self-Relations Questionnaire.* Available from the author at www.body-images.com

Cash (2011). Crucial considerations in the assessment of body image. In T. F. Cash & L. Smolak (Eds.), *Body image. A handbook of science, practice, and prevention* (pp. 129-137). New York, NY: The Guilford Press.

Chaba, L., d'Arripe-Longueville, F., Lentillon-Kaestner, V., & Scoffier-Mériaux, S. (2018). Adaptation and validation of a short French version of the Drive for Muscularity Scale in male athletes (DMS-FR). *PLoS ONE*, *13*, 1–15. http://doi.org/10.1371/journal.pone.0196608

Chan, C. W., & Leung, S. F. (2015). Validation of the Eating Disorder Examination Questionnaire: An online version. *Journal of Human Nutrition and Dietetics, 28*, 659–665. http://doi.org/10.1111/jhn.12275

Clausen, L., Rosenvinge, J. H., Friborg, O., & Rokkedal, K. (2011). Validating the Eating Disorder Inventory-3 (EDI-3): A comparison between 561 female eating disorders patients and 878 females from the general population. *Journal of Psychopathology and Behavioral Assessment, 33*, 101–110. http://doi.org/10.1007/s10862-010-9207-4

Compte, E. J., Sepúlveda, A. R., de Pellegrin, Y., & Blanco, M. (2015). Confirmatory factor analysis of the Drive for Muscularity Scale-S (DMS-S) and Male Body Attitudes Scale-S (MBAS-S) among male university students in Buenos Aires. *Body Image*, *14*, 13–19. http://doi.org/10.1016/j.bodyim.2015.02.005

Confalonieri, E., Gatti, E., Ionio, C., & Traficante, D. (2008). Body Esteem Scale: A validation on Italian adolescents. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *15*, 153–165.

Conti, M. A., Cordás, T. A., & Latorre, M. R. D. O. (2009). A study of the validity and reliability of the Brazilian version of the Body Shape Questionnaire (BSQ) among adolescents. *Revista Brasileira De Saúde Materno Infantil, 9*, 331-338. https://doi.org/10.1590/S1519-38292009000300012

Cooper, Z., & Fairburn, C. (1987). The Eating Disorder Examination: A semi-structured interview for the assessment of the specific psychopathology of eating disorders. *International Journal of Eating Disorders, 6*, 1-8. https://doi.org/10.1002/1098-108X(198701)6:1%3C1::AID-EAT2260060102%3E3.0.CO;2-9

Cooper, P. J., Taylor, M. J., Cooper, Z., & Fairburn, C. G. (1987). The development and validation of the Body Shape Questionnaire. *International Journal of Eating Disorders*, *6*, 485–494. http://doi.org/10.1002/1098-108X(198707)6:4<485::AID-EAT2260060405>3.0.CO;2-O

Cordero, E. D., Julian, A. K., & Murray, K. E. (2013). Measurement of disordered eating in Latina college women. *Eating Behaviors*, *14*, 220–223. http://doi.org/10.1016/j.eatbeh.2012.12.006

Cotter, E. W., Kelly, N. R., Mitchell, K. S., & Mazzeo, S. E. (2015). An investigation of body appreciation, ethnic identity, and eating disorder symptoms in Black women. *Journal of Black Psychology*, *41*, 3–25. http://doi.org/10.1177/0095798413502671

Cragun, D., Debate, R. D., Ata, R. N., & Thompson, J. K. (2013). Psychometric properties of the Body Esteem Scale for Adolescents and Adults in an early adolescent sample. *Eating and Weight Disorders*, *18*, 275–282. http://doi.org/10.1007/s40519-013-0031-1

Cruzat-Mandich, C., Díaz-Castrillón, F., Pérez-Villalobos, C. E., Lizana, P., Moore, C., Simpson, S., & Oda-Montecinos, C. (2019). Factor structure and reliability of the Multidimensional Body–Self Relations Questionnaire in Chilean youth. *Eating and Weight Disorders*, *24*, 339-350. http://doi.org/10.1007/s40519-017-0411-z

Cumella, E. J. (2006). Review of the Eating Disorder Inventory-3. *Journal of Personality Assessment, 87*, 116-117. https://doi.org/10.1207/s15327752jpa8701_11

Silva, W. R., Dias, J. C. R., Maroco, J., & Campos, J. A. D. B. (2014). Confirmatory factor analysis of different versions of the Body Shape Questionnaire applied to Brazilian university students. *Body Image, 11*, 384–390. http://doi.org/10.1016/j.bodyim.2014.06.001

Dadgostar, H., Nedjat, S., Dadgostar, E., & Soleimany, G. (2017). Translation and evaluation of the reliability and validity of Eating Disorder Inventory-3 questionnaire among Iranian university students. *Asian Journal of Sports Medicine*. Advance online publication. http://doi.org/10.5812/asjsm.13950

Darcy, A. M., Hardy, K. K., Crosby, R. D., Lock, J., & Peebles, R. (2013). Factor structure of the Eating Disorder Examination Questionnaire (EDE-Q) in male and female college athletes. *Body Image, 10*, 399–405. http://doi.org/10.1016/j.bodyim.2013.01.008

Deblaere, C., & Brewster, M. E. (2017). A confirmation of the Drive for Muscularity Scale with sexual minority men. *Psychology of Sexual Orientation and Gender Diversity, 4*, 227–232. http://doi.org/10.1037/sgd0000224

Di Pietro, M., & Da Silveira, D. X. (2009). Internal validity, dimensionality and performance of the Body Shape Questionnaire in a group of Brazilian college students. *Revista Brasileira de Psiquiatria, 31*, 21–24. http://doi.org/10.1590/S1516-44462008005000017

Dowson, J., & Henderson, L. (2001). The validity of a short version of the Body Shape Questionnaire. *Psychiatry Research, 102*, 263–271. http://doi.org/10.1016/S0165-1781(01)00254-2

Elder, K. A., & Grilo, C. M. (2007). The Spanish language version of the Eating Disorder Examination Questionnaire: Comparison with the Spanish language version of the

Eating Disorder Examination and test-retest reliability. *Behaviour Research and Therapy*, *45*, 1369–1377. http://doi.org/10.1016/j.brat.2006.08.012

Elosua, P., & Hermosilla, D. (2013). Does body dissatisfaction have the same meaning for males and females? A measurement invariance study. *Revue Europeenne de Psychologie Appliquee*, *63*, 315–321. http://doi.org/10.1016/j.erap.2013.06.002

Elosua, P., & López-Jáuregui, A. (2012). Internal structure of the Spanish adaptation of the Eating Disorder Inventory-3. *European Journal of Psychological Assessment*, *28*, 25–31. http://doi.org/10.1027/1015-5759/a000087

Escoto, C., Alvarez-Rayón, G., Mancilla-Díaz, J. M., Camacho Ruiz, E. J., Franco Paredes, K., & Juárez Lugo, C. S. (2013). Psychometric properties of the Drive for Muscularity Scale in Mexican males. *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity*, *18*, 23–28. http://doi.org/10.1007/s40519-013-0010-6

Evans, C., & Dolan, B. (1993). Body Shape Questionnaire: Derivation of shortened "alternative forms." *International Journal of Eating Disorders, 13,* 315-321. https://doi.org/10.1002/1098-108X(199304)13:3

Evans, K., Spiby, H., & Morrell, C. J. (2015). A psychometric systematic review of self-report instruments to identify anxiety in pregnancy. *Journal of Advanced Nursing, 71,* 1986-2001. doi: 10.1111/jan.12649.

Fairburn, C. G., & Bèglin, S. J. (1994). Assessment of eating disorders: Interview or self-report questionnaire? *International Journal of Eating Disorders, 16*, 363-370.

Ferreira, L., Neves, A. N., & da Consolação Gomes Cunha Fernandes Tavares, M. (2014). Validity of body image scales for Brazilian older adults. *Motriz. Revista de Educacao Fisica*, *20*, 359–373. http://doi.org/10.1590/S1980-65742014000400002

Franko, D. L., Jenkins, A., Roehrig, J. P., Luce, K. H., Crowther, J. H., & Rodgers, R. F. (2012). Psychometric properties of measures of eating disorder risk in Latina college women. *International Journal of Eating Disorders*, *45,* 592–596. http://doi.org/10.1002/eat.20979

Furlan, A. D., Pennick, V., Bombardier, C., & Van Tulder, M. (2009). 2009 Updated method guidelines for systematic reviews in the Cochrane back review group. *Spine, 34*, 1929-1941. https://doi.org/10.1097/BRS.0b013e3181b1c99f

Gallini, S. M. A. (2008). *The relationship of children's body dissatisfaction with five domains of life satisfaction* (Doctoral dissertation). Available from ProQuest Dissertations and Theses. (3283109)

Gardner, R. M., & Brown, D. L. (2010). Body image assessment: A review of figural drawing scales. *Personality and Individual Differences, 48*, 107-111. https://doi.org/10.1016/j.paid.2009.08.017

Garner, D. M. (2004). Eating Disorder Inventory-3 professional manual. Odessa, FL: Psychological Assessment Resources.

Ghaderi, A., & Scott, B. (2004). The reliability and validity of the Swedish version of the Body Shape Questionnaire. *Scandinavian Journal of Psychology*, *45*, 319–324. http://doi.org/10.1111/j.1467-9450.2004.00411.x

Gideon, N., Hawkes, N., Mond, J., Saunders, R., Tchanturia, K., & Serpell, L. (2016). Development and psychometric validation of the EDE-QS, a 12-item short form of the Eating Disorder Examination Questionnaire (EDE-Q). *PLoS ONE, 11*, 1–19. http://doi.org/10.1371/journal.pone.0152744

Giovazolias, T., Tsaousis, I., & Vallianatou, C. (2013). The factor structure and psychometric properties of the Greek version of the eating disorders examination questionnaire (EDE-Q). *European Journal of Psychological Assessment*, *29*, 189–196. http://doi.org/10.1027/1015-5759/a000138

Griffiths, C., Guest, E., White, P., Gaskin, E., Rumsey, N., Pleat, J., & Harcourt, G. (2017). A systematic review of patient-reported outcome measures used in adult burn research. *Journal of Burn Care & Research, 38,* 521-545. https://doi.org/10.1097/BCR.0000000000000474

Grilo, C. M., Henderson, K. E., Bell, R. L., & Crosby, R. D. (2013). Eating Disorder Examination-Questionnaire factor structure and construct validity in bariatric surgery candidates. *Obesity Surgery*, *23*, 657–662. http://doi.org/10.1007/s11695-012-0840-8

Grilo, C. M., Reas, D. L., Hopwood, C. J., & Crosby, R. D. (2015). Factor structure and construct validity of the Eating Disorder Examination-Questionnaire in college students: Further support for a modified brief version. *International Journal of Eating Disorders*, *48*, 284–289. http://doi.org/10.1002/eat.22358

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2011). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336,* 924-926. doi: https://doi.org/10.1136/bmj.39489.470347.AD

Heiss, S., Boswell, J. F., & Hormes, J. M. (2018). Confirmatory factor analysis of the Eating Disorder Examination-Questionnaire: A comparison of five factor solutions across vegan and omnivore participants. *International Journal of Eating Disorders*, *51*, 418–428. http://doi.org/10.1002/eat.22848

Hilbert, A., de Zwaan, M., & Braehler, E. (2012). How frequent are eating disturbances in the population? Norms of the Eating Disorder Examination-Questionnaire. *PLoS ONE*, *7*(1). http://doi.org/10.1371/journal.pone.0029125

Hrabosky, J. I., White, M. A., Masheb, R. M., Rothschild, B. S., Burke-Martindale, C. H., & Grilo, C. M. (2008). Psychometric evaluation of the Eating Disorder Examination-

Questionnaire for bariatric surgery candidates. *Obesity*, *16*, 763–769. http://doi.org/10.1038/oby.2008.3

Isomaa, R., Lukkarila, I. L., Ollila, T., Nenonen, H., Charpentier, P., Sinikallio, S., & Karhunen, L. (2016). Development and preliminary validation of a Finnish version of the Eating Disorder Examination Questionnaire (EDE-Q). *Nordic Journal of Psychiatry*, *70*, 542–546. http://doi.org/10.1080/08039488.2016.1179340

Jones, B. A., Haycraft, E., Murjan, S., & Arcelus, J. (2016). Body dissatisfaction and disordered eating in trans people: A systematic review of the literature. *International Review of Psychiatry, 28*, 81-94. http://doi.org/10.3109/09540261.2015.1089217

Jónsdóttir, S. R., Arnarson, E. Ö., & Smári, J. (2008). Body esteem, perceived competence and depression in Icelandic adolescents. *Nordic Psychology*, *60*, 58–71. http://doi.org/10.1027/1901-2276.60.1.58

Kapstad, H., Nelson, M., Øverås, M., & Rø, Ø. (2015). Validation of the Norwegian short version of the Body Shape Questionnaire (BSQ-14). *Nordic Journal of Psychiatry*, *69*, 509–514. http://doi.org/10.3109/08039488.2015.1009486

Kashubeck-West, S., Coker, A. D., Awad, G. H., Stinson, R. D., Bledman, R., & Mintz, L. (2013). Do measures commonly used in body image research perform adequately with African American college women? *Cultural Diversity and Ethnic Minority Psychology*, *19*, 357–368. http://doi.org/10.1037/a0031905

Kashubeck-West, S., Mintz, L. B., & Saunders, K. J. (2001). Assessment of eating disorders in women. *The Counseling Psychologist, 29*, 662-694. https://doi.org/10.1177/0011000001295003

Kelly, N. R., Mitchell, K. S., Gow, R. W., Trace, S. E., Lydecker, J. A., Bair, C. E., & Mazzeo, S. (2012). An evaluation of the reliability and construct validity of eating disorder measures in white and black women. *Psychological Assessment, 24*, 608–617. http://doi.org/10.1037/a0026457

Kertechian, S., & Swami, V. (2017). An examination of the factor structure and sex invariance of a French translation of the Body Appreciation Scale-2 in university students. *Body Image*, *21,* 26–29. http://doi.org/10.1016/j.bodyim.2017.02.005

Keum, B. T. H., Wong, S. N., DeBlaere, C., & Brewster, M. E. (2015). Body image and Asian American men: Examination of the Drive for Muscularity Scale. *Psychology of Men and Masculinity*, *16,* 284–293. http://doi.org/10.1037/a0038180

Kim, T-S., & Chee, I-S. (2018). The reliability and validity of the Korean version of the Body Shape Questionnaire. *Anxiety and Mood, 14*, 36-43. https://doi.org/10.24986/anxmod.2018.14.1.36

Lehmann, V., Ouwens, M. A., Braeken, J., Danner, U. N., van Elburg, A. A., Bekker, M. H. J., … van Strien, T. (2013). Psychometric properties of the Dutch version of the Eating Disorder Inventory–3. *SAGE Open*, *3*(4). http://doi.org/10.1177/2158244013508415

Lemoine, J. E., Konradsen, H., Lunde Jensen, A., Roland-Lévy, C., Ny, P., Khalaf, A., & Torres, S. (2018). Factor structure and psychometric properties of the Body Appreciation Scale-2 among adolescents and young adults in Danish, Portuguese, and Swedish. *Body Image*, *26,* 1–9. http://doi.org/10.1016/j.bodyim.2018.04.004

Lentillon-Kaestner, V., Berchtold, A., Rousseau, A., & Ferrand, C. (2014). Validity and reliability of the French versions of the Body Shape Questionnaire. *Journal of Personality Assessment*, *96,* 471–477. http://doi.org/10.1080/00223891.2013.843537

Lewis-Smith, H., Diedrichs, P. C., Rumsey, N., & Harcourt, D. (2018). Efficacy of psychosocial and physical activity-based interventions to improve body image among women treated for breast cancer: A systematic review. *Psycho-Oncology*, *27,* 2687-2699. http://dx.doi.org.ezproxy.ub.gu.se/10.1002/pon.4870

Lobera, I. J., & Ríos, P. B. (2011). Spanish version of the Body Appreciation Scale (BAS) for adolescents. *The Spanish Journal of Psychology, 14*, 411-420. https://doi.org/10.5209/rev_SJOP.2011.v14.n1.37

Luce, K. H., & Crowther, J. H. (1999). The reliability of the Eating Disorder Examination—Self-report Questionnaire version (EDE-Q). *International Journal of Eating Disorders, 25*, 349-351. https://doi.org/10.1002/(SICI)1098-108X(199904)25:3

Lydecker, J. A., White, M. A., & Grilo, C. M. (2016). Black patients with binge-eating disorder: Comparison of different assessment methods. *Psychological Assessment*, *28*, 1319–1324. http://doi.org/10.1037/pas0000246

Machado, P. P. P., Grilo, C. M., & Crosby, R. D. (2018). Replication of a modified factor structure for the Eating Disorder Examination-Questionnaire: Extension to clinical eating disorder and non-clinical samples in Portugal. *European Eating Disorders Review*, *26*, 75–80. http://doi.org/10.1002/erv.2569

Machado, P. P. P., Martins, C., Vaz, A. R., Conceição, E., Bastos, A. P., & Gonçalves, S. (2014). Eating Disorder Examination Questionnaire: Psychometric properties and norms for the Portuguese population. *European Eating Disorders Review, 22*, 448–453. http://doi.org/10.1002/erv.2318

Mahmoodi, M., Moloodi, R., Ghaderi, A., Babai, Z., Saleh, Z., Alasti, H., … Mohammadpour, Z. (2016). The Persian version of Eating Disorder Examination Questionnaire and Clinical Impairment Assessment: Norms and psychometric properties for undergraduate women. *Iranian Journal of Psychiatry*, *11*, 67–74. http://doi.org/10.1111/j.1755-5949.2008.00071.x

Marco, J. H., Perpiñá, C., Roncero, M., & Botella, C. (2017). Confirmatory factor analysis and psychometric properties of the Spanish version of the Multidimensional Body-Self

Relations Questionnaire-Appearance Scales in early adolescents. *Body Image, 21*, 15–18. http://doi.org/10.1016/j.bodyim.2017.01.003

Matarese, M., Lommi, M., & De Marinis, M. G. (2017). Systematic review of measurement properties of self-reported instruments for evaluating self-care in adults. *Journal of Advanced Nursing, 73*, 1272-1287. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1111/jan.13204

Mazzeo, S. E. (1999). Modification of an existing measure of body image preoccupation and its relationship to disordered eating in female college students. *Journal of Counseling Psychology, 46*, 42–50. http://doi.org/10.1037/0022-0167.46.1.42

McCreary, D. R., & Sasse, D. K. (2000). An exploration of the drive for muscularity in adolescent boys and girls. *Journal of the American College Health Association*, *48*, 297–304. http://doi.org/10.1080/07448480009596271

McCreary, D. R., Sasse, D. K., Saucier, D. M., & Dorsch, K. D. (2004). Measuring the drive for muscularity: Factorial validity of the Drive for Muscularity Scale in men and women. *Psychology of Men and Masculinity*, *5*, 49–58. http://doi.org/10.1037/1524-9220.5.1.49

McPherson, K. E., McCarthy, P., McCreary, D. R., & McMillan, S. (2010). Psychometric evaluation of the Drive for Muscularity Scale in a community-based sample of Scottish men participating in an organized sporting event. *Body Image, 7*, 368–371. http://doi.org/10.1016/j.bodyim.2010.06.001

Mendelson, B. K., Mendelson, M. J., & White, D. R. (2001). Body-Esteem Scale for Adolescents and Adults. *Journal of Personality Assessment, 76*, 90-106. https://doi.org/10.1207/S15327752JPA7601_6

Menzel, J. E., Krawczyk, R., & Thompson, J. K. (2011). Attitudinal assessment of body image for adolescents and adults. In T. F. Cash & L. Smolak (Eds.), *Body image: A* *handbook of science, practice, and prevention* (pp. 154-169). New York, NY: The Guilford Press

Mitsui, T., Yoshida, T., & Komaki, G. (2017). Psychometric properties of the Eating Disorder Examination-Questionnaire in Japanese adolescents. *BioPsychoSocial Medicine, 11*(1), 1–9. http://doi.org/10.1186/s13030-017-0094-8

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

Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., . . . De Vet, H. C. W. (2010). Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) Checklist. *BMC Medical Research Methodology, 10*(1), 82. https://doi.org/10.1186/1471-2288-10-82

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., . . . De Vet, H. C. W. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology, 10*, 22. https://doi.org/10.1186/1471-2288-10-22

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. W. (2010a). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 19*, 539-549. https://doi.org/10.1007/s11136-010-9606-8

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P.W., Knol, D. L., . . . de Vet, H. C. W. (2010b). 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*, 737-745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Mond, J. M., Hay, P. J., Rodgers, B., Owen, C., & Beumont, P. J. V. (2004a). Temporal stability of the Eating Disorder Examination Questionnaire. *International Journal of Eating Disorders*, *36*, 195–203. http://doi.org/10.1002/eat.20017

Mond, J. M., Hay, P. J., Rodgers, B., Owen, C., & Beumont, P. J. V. (2004b). Validity of the Eating Disorder Examination Questionnaire (EDE-Q) in screening for eating disorders in community samples. *Behaviour Research and Therapy*, *42*, 551–567. http://doi.org/10.1016/S0005-7967(03)00161-X

Moreira, G. S. X., Lorenzato, L., Neufeld, C. B., & Almeida, S. S. (2018). Brazilian version of the Body Appreciation Scale (BAS) for young adolescents. *The Spanish Journal of Psychology*, *21*(June), E21. http://doi.org/10.1017/sjp.2018.20

Moss, T. P., & Rosser, B. A. (2012). The moderated relationship of appearance valence on appearance self consciousness: Development and testing of new measures of appearance schema components. *PLoS ONE*, *7*(11). http://doi.org/10.1371/journal.pone.0050605

Moss, T. P., Lawson, V., White, P., Rumsey, N., Byron-Daniel, J., Charlton, R., … Walsh, E. (2014). Salience and valence of appearance in a population with a visible difference of appearance: Direct and moderated relationships with self-consciousness, anxiety and depression. *PLoS ONE*, *9*(2), 1–8. http://doi.org/10.1371/journal.pone.0088435

Mumford, D. B., Whitehouse, A. M., & Choudry, I. Y. (1992). Survey of eating disorders in English-medium schools in Lahore, Pakistan. *International Journal of Eating Disorders*, *11*, 173–184. http://doi.org/10.1002/1098-108X(199203)11:2<173: AID-EAT2260110208>3.0.CO;2-L

Mumford, D. B., Whitehouse, A. M., & Platts, M. (1991). Sociocultural correlates of eating disorders among Asian schoolgirls in Bradford. *British Journal of Psychiatry*, *158*, 222–228. http://doi.org/10.1192/bjp.158.2.222

Naqvi, I., & Kamal, A. (2017). Translation and validation of Multidimensional Body Self-Relations Questionnaire-appearance scale for young adults. *Pakistan Journal of Psychological Research, 32*(2), 465-485.

Nerini, A., Matera, C., Baroni, D., & Stefanile, C. (2016). Drive for muscularity and sexual orientation: Psychometric properties of the Italian version of the Drive for Muscularity Scale (DMS) in straight and gay men. *Psychology of Men and Masculinity*, *17*, 137–146. http://doi.org/10.1037/a0039675

Nevill, A. M., Lane, A. M., & Duncan, M. J. (2015). Are the Multidimensional Body Self-Relations Questionnaire Scales stable or transient? *Journal of Sports Sciences*, *33*, 1881–1889. http://doi.org/10.1080/02640414.2015.1018930

Ng, S., Barron, D., & Swami, V. (2015). Factor structure and psychometric properties of the Body Appreciation Scale among adults in Hong Kong. *Body Image, 13*, 1-8. http://doi.org/10.1016/j.bodyim.2014.10.009

Nyman-Carlsson, E., Engström, I., Norring, C., & Nevonen, L. (2015). Eating Disorder Inventory-3, validation in Swedish patients with eating disorders, psychiatric outpatients and a normal control sample. *Nordic Journal of Psychiatry*, *69*, 142–151. http://doi.org/10.3109/08039488.2014.949305

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews, 5*. doi:10.1186/s13643-016-0384-4

Parker, K., Mitchell, S., O'Brien, P., & Brennan, L. (2015). Psychometric evaluation of disordered eating measures in bariatric surgery patients. *Eating Behaviors*, *19*, 39–48. http://doi.org/https://doi.org/10.1016/j.eatbeh.2015.05.007

Parker, K., Mitchell, S., O'Brien, P., & Brennan, L. (2016). Psychometric evaluation of disordered eating measures inbariatric surgery candidates. *Obesity Surgery*, *26*, 563–575. http://doi.org/10.1007/s11695-015-1780-x

Peláez-Fernández, M. A., Labrador, F. J., & Raich, R. M. (2012). Validation of Eating Disorder Examination Questionnaire (EDE-Q) –Spanish Version– for screening eating disorders. *The Spanish Journal of Psychology*, *15*, 817–824. http://doi.org/10.5209/rev_SJOP.2012.v15.n2.38893

Penelo, E., Negrete, A., Portell, M., & Raich, R. M. (2013). Psychometric properties of the Eating Disorder Examination Questionnaire (EDE-Q) and norms for rural and urban adolescent males and females in Mexico. *PLoS ONE*, *8*, 1–11. http://doi.org/10.1371/journal.pone.0083245

Penelo, E., Villarroel, A. M., Portell, M., & Raich, R. M. (2012). Eating Disorder Examination Questionnaire (EDE-Q): An initial trial in Spanish male undergraduates. *European Journal of Psychological Assessment*, *28*, 76-83. http://doi.org/10.1027/1015-5759/a000093

Peterson, C. B., Crosby, R. D., Wonderlich, S. A., Joiner, T., Crow, S. J., Mitchell, J. E., . . . le Grange, D. (2007). Psychometric properties of the Eating Disorder Examination-Questionnaire: Factor structure and internal consistency. *International Journal of Eating Disorders*, *40*, 386-389. http://doi.org/10.1002/eat.20373

Phillips, K. E., Jennings, K. M., & Gregas, M. (2018). Factor structure of the Eating Disorder Examination-Questionnaire in a clinical sample of adult women with anorexia nervosa.

Journal of Psychosocial Nursing and Mental Health Services, *56*, 33–39. http://doi.org/10.3928/02793695-20180108-03

Pook, M., Tuschen-Caffier, B., & Brähler, E. (2008). Evaluation and comparison of different versions of the Body Shape Questionnaire. *Psychiatry Research*, *158*, 67–73. http://doi.org/10.1016/j.psychres.2006.08.002

Popkess-Vawter, S., & Banks, N. (1992). Body image measurement in overweight females. *Clinical Nursing Research, 1*, 402-417. http://doi.org/10.1177/105477389200100408

Pretorius, N., Waller, G., Gowers, S., & Schmidt, U. (2009). Validity of the Eating Disorder Examination-Questionnaire when used with adolescents with bulimia nervosa and atypical bulimia nervosa. *Eating and Weight Disorders*, *14*, 243–248. http://doi.org/10.1007/BF03325125

Probst, M., Pieters, G., & Vanderlinden, J. (2009). Body experience assessment in non-clinical male and female subjects. *Eating and Weight Disorders*, *14*, 1–6. http://doi.org/10.1007/BF03354623

Razmus, M., & Razmus, W. (2017). Evaluating the psychometric properties of the Polish version of the Body Appreciation Scale-2. *Body Image*, *23*, 45–49. http://doi.org/10.1016/j.bodyim.2017.07.004

Reas, D. L., Grilo, C. M., & Masheb, R. M. (2006). Reliability of the Eating Disorder Examination-Questionnaire in patients with binge eating disorder. *Behaviour Research and Therapy*, *44*, 43–51. http://doi.org/10.1016/j.brat.2005.01.004

Reas, D. L., Øverås, M., & Øyvind, R. (2012). Norms for the Eating Disorder Examination Questionnaire (EDE-Q) among high school and university men. *Eating Disorders, 20*, 437–443. http://doi.org/10.1080/10640266.2012.715523

Reilly, E. E., Anderson, L. M., Schaumberg, K., & Anderson, D. A. (2014). Gender-based differential item functioning in common measures of body dissatisfaction. *Body Image, 11*, 206–209. http://doi.org/10.1016/j.bodyim.2014.02.001

Rodgers, R. F., Franko, D. L., Lovering, M. E., Luk, S., Pernal, W., & Matsumoto, A. (2018). Development and validation of the Female Muscularity Scale. *Sex Roles, 78,* 18-26. http://dx.doi.org.ezproxy.ub.gu.se/10.1007/s11199-017-0775-6

Rø, Ø., Reas, D. L., & Lask, B. (2010). Norms for the Eating Disorder Examination Questionnaire among female university students in Norway. *Nordic Journal of Psychiatry, 64*, 428–432. http://doi.org/10.3109/08039481003797235

Roncero, M., Perpiñá, C., Marco, J. H., & Sánchez-Reales, S. (2015). Confirmatory factor analysis and psychometric properties of the Spanish version of the Multidimensional Body-Self Relations Questionnaire-Appearance Scales. *Body Image, 14*, 47-53. http://doi.org/10.1016/j.bodyim.2015.03.005

Rose, J. S., Vaewsorn, A., Rosselli-Navarra, F., Wilson, G. T., & Weissman, R. S. (2013). Test-retest reliability of the Eating Disorder Examination-Questionnaire (EDE-Q) in a college sample. *Journal of Eating Disorders, 1*(1). http://doi.org/10.1186/2050-2974-1-42

Rosen, J. C., Jones, A., Ramirez, E., & Waxman, S. (1996). Body Shape Questionnaire: Studies of validity and reliability. *International Journal of Eating Disorders, 20,* 315-319. http://doi.org/10.1002/(SICI)1098-108X(199611)20:3

Rothstein, L. A., Sbrocco, T., & Carter, M. M. (2017). Factor analysis of EDI-3 eating disorder risk subscales among African American women. *Journal of Black Psychology, 43*, 767–777. http://doi.org/10.1177/0095798417708506

Rumsey, N., & Harcourt, D. (2012). *Oxford handbook of the psychology of appearance*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199580521.001.0001

Rusticus, S. A., & Hubley, A. M. (2006). Measurement invariance of the Multidimensional Body-Self Relations Questionnaire: Can we compare across age and gender? *Sex Roles, 55*, 827–842. http://doi.org/10.1007/s11199-006-9135-7

Sabiston, C. M., Rusticus, S., Brunet, J., McDonough, M. H., Hadd, V., Hubley, A. M., & Crocker, P. R. E. (2010). Invariance test of the Multidimensional Body Self-Relations Questionnaire: Do women with breast cancer interpret this measure differently? *Quality of Life Research*, *19,* 1171–1180. http://doi.org/10.1007/s11136-010-9680-y

Sepulveda, A. R., Parks, M., de Pellegrin, Y., Anastasiadou, D., & Blanco, M. (2016). Validation of the Spanish version of the Drive for Muscularity Scale (DMS) among males: Confirmatory factor analysis. *Eating Behaviors*, *21,* 116–122. http://doi.org/10.1016/j.eatbeh.2016.01.010

Silva, W. R., Costa, D., Pimenta, F., Maroco, J., & Campos, J. A. D. B. (2016). Psychometric evaluation of a unified Portuguese-language version of the Body Shape Questionnaire in female university students. *Cadernos de Saúde Pública*, *32*, 1–12. http://doi.org/10.1590/0102-311X00133715

Skrzypek, S., Wehmeier, P. M., & Remschmidt, H. (2001). Body image assessment using body size estimation in recent studies on anorexia nervosa. A brief review. *European Child & Adolescent Psychiatry*, *10,* 215-221. http://doi.org/10.1007/s007870170010

Smith, A. R., & Davenport, B. R. (2012). An evaluation of body image assessments in Hispanic college women: The Multidimensional Body-Self Relations Questionnaire and the Appearance Schemas Inventory-Revised. *Journal of College Counseling*, *15*, 198-214. https://doi.org/10.1002/j.2161-1882.2012.00016.x

Speyer, R., Kim, J-H., Doma, K., Chen, Y-W., Denman, D., Phyland, D., ... Cordier, R. (2019). Measurement properties of self-report questionnaires on health-related quality of life and functional health status in dysphonia: A systematic review using the COSMIN taxonomy. *Quality of Life Research, 28,* 283–296. https://doi.org/10.1007/s11136-018-2001-6

Stein, K. F., Riley, B. B., Hoyland-Domenico, L., & Lee, C. K. (2015). Measurement of body dissatisfaction in college-enrolled Mexican American Women: A Rasch-based examination of the validity and reliability of the EDI-III. *Eating Behaviors, 19,* 5–8. http://doi.org/10.1016/j.eatbeh.2015.06.001

Stock, N. M., Billaud Feragen, K., & Rumsey, N. (2018). Adults' narratives of growing up with a cleft lip and/or palate: Factors associated with psychological adjustment. *The Cleft Palate-Craniofacial Journal, 53*, 222–239. https://doi.org/10.1597/14-269

Swami, V., & Chamorro-Premuzic, T. (2008). Factor structure of the Body Appreciation Scale among Malaysian women. *Body Image*, *5,* 409–413. http://doi.org/10.1016/j.bodyim.2008.04.005

Swami, V., & Jaafar, J. L. (2012). Factor structure of the Body Appreciation Scale among Indonesian women and men: Further evidence of a two-factor solution in a non-Western population. *Body Image*, *9*, 539–542. http://doi.org/10.1016/j.bodyim.2012.06.002

Swami, V., Ng, S. K., & Barron, D. (2016a). Translation and psychometric evaluation of a Standard Chinese version of the Body Appreciation Scale-2. *Body Image*, *18,* 23–26. http://doi.org/10.1016/j.bodyim.2016.04.005

Swami, V., Barron, D., Lau, P. L., & Jaafar, J. L. (2016b). Psychometric properties of the Drive for Muscularity Scale in Malay men. *Body Image*, *17*, 111–116. http://doi.org/10.1016/j.bodyim.2016.03.004

Swami, V., Özgen, L., Gökçen, E., & Petrides, K. V. (2015). Body image among female university students in Turkey: Concurrent translation and validation of three body image measures. *International Journal of Culture and Mental Health, 8,* 176–191. http://doi.org/10.1080/17542863.2014.917117

Swami, V., Stieger, S., Haubner, T., & Voracek, M. (2008). German translation and psychometric evaluation of the Body Appreciation Scale. *Body Image*, *5,* 122–127. http://doi.org/10.1016/j.bodyim.2007.10.002

Swami, V., Tudorel, O., Goian, C., Barron, D., & Vintila, M. (2017). Factor structure and psychometric properties of a Romanian translation of the Body Appreciation Scale-2. *Body Image*, *23,* 61–68. http://doi.org/10.1016/j.bodyim.2017.08.001

Swami, V., Vintila, M., Tudorel, O., Goian, C., & Barron, D. (2018). Factor structure and psychometric properties of a Romanian translation of the drive for Muscularity Scale (DMS) in university men. *Body Image, 25,* 48–55. http://doi.org/10.1016/j.bodyim.2018.02.004

Taylor, D., Szpakowska, I., & Swami, V. (2013). Weight discrepancy and body appreciation among women in Poland and Britain. *Body Image*, *10*, 628–631. http://doi.org/10.1016/j.bodyim.2013.07.008

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & de Vet, Henrica 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: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *21*, 651-657. https://doi.org/10.1007/s11136-011-9960-1

Thoma, A., Sprague, S., Veltri, K., Duku, E., & Furlong, W. (2005). Methodology and measurement properties of health-related quality of life instruments: A prospective study of patients undergoing breast reduction surgery. *Health and Quality of Life Outcomes, 3*(1), 44. http://doi.org/10.1186/1477-7525-3-44

Thompson, J. K., Burke, N. L., & Krawczyk, R. (2012). Mesurement of body image in adolescence and adulthood. In T. F. Cash (Ed.), *Encyclopedia of body image and human appearance* (Vol. 2, pp. 512-520). San Diego, CA: Academic Press/Elsevier.

Thompson, J. K., Penner, L. A., & Altabe, M. N. (1990). Procedures, problems, and progress in the assessment of body images. In T. F. Cash & T. Pruzinsky (Eds.), *Body images. Development, deviance and change* (pp. 21-50). New York, NY: The Guilford Press.

Tiggemann, M. (2001). Person × situation interactions in body dissatisfaction. *International Journal of Eating Disorder*s, *29*, 65-70. doi: 10.1002/1098-108X(200101)29:1<65:AID-EAT10>3.0.CO;2-Y

Tod, D., Morrison, T. G., & Edwards, C. (2012). Evaluating validity and test-retest reliability in four drive for muscularity questionnaires. *Body Image*, *9*, 425–428. http://doi.org/10.1016/j.bodyim.2012.02.001

Túry, F., Güleç, H., & Kohls, E. (2010). Assessment methods for eating disorders and body image disorders. *Journal of Psychosomatic Research, 69*, 601-611. https://doi.org/10.1016/j.jpsychores.2009.05.012

Tylka, T. L. (2013). Evidence for the Body Appreciation Scale’s measurement equivalence/invariance between U.S. college women and men. *Body Image*, *10*, 415–418. http://doi.org/10.1016/j.bodyim.2013.02.006

Tylka, T. L. (2018). Body image: Celebrating the past, appreciating the present, and envisioning the future. *Body Image, 24,* A1-A3. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1016/j.bodyim.2018.01.003

Tylka, T. L., & Wood-Barcalow, N. L. (2015a). What is and what is not positive body image? conceptual foundations and construct definition. *Body Image*, *14*, 118-129. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1016/j.bodyim.2015.04.001

Tylka, T. L., & Wood-Barcalow, N. L. (2015b). The Body Appreciation Scale-2: Item refinement and psychometric evaluation. *Body Image*, *12*, 53–67. http://doi.org/10.1016/j.bodyim.2014.09.006

Unikel Santoncini, C., Bojorquez Chapela, I., Díaz de León Vázquez, C., Vázquez Velázquez, V., Rivera Márquez, J. A., Galván Sánchez, G., & Rocha Velis, I. (2018). Validation of Eating Disorders Examination Questionnaire in Mexican women. *International Journal of Eating Disorders*, *51*, 146–154. http://doi.org/10.1002/eat.22819

Untas, A., Koleck, M., Rascle, N., & Borteyrou, X. (2009). Psychometric properties of the French adaptation of the Multidimensional Body Self Relations Questionnaire–Appearance Scales. *Psychological Reports*, *105*, 461–471. http://doi.org/10.2466/PR0.105.2.461-471

Villarroel, A. M., Penelo, E., Portell, M., & Raich, R. M. (2011). Screening for eating disorders in undergraduate women: Norms and validity of the Spanish version of the Eating Disorder Examination Questionnaire (EDE-Q). *Journal of Psychopathology and Behavioral Assessment*, *33*, 121–128. http://doi.org/10.1007/s10862-009-9177-6

Vossbeck-Elsebusch, A. N., Waldorf, M., Legenbauer, T., Bauer, A., Cordes, M., & Vocks, S. (2014). German version of the Multidimensional Body-Self Relations Questionnaire - Appearance Scales (MBSRQ-AS): Confirmatory factor analysis and validation. *Body Image*, *11*, 191–200. http://doi.org/10.1016/j.bodyim.2014.02.002

Warren, C. S., Cepeda-Benito, A., Gleaves, D. H., Moreno, S., Rodriguez, S., Fernandez, M. C., … Pearson, C. A. (2008). English and Spanish versions of the Body Shape

Questionnaire: Measurement equivalence across ethnicity and clinical status. *International Journal of Eating Disorders*, *41*, 265–272. http://doi.org/10.1002/eat.20492

Webb, J. B., Wood-Barcalow, N. L., & Tylka, T. L. (2015). Assessing positive body image: Contemporary approaches and future directions. *Body Image, 14,* 130-145. https://doi.org/10.1016/j.bodyim.2015.03.010

Welch, E., Lagerström, M., & Ghaderi, A. (2012). Body Shape Questionnaire: Psychometric properties of the short version (BSQ-8C) and norms from the general Swedish population. *Body Image*, *9*, 547–550. http://doi.org/10.1016/j.bodyim.2012.04.009

Weldam, S. W. M., Schuurmans, M. J., Liu, R., & Lammers, J. J. (2013). Evaluation of quality of life instruments for use in COPD care and research: A systematic review. *International Journal of Nursing Studies*, *50*, 688-707. http://dx.doi.org.ezproxy.ub.gu.se/10.1016/j.ijnurstu.2012.07.017

White, H. J., Haycraft, E., Goodwin, H., & Meyer, C. (2014). Eating Disorder Examination Questionnaire: Factor structure for adolescent girls and boys. *International Journal of Eating Disorders*, *47*, 99–104. http://doi.org/10.1002/eat.22199

Wojtowicz, A. E., & Von Ranson, K. M. (2006). Psychometric evaluation of two scales examining muscularity concerns in men and women. *Psychology of Men and Masculinity*, *7*, 56–66. http://doi.org/10.1037/1524-9220.7.1.56

Yucel, B., Polat, A., Ikiz, T., Dusgor, B. P., Yavuz, A. E., & Berk, O. S. (2011). The Turkish version of the Eating Disorder Examination Questionnaire: Reliability and validity in adolescents. *European Eating Disorders Review*, *19*, 509–511. http://doi.org/10.1002/erv.1104

1

Figure 1.

Details of studies obtained and excluded in search Step 2, following recommendations by Moher, Liberati, Tetzlaff, and Altman (2009).

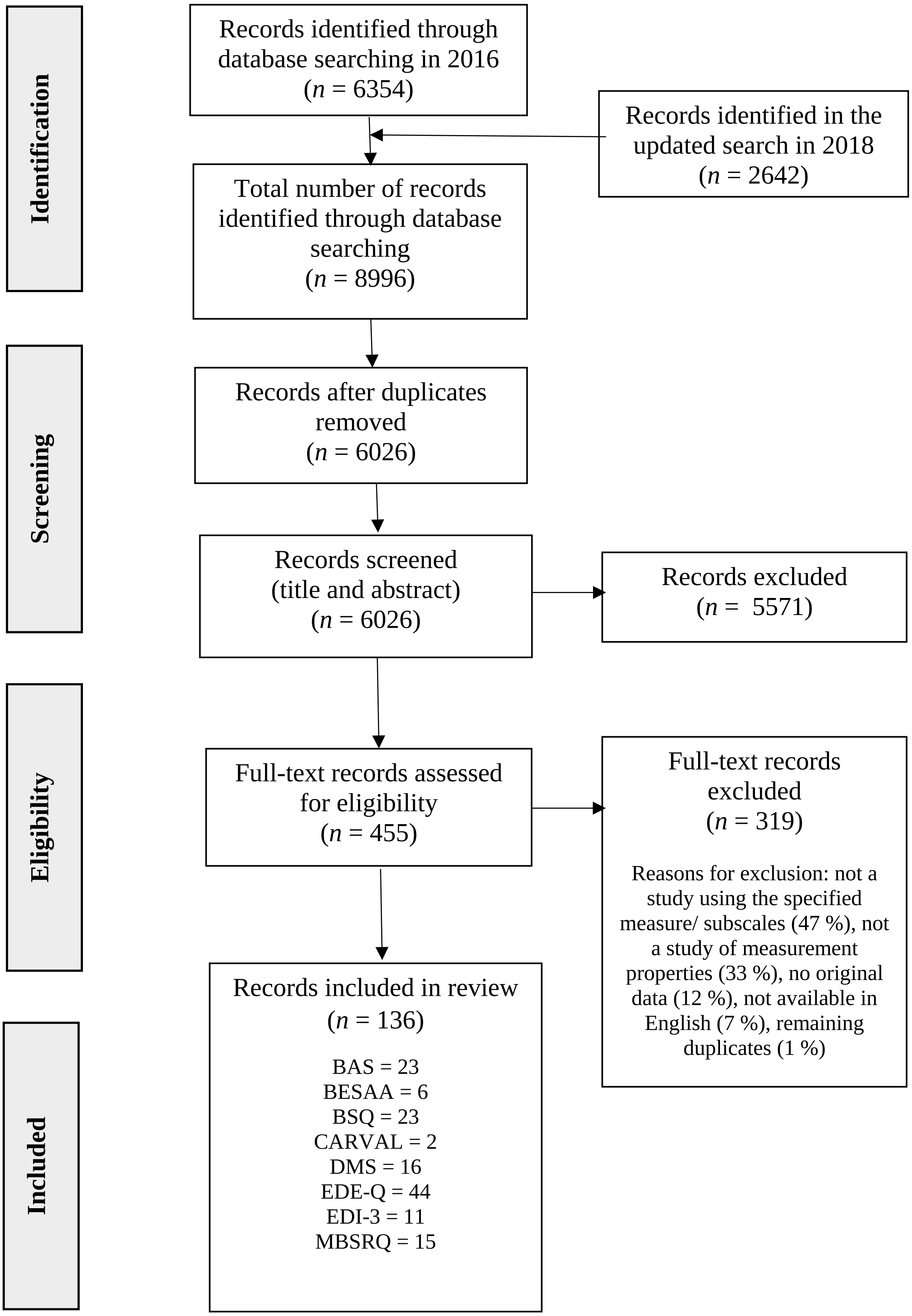
Identification
Screening
Eligibility
Included
Records identified through database searching in 2016 (n = 6354)
Records identified in the updated search in 2018 (n = 2642)
Total number of records identified through database searching (n = 8996)
Records after duplicates removed (n = 6026)
Records screened (title and abstract) (n = 6026)
Records excluded (n = 5571)
Full-text records assessed for eligibility (n = 455)
Full-text records excluded (n = 319)
Reasons for exclusion: not a study using the specified measure/ subscales (47 %), not a study of measurement properties (33 %), no original data (12 %), not available in English (7 %), remaining duplicates (1 %)
Records included in review (n = 136)
BAS = 23
BESAA = 6
BSQ = 23
CARVAL = 2
DMS = 16
EDE-Q = 44
EDI-3 = 11
MBSRQ = 15

Table 1.
Measures included in the present systematic review.

Measure	Abbreviation(s)	Included subscales	Authors and year of publication	Concept(s) measured	Description	Example item(s)
Body Appreciation Scale	BAS; BAS-2		Avalos, Tylka, & Wood-Barcalow, 2005; Tylka & Wood-Barcalow, 2015b	Body appreciation; positive body image	The 13-item BAS assesses individuals' acceptance of, favorable opinions toward, and respect for their bodies. The revised measure, BAS-2, consists of 10 items, five of which were retained from the parent scale. Items are rated on a 5-point scale (1= *never*, 5= *always*).	I feel good about my body

Body Esteem Scale for Adolescents and Adults	BESAA	Appearance esteem (AE);
		Weight esteem (WE);
		Attribution (A)
			Mendelson, Mendelson, & White, 2001	Body esteem, body image	Appearance subscale (10 items) assesses satisfaction with appearance, Weight subscale (8 items) captures satisfaction with weight, and Attribution subscale (5 items) assesses how one believes other people think about one's appearance. Items are rated from 0 (never) to 4 (always).	AE: I like what I see when I look in the mirror
						WE: I really like what I weigh
						A: People my own age like my looks

Measure	Abbrevia-tion(s)	Included subscales	Authors and year of publication	Concept(s) measured	Description	Example item(s)
Body Shape Questionnaire	BSQ		Cooper, Taylor, Cooper, & Fairburn, 1987	Body dissatisfaction, body shape preoccupations	The original BSQ includes 34 items on body shape, in particular the experience of “feeling fat.” The items refer to the past four weeks and are answered on a 6-point scale, from *never* to *always*.	Have you felt so bad about your shape that you have cried?
Centre for Appearance Research Valence Scale	CARVAL		Moss & Rosser, 2012	Appearance valence, body dissatisfaction	The CARVAL contains 8 items measuring the extent to which the respondent evaluates her/his appearance in a positive/negative way. Response categories ranging from 1 (*strongly disagree*) to 6 (*strongly agree*).	I feel bad about my body and my appearance

Measure	Abbrevia-tion(s)	Included subscales	Authors and year of publication	Concept(s) measured	Description	Example item(s)
Drive for Muscularity Scale	DMS		McCreary & Sasse, 2000	Muscularity-related attitudes and behaviors, muscularity dissatisfaction	The DMS is a 15-item scale that measures desire for a more muscular body. Participants indicate how each item reflects their own behaviors and attitudes using a 6-point scale from *Always* (scored as 6) to *Never* (1).	I wish that I were more muscular
Eating Disorders Examination Questionnaire	EDE-Q	Weight Concern; Shape Concern	Fairburn & Beglin, 1994	Weight and shape concerns, Body dissatisfaction	The EDE-Q assesses key attitudes and behavioral features of eating disorders over the past 28-days using a 7-point scale (*No days* [0] to *Every day* [6]). Body image is assessed by the subscales Weight concern (5 items) and Shape concern (8 items)	WC: How dissatisfied have you been with your weight? SC: Has your shape influenced how you think about (judge) yourself as a person?

Measure	Abbrevia-tion(s)	Included subscales	Authors and year of publication	Concept(s) measured	Description	Example item(s)
Eating Disorders Inventory-3	EDI-3	Body Dissatisfaction	Garner, 2004	Body dissatisfaction	The EDI-3 assesses eating disorder symptomology. Body image is assessed with the subscale Body dissatisfaction (10 items). Responses are answered on a 6-point scale, from *always* to *never*.	I think that my stomach is too big.
Multidimensional Body Relations Questionnaire	MBSRQ	Appearance Evaluation (AE); Body Areas Satisfaction Scale (BASS)	Brown, Cash, & Mikulka, 1990	Body image, appearance evaluation, body satisfaction/dissatis faction	The MBSRQ assesses self-attitudinal body image. The Appearance evaluation subscale includes 7 items measuring appearance satisfaction/dissatisfaction on a 5-point scale from *Definitely disagree* (1) to *Definitely agree* (5). The Body areas satisfaction scale includes 9 items measuring satisfaction/dissatisfaction with body areas on a 5-point scale from *Very dissatisfied* (1) to *Very satisfied* (5).	AE: I like my looks just the way they are BASS: Face (facial features, complexion); Muscle tone

Measure	Abbreviation(s)	Included subscales	Authors and year of publication	Concept(s) measured	Description	Example item(s)
Body Appreciation Scale	BAS; BAS-2		Avalos, Tylka, & Wood-Barcalow, 2005; Tylka & Wood-Barcalow, 2015b	Body appreciation; positive body image	The 13-item BAS assesses individuals' acceptance of, favorable opinions toward, and respect for their bodies. The revised measure, BAS-2, consists of 10 items, five of which were retained from the parent scale. Items are rated on a 5-point scale (1= never, 5= *always*).	I feel good about my body

Table 2.
Levels of evidence for the quality of the measurement properties (Furlan et al., 2009).

Level	Rating	Criteria
Strong	+ + + or - - -	Consistent findings (positive or negative) in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	+ + or - -	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or -	One study of fair methodological quality
Conflicting	+ / -	Conflicting findings
Unknown	?	Only studies of poor methodological quality OR the results are indeterminate for other reasons

+ = positive rating, ? = indeterminate rating, - = negative rating

Table 3.
Overall evidence rating

	BAS (BAS-2)	BESAA	BSQ	CARVAL	DMS	EDE-Q (SC & WC)	EDI-3 (BD)	MBSRQ (AE & BASS)
Internal consistency	++ (+++)	+++	+++	+++	+++	+++	++	+++
Reliability	++ (++)	++	++	--	++	++	+	+/-
Content validity	+++ (?)	?	+++	?	--	?	+++	
Structural validity	+/- (+++)	+++	+++	+++	+++	---	+/-	+/-
Hypotheses testing	++ (+++)	++	++	+++	+++	+++	++	++
Criterion validity	(-)		?	?	?	++	++	?
Responsiveness			+					+

Appendix 1

Methodological quality by sample and measurement property for the original and revised Body Appreciation Scale (BAS and BAS-2; Avalos, Tylka, & Wood-Barcalow, 2005; Tylka & Wood-Barcalow, 2015b). Results for BAS-2 are followed by (2).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsive-ness
Alcaraz-Ibañes et al., 2017, sample a	Fair (2)			Fair (2)	Fair (2)	Poor (2)		
Alcaraz-Ibañes et al., 2017, sample b	Fair (2)			Fair (2)	Fair (2)			
Alcaraz-Ibañes et al., 2017, sample c		Fair (2)						
Alexias et al., 2016	Fair	Fair		Fair	Fair	Fair		
Alleva et al., 2016	Excellent (2)			Good (2)	Good (2)	Poor (2)		
Atari et al., 2016, sample a	Fair (2)			Fair (2)	Fair (2)	Poor (2)		
Atari et al., 2016, sample b	Fair (2)			Fair (2)	Fair (2)			
Avalos et al., 2005, sample a	Fair		Fair	Fair	Fair			
Avalos et al., 2005, sample b				Fair				

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsive-ness
Avalos et al., 2005, sample c					Fair			
Avalos et al., 2005, sample d	Poor	Poor						
Cotter et al., 2015	Fair			Fair	Fair			
Ferreira et al., 2014	Fair		Excellent	Good	Fair	Fair		
Jauregui et al., 2011, sample a	Fair			Fair	Poor	Poor		
Jauregui et al., 2011, sample b	Fair	Fair						
Kertechian & Swami, 2017, sample a	Fair (2)			Fair (2)	Fair (2)	Poor (2)		
Kertechian & Swami, 2017, sample b	Fair (2)			Fair(2)	Fair (2)			
Kertechian & Swami, 2017, sample c	Fair (2)			Fair (2)	Fair (2)			
Lemoine et al., 2018, sample a	Good (2)			Good (2)	Good (2)	Fair (2)		

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsive-ness
Lemoine et al., 2018, sample b	Good (2)			Fair (2)	Good (2)			
Lemoine et al., 2018, sample c	Good (2)			Excellent (2)	Excellent (2)			
Lemoine et al., 2018, sample d	Excellent (2)			Excellent (2)	Excellent (2)			
Lemoine et al., 2018, sample e	Excellent (2)			Excellent (2)	Excellent (2)			
Lemoine et al., 2018, sample f	Excellent (2)			Excellent (2)	Excellent (2)			
Moreira et al., 2018	Fair		Fair	Fair	Fair			
Ng et al., 2015	Fair			Excellent	Good			
Razmus & Razmus, 2017, sample a	Excellent (2)			Excellent (2)		Fair (2)		
Razmus & Razmus, 2017, sample b	Excellent (2)			Excellent (2)				
Razmus & Razmus, 2017, sample c	Excellent (2)			Excellent (2)	Good (2)			

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsive-ness
Swami & Chamorro-Premuzic, 2008	Fair			Fair	Fair	Poor		
Swami & Jaafar, 2012, sample a	Fair			Fair	Fair	Poor		
Swami & Jaafar, 2012, sample b	Fair			Fair	Fair			
Swami & Ng 2015, sample a	Fair (2)			Fair (2)	Fair (2)	Poor (2)		
Swami & Ng, 2015, sample b	Fair (2)			Fair (2)	Fair (2)			
Swami et al., 2008, sample a	Fair			Fair	Poor	Poor		
Swami et al., 2008, sample b	Fair			Fair	Poor			
Swami et al., 2015	Fair			Fair	Fair	Poor		
Swami et al., 2016a, sample a	Fair (2)			Fair (2)	Fair (2)	Poor (2)		
Swami et al., 2016a, sample b	Fair (2)			Fair (2)	Fair (2)			
Swami et al., 2016a, sample c	Fair (2)			Fair (2)				

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsive-ness
Swami et al., 2017, sample a	Excellent (2)			Excellent (2)	Good (2)	Poor (2)		
Swami et al., 2017, sample b	Excellent (2)			Excellent (2)	Good (2)			
Swami et al., 2017, sample c	Excellent (2)			Fair (2)	Good (2)			
Swami et al., 2017, sample d		Fair (2)						
Taylor et al., 2013	Fair			Fair	Fair	Fair		
Tylka, 2013	Good			Good	Fair			
Tylka & Wood-Barcalow, 2015b, sample a	Excellent (2)	Fair (2)	Poor (2)	Excellent (2)	Excellent (2)		Fair (2)	
Tylka & Wood-Barcalow, 2015b, sample b	Excellent (2)			Excellent (2)	Good (2)			
Tylka & Wood-Barcalow, 2015b, sample c	Excellent (2)			Excellent (2)	Good (2)			
Tylka & Wood-Barcalow, 2015b, sample d	Excellent (2)			Excellent (2)	Poor (2)			

Appendix 2

Methodological quality by sample and measurement property for The Body Esteem Scale for Adolescents and Adults (BESAA; Mendelson, Mendelson & White, 2001).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Confalonieri et al., 2008	Good			Fair	Fair	Poor		
Cragun et al., 2013, sample a	Good		Poor	Good	Poor			
Cragun et al., 2013, sample b	Good			Good	Poor			
Franko et al., 2012	Fair	Fair		Fair	Fair			
Gallini, 2008	Fair		Poor	Excellent	Fair			
Jónsdóttir et al., 2008	Fair				Fair			
Mendelson et al., 2001	Fair	Fair		Fair	Fair			

Appendix 3

Methodological quality by sample and measurement property for the Body Shape Questionnaire (BSQ; Cooper, Taylor, Cooper, & Fairburn, 1987).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Akdemir et al., 2012	Excellent	Fair		Fair	Fair	Poor		
Conti et al., 2009	Poor	Fair			Fair		Poor	
Cooper et al., 1987, sample a					Poor			
Cooper et al., 1987, sample b					Poor			
Di Pietro et al., 2009	Poor				Poor	Poor		
Dowson & Henderson, 2001	Poor				Fair			
Evans & Dolan, 1993	Excellent			Good	Poor			
Franko et al., 2012	Fair	Fair		Fair	Fair			
Ghaderi & Scott, 2004, sample a	Excellent			Excellent	Fair			
Ghaderi & Scott, 2004, sample b	Poor	Fair			Fair			

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Ghaderi & Scott, 2004, sample c	Poor				Poor			
Kapstad et al., 2015, sample a	Fair	Poor			Poor		Poor	
Kapstad et al., 2015, sample b	Poor				Fair			
Kim & Chee, 2018	Excellent	Fair		Excellent	Fair	Poor		
Lentillon-Kaestner et al., 2014, sample a	Poor			Poor	Fair			
Lentillon-Kaestner et al., 2014, sample b	Fair	Fair		Fair	Fair			
Mazzeo, 1999, sample a		Poor		Fair			Poor	
Mazzeo, 1999, sample b	Fair			Fair			Poor	
Mumford et al., 1991				Fair	Poor			
Mumford et al., 1992				Good	Good			
Pook et al., 2008, sample a	Fair			Fair				

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Pook et al., 2008 sample b								Fair
Popkess-Vawter et al., 1992	Poor	Fair					Fair	
Probst et al., 2009	Fair				Fair			
Reilly et al., 2014	Fair							
Rosen et al., 1996, sample a					Poor		Fair	
Rosen et al., 1996, sample b					Poor		Poor	
Rosen et al., 1996, sample c		Fair			Poor		Poor	
Rosen et al., 1996, sample d					Poor		Poor	
Silva et al., 2014	Excellent		Excellent	Excellent	Fair			
Silva et al., 2016, sample a	Fair		Excellent	Fair	Fair	Poor		
Silva et al., 2016, sample b	Fair			Fair	Fair			
Warren et al., 2008, sample a	Fair			Fair				

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Warren et al., 2008, sample b	Fair			Fair				
Warren et al., 2008, sample c	Fair			Fair				
Warren et al., 2008, sample d	Fair			Fair				
Welch et al., 2012, sample a	Fair	Fair				Poor		
Welch et al., 2012, sample b	Fair			Fair	Fair			

Appendix 4

Methodological quality by sample and measurement property for the Centre for Appearance Research Valence Scale (CARVAL; Moss & Rosser, 2012).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Moss & Rosser, 2012, sample a	Good		Poor		Good		Fair	
Moss & Rosser, 2012 sample b		Fair						
Moss et al., 2014	Good	Fair		Good	Good			

Appendix 5

Methodological quality by sample and measurement property for the Drive for Muscularity Scale (DMS; McCreary & Sasse, 2000).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Cafri & Thompson, 2004, sample a	Fair	Fair			Fair		Fair	
Cafri & Thompson, 2004, sample b	Fair				Fair			
Campana et al., 2013	Good		Good	Good	Poor	Good		
Chaba et al., 2018, sample a	Fair			Fair		Good		
Chaba et al., 2018, sample b	Fair	Fair		Fair	Fair			
Compte et al., 2015	Good			Excellent	Good			
DeBlaere et al., 2017	Good			Good	Good			

Escoto et al., 2013	Fair		Fair		Fair
Keum et al., 2015	Good		Good		
McCreary & Sasse, 2000	Poor			Fair	
McCreary et al., 2004, sample a	Fair		Fair		
McCreary et al., 2004, sample b	Fair		Fair		
McPherson et al., 2010	Fair	Fair	Fair	Fair	
Nerini et al., 2016, sample a	Excellent		Excellent	Fair	Poor
Nerini et al., 2016, sample b	Excellent		Excellent	Fair	Poor
Sepulveda et al., 2016	Good		Good	Good	Poor
Swami et al., 2016b	Fair		Fair	Fair	Poor

Swami et al., 2018	Excellent			Excellent	Good	Poor
Tod et al., 2012, sample a	Poor				Fair	
Tod et al., 2012, sample b	Poor	Fair				
Wojtowicz & von Ranson, 2006, sample a	Poor				Fair	
Wojtowicz & von Ranson, 2006, sample b	Poor				Fair	

Appendix 6

Methodological quality by sample and measurement property for the Weight and Shape concerns subscales of the Eating Disorders Examination Questionnaire (EDE-Q; Fairburn & Beglin, 1994).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Allen et al., 2011, sample a	Fair			Fair			Fair	
Allen et al., 2011,sample b	Fair			Fair			Fair	
Bardone-Cone & Boyd, 2007, sample a	Poor	Fair			Fair			
Bardone-Cone & Boyd, 2007, sample b	Poor	Fair			Fair			
Barnes et al., 2012	Good			Good				
Becker et al., 2010	Excellent	Poor		Excellent	Fair	Poor		
Binford et al., 2005							Fair	
Calugi et al., 2016	Fair	Poor		Good	Fair	Poor		
Carrard et al., 2015, sample a	Fair			Fair		Poor		
Carrard et al., 2015, sample b	Fair			Fair		Poor		
Chan & Leung, 2015	Poor			Fair	Fair			

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Darcy et al., 2013, sample a				Good				
Darcy et al., 2013, sample b				Good				
Darcy et al., 2013, sample c				Good				
Darcy et al., 2013, sample d				Good				
Elder & Grilo, 2007		Fair				Poor	Fair	
Franko et al., 2012	Fair	Fair		Fair	Fair			
Gideon et al., 2016	Excellent		Poor					
Giovazolias et al., 2013, sample a	Fair			Fair		Fair		
Giovazolias et al., 2013, sample b					Fair			
Grilo et al., 2013	Good			Good	Fair			
Grilo et al., 2015	Excellent			Excellent	Fair			
Heiss et al., 2018, sample a	Excellent			Fair				
Heiss et al., 2018, sample b	Excellent			Excellent				

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Hilbert et al., 2012, sample a	Excellent				Good	Poor		
Hilbert et al., 2012, sample b	Fair				Fair	Poor		
Hrabosky et al., 2008	Fair			Fair	Fair			
Isomaa et al., 2016, sample a	Poor				Fair	Fair		
Isomaa et al., 2016, sample b	Poor				Fair	Fair		
Isomaa et al., 2016, sample c	Poor				Fair	Fair		
Luce & Crowther, 1999	Poor	Fair						
Lydecker et al., 2016, sample a	Poor						Fair	
Lydecker et al., 2016, sample b	Poor						Fair	
Machado et al., 2014, sample a	Poor					Fair		
Machado et al., 2014, sample b							Fair	

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Machado et al., 2018, sample a	Excellent			Excellent				
Machado et al., 2018, sample b	Excellent			Excellent				
Mahmoodi et al., 2016	Poor				Fair	Fair		
Mitsui et al., 2017, sample a	Good			Good		Poor		
Mitsui et al., 2017, sample b					Fair			
Mitsui et al., 2017, sample c					Fair			
Mitsui et al., 2017, sample d					Fair			
Mond et al., 2004a	Poor	Fair						
Mond et al., 2004b							Fair	
Parker et al., 2015	Poor			Poor	Fair			
Parker et al., 2016	Good			Good	Fair		Good	
Peláez-Fernández et al., 2012	Poor				Fair		Fair	
Penelo et al., 2012	Excellent			Excellent	Excellent			

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Penelo et al., 2013	Fair	Fair		Fair	Poor			
Peterson et al., 2007	Fair			Fair				
Phillips et al., 2018	Fair			Fair				
Pretorius et al., 2009							Fair	
Reas et al., 2006		Poor						
Reas et al., 2012	Poor				Poor			
Reilly et al., 2014	Fair				Fair			
Rø et al., 2010	Fair	Poor				Poor		
Rose et al., 2013	Fair	Poor						
Unikel Santoncini et al., 2018, sample a	Fair			Fair	Fair	Fair		
Unikel Santoncini et al., 2018, sample b	Fair			Fair	Fair	Fair		
Villarroel et al., 2011	Excellent			Excellent	Fair			
White et al., 2014, sample a				Excellent	Fair			
White et al., 2014, sample b				Excellent	Fair			
Yucel et al., 2011	Poor	Fair			Fair	Poor		

Appendix 7

Methodological quality by sample and measurement property for the Body Dissatisfaction subscale of the Eating Disorder Inventory 3 (EDI-3; Garner, 2004).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Belon et al., 2015	Poor			Poor				
Clausen et al., 2011, sample a	Fair			Fair	Poor	Poor	Poor	
Clausen et al., 2011, sample b	Fair			Fair	Poor	Poor		
Cordero et al., 2013	Fair			Fair				
Dadgostar et al., 2017	Poor	Poor	Excellent			Fair		
Elosua & Hermosilla, 2013, sample a	Good			Good				
Elosua & Hermosilla, 2013, sample b	Good			Good				
Elosua & López-Jáuregui, 2012	Poor	Fair		Poor		Poor		
Kashubeck-West et al., 2013	Good			Excellent	Fair			

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Lehmann et al., 2013	Fair			Fair			Fair	
Nyman-Carlsson et al., 2015, sample a	Fair				Fair		Fair	
Nyman-Carlsson et al., 2015, sample b	Fair				Poor			
Nyman-Carlsson et al., 2015, sample c	Fair				Poor			
Rothstein et al., 2017, sample a	Poor			Fair				
Rothstein et al., 2017, sample b	Poor			Poor	Fair			
Stein et al., 2015				Fair	Fair		Fair	

Appendix 8

Methodological quality by sample and measurement property for the Appearance Evaluation subscale and Body Areas Satisfaction Scale of the Multidimensional Body Relations Questionnaire (MBSRQ; Brown, Cash, & Mikulka, 1990; Cash, 2000).

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Argyrides & Kkeli, 2013	Fair	Fair		Fair	Fair	Fair	Fair	
Brytek-Matera & Rogoza, 2015	Poor			Poor				
Cruzat-Mandich et al., 2019	Fair			Fair				
Kashubeck-West et al., 2013	Good			Excellent	Fair			
Kelly et al., 2012, sample a	Good			Good	Fair			
Kelly et al., 2012, sample b	Good			Good	Fair			
Marco et al., 2017	Fair			Fair	Poor			
Naqvi & Kamal, 2017, sample a		Fair				Good		
Naqvi & Kamal, 2017, sample b	Fair			Fair				

Sample	Internal consistency	Reliability	Content validity	Structural validity	Hypotheses testing	Translation validity	Criterion validity	Responsiveness
Naqvi & Kamal, 2017, sample c				Fair				
Nevill et al., 2015		Poor						
Roncero et al., 2015	Fair			Fair	Fair	Poor		
Rusticus & Hubley, 2006				Fair	Poor			
Sabiston et al., 2010, sample a	Good			Fair	Fair			
Sabiston et al., 2010, sample b	Good			Fair	Fair			
Smith & Davenport, 2012	Poor				Fair			
Thoma et al., 2005		Fair			Fair			Fair
Untas et al., 2009	Fair	Fair		Fair	Poor	Fair		
Vossbeck- Elsebusch et al., 2014	Good	Fair		Good	Good	Poor		

Appendix 9

Body Appreciation Scale (BAS): Sample characteristics

Sample	Country	Language	Setting	*n* (female)	Age: Mean (*SD*)	Other characteristics
Alcaraz-Ibañes et al., 2017, sample a	Brazil	Brazilian Portuguese	School	438 (0%)	15.50 (1.20)	
Alcaraz-Ibañes et al., 2017, sample b	Brazil	Brazilian Portuguese	School	402 (100%)	15.51 (1.18)	
Alcaraz-Ibañes et al., 2017, sample c	Brazil	Brazilian Portuguese	School	46 (59%)	14.02 (0.93)	
Alexias et al., 2016	Greece	Greek	General population	2312 (71%)	31 (11.69)	
Alleva et al., 2016	Netherlands	Dutch	University	310 (100%)	21.31 (3.04)	
Atari, 2016, sample a	Iran	Persian	University	568 (0%)	26.16 (4.08)	
Atari, 2016, sample b	Iran	Persian	University	525 (100%)	25.54 (5.21)	
Avalos et al., 2005, sample a	USA	English	University	181 (100%)	20.24 (5.17)	82 % Caucasian American
Avalos et al., 2005, sample b	USA	English	University	327 (100 %)	18.45 (1.04)	88 % Caucasian American

Sample	Country	Language	Setting	*n* (female)	Age: Mean (*SD*)	Other characteristics
Avalos et al., 2005, sample c	USA	English	University	424 (100%)	19.86 (4.64)	78 % Caucasian American
Avalos et al., 2005, sample d	USA	English	University	177 (100%)	22.34 (6.93)	94 % Caucasian American
Cotter et al., 2015	USA	English	University	228 (100%)	19.89 (4.57)	Black
Ferreira et al., 2014	Brazil	Brazilian Portuguese	General population	424 (70%)	68.7 (0.98)	Older adults
Jauregui et al., 2011, sample a	Spain	Spanish	School	312 (47%)	14.81 (1.94)	Adolescents
Jauregui et al., 2011, sample b	Spain	Spanish	School	160 (49%)	15.01 (1.67)	Adolescents
Kertechian & Swami, 2017, sample a	France	French	University	174 (100%)	21.33 (3.18; females and males combined)	
Kertechian & Swami, 2017, sample b	France	French	University	152 (0%)	21.33 (3.18; females and males combined)	
Kertechian & Swami, 2017, sample c	France	French	University	326 (46%)	21.33 (3.18; females and males combined)	

Sample	Country	Language	Setting	*n* (female)	Age: Mean (*SD*)	Other characteristics
Lemoine et al., 2018, sample a	Denmark	Danish	School	79 (100%)	14.4 (2.1; females and males combined)	
Lemoine et al., 2018, sample b	Denmark	Danish	School	50 (0%)	14.4 (2.1; females and males combined)	
Lemoine et al., 2018, sample c	Portugal	Portuguese	School	296 (100%)	15.0 (2.1; females and males combined)	
Lemoine et al., 2018, sample d	Portugal	Portuguese	School	217 (0%)	15.0 (2.1; females and males combined)	
Lemoine et al., 2018, sample e	Sweden	Swedish	School	155 (100%)	15.5 (1.3; females and males combined)	
Lemoine et al., 2018, sample f	Sweden	Swedish	School	215 (0%)	15.5 (1.3; females and males combined)	
Moreira et al., 2018	Brazil	Brazilian Portuguese	School	347 (51%)	11.10 (0.81)	
Ng et al., 2015	China (Hong Kong)	Cantonese	University	2403 (55%)	23.52 (10.26)	
Razmuz & Razmuz, 2017, sample a	Poland	Polish	Unclear	171 (100%)	34.95 (10.83; females and males combined)	

Sample	Country	Language	Setting	*n* (female)	Age: Mean (*SD*)	Other characteristics
Razmuz & Razmuz, 2017, sample b	Poland	Polish	Unclear	165 (0%)	34.95 (10.83; females and males combined)	
Razmuz & Razmuz, 2017, sample c	Poland	Polish	Unclear	385 (55%)	35.38 (10.83)	
Swami & Chamorro-Premuzic, 2008	Malaysia	Malay	General population	591 (100%)	42.96 (12.98; Malay); 43.18 (13.30; Chinese)	53% Malay, 47% Malaysian Chinese
Swami & Jaafar, 2012, sample a	Indonesia	Indonesian (Bahasa Indonesia)	General population	262 (100%)	43.19 (12.95; females and males combined)	48% Javanese, 44% Sundanese, and 8% Chinese ancestry (females and males combined)
Swami & Jaafar, 2012, sample b	Indonesia	Indonesian (Bahasa Indonesia)	General population	278 (0%)	43.19 (12.95; females and males combined)	48% Javanese, 44% Sundanese, and 8% Chinese ancestry (females and males combined)
Swami & Ng 2015, sample a	China (Hong Kong)	Cantonese	University	457 (100%)	19.97 (4.58; females and males combined)	
Swami & Ng, 2015, sample b	China (Hong Kong)	Cantonese	University	417 (0%)	19.97 (4.58; females and males combined)	

Sample	Country	Language	Setting	*n* (female)	Age: Mean (*SD*)	Other characteristics
Swami et al., 2008, sample a	Austria	German	General population	156 (100%)	31.66 (13.60)	
Swami et al., 2008, sample b	Austria	German	General population	144 (0%)	31.31 (15.05)	
Swami et al., 2015	Turkey	Turkish	University	501 (100%)	22.05 (1.81)	
Swami et al., 2016a, sample a	China	Standard Chinese	University	191 (100%)	22.41 (5.30; females and males combined)	
Swami et al., 2016a, sample b	China	Standard Chinese	University	154 (0%)	22.41 (5.30; females and males combined)	
Swami et al., 2016a, sample c	China	Standard Chinese	University	345 (55%)	22.41 (5.3; females and males combined)	
Swami et al., 2017, sample a	Romania	Romanian	University	100 (100%)	23.57 (7.86; females and males combined)	
Swami et al., 2017, sample b	Romania	Romanian	University	100 (0%)	23.57 (7.86; females and males combined)	
Swami et al., 2017, sample c	Romania	Romanian	University	243 (46%)	23.57 (7.86; females and males combined)	

Sample	Country	Language	Setting	*n* (female)	Age: Mean (*SD*)	Other characteristics
Swami et al., 2017, sample d	Romania	Romanian	University	109 (52%)	25.02 (8.91)	
Taylor et al., 2013	Poland/UK	Polish	General population	306 (100%)	Polish: 33.45 (13.05); British-Polish: 34.63 (13.11)	50% Polish, 50% British-Polish
Tylka, 2013	USA	English	University	930 (57%)	19.91 (3.47)	White
Tylka & Wood-Barcalow, 2015b, sample a	USA	English	University	675 (54%)	20.34 (5.08)	79% White
Tylka & Wood-Barcalow, 2015b, sample b	USA	English	University	263 (61%)	20.43 (6.04)	81% White
Tylka & Wood-Barcalow, 2015b, sample c	USA	English	General population	317 (47%)	32.89 (10.10)	80% White
Tylka & Wood-Barcalow, 2015b, sample d	USA	English	General population	382 (50%)	33.38 (11.08)	72% White

Body Appreciation Scale (BAS): Measurement properties by sample

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Alcaraz-Ibañes et al., 2017, sample a	BAS-2	.93		1 factor (CFA*)	Convergent and discriminant validity supported. Males scored significantly higher than females. ** Underweight/normal weight participants scored significantly higher than overweight participants.	Factor structure invariant across gender and weight status.
Alcaraz-Ibañes et al., 2017, sample b	BAS-2	.93		1 factor (CFA)	Convergent and discriminant validity supported.	Factor structure invariant across gender and weight status.
Alcaraz-Ibañes et al., 2017, sample c	BAS-2		ICC = .98			Time interval for test-retest: 2 weeks
Alexias et al., 2016	BAS	.87	*r* = .88	1 factor (CFA)	Convergent and discriminant validity supported.	Time interval for test-retest: 3 weeks
Alleva et al., 2016	BAS-2	.90		1 factor (EFA)	Convergent and incremental validity supported.	

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Atari, 2016, sample a	BAS-2	.89		1 factor (EFA)	Convergent validity supported. No significant difference between males and females.	
Atari, 2016, sample b	BAS-2	.87		1 factor (EFA)	Convergent validity supported.	
Avalos et al., 2005, sample a	BAS	.94			Convergent and incremental validity supported.	Content validity assessed and supported.
Avalos et al., 2005, sample b	BAS			1 factor (EFA)		
Avalos et al., 2005, sample c	BAS				Convergent, discriminant and incremental validity supported.	
Avalos et al., 2005, sample d	BAS	.91/.93	*r* = .90			Time interval for test-retest: 3 weeks
Cotter et al., 2015	BAS	.92		1 factor (EFA and CFA)	Convergent and discriminant validity supported.	

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Ferreira et al., 2014	BAS	.79 (BV); .82 (BC)		2 factors (CFA): Body Valorization (BV) and Body Care (BC)	Convergent, discriminant, and concurrent validity supported for the 2-factor (BV and BC) model.	Content validity assessed and supported.
Jauregui et al., 2011, sample a	BAS	.91		1 factor (PCA)	Convergent and discriminant validity supported. Males scored significantly higher than females.	
Jauregui et al., 2011, sample b	BAS	.88/.90	*r* = .87			Time interval for test-retest: 3 weeks
Kertechian & Swami, 2017, sample a	BAS-2	.92		1 factor (EFA)	Females scored significantly lower than males.	
Kertechian & Swami, 2017, sample b	BAS-2	.92		1 factor (EFA)		
Kertechian & Swami, 2017, sample c	BAS-2	Females: .91; Males: .92		1 factor (CFA)		Factor structure invariant across gender.

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Lemoine et al., 2018, sample a	BAS-2	.93		1 factor (CFA)	Convergent validity supported. Females scored significantly lower than males.	Factor structure partly invariant across Danish, Portuguese, and Swedish samples.
Lemoine et al., 2018, sample b	BAS-2	.92		1 factor (CFA)	Convergent validity supported.	
Lemoine et al., 2018, sample c	BAS-2	.94		1 factor (CFA)	Convergent validity supported. Females scored significantly lower than males.	
Lemoine et al., 2018, sample d	BAS-2	.91		1 factor (CFA)	Convergent validity supported.	
Lemoine et al., 2018, sample e	BAS-2	.95		1 factor (CFA)	Convergent validity supported. Females scored significantly lower than males.	
Lemoine et al., 2018, sample f	BAS-2	.94		1 factor (CFA)	Convergent validity supported.	

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Moreira et al., 2018	BAS	.80 (females: .85; males: .75)		1 factor (CFA)	Convergent and criterion validity supported. Normal weight participants scored significantly higher than obese participants.	Content validity assessed and supported.
Ng et al., 2015	BAS	Females: .92 (GBA); .64 (BII). Males: .90 (GBA); .61 (BII).		2 factors (CFA): General Body Appreciation (GBA) and Body Image Investment (BII)	Convergent and discriminant validity supported for General Body Appreciation subscale.	Factor structure invariant across gender.
Razmuz & Razmuz, 2017, sample a	BAS-2	.94		1 factor (EFA)		
Razmuz & Razmuz, 2017, sample b	BAS-2	.96		1 factor (EFA)		
Razmuz & Razmuz, 2017, sample c	BAS-2	.94 (females: .93; males: .95)		1 factor (CFA)	Convergent validity supported. No significant difference between females and males.	Factor structure invariant across gender.

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information

Swami & Chamorro-Premuzic, 2008	BAS	.95 (GBA);
		.74 (BII, Malaysian);
		.71 (BII, Malaysian Chinese)
				2 factors (PCA/CFA): General Body Appreciation (GBA) and Body Image Investment (BII)	Convergent and discriminant validity supported. No significant ethnic differences.	CFA failed to replicate the original structure.
						Factor structure invariant across ethnic groups.

Swami & Jaafar, 2012, sample a	BAS	.93 (GBA); .72 (BII)		2 factors (EFA): General Body Appreciation (GBA) and Body Image Investment (BII)	Males scored significantly higher than females. No significant ethnic differences.	
Swami & Jaafar, 2012, sample b	BAS	.90 (GBA); .68 (BII)		2 factors (EFA): General Body Appreciation (GBA) and Body Image Investment (BII)		

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Swami & Ng 2015, sample a	BAS-2	.90		1 factor (EFA)	Convergent and discriminant validity supported. Males scored significantly higher than females.	
Swami & Ng, 2015, sample b	BAS-2	.91		1 factor (EFA)	Convergent and discriminant validity supported.	
Swami et al., 2008, sample a	BAS	.90		1 factor (EFA)	Convergent validity supported. Lower BMIs associated with greater BAS scores. Males scored significantly higher than females.	
Swami et al., 2008, sample b	BAS	.85		1 factor (EFA)	Convergent validity supported.	
Swami et al., 2015	BAS	.88		1 factor (EFA)	Convergent validity supported.	
Swami et al., 2016a, sample a	BAS-2	.89		1 factor (CFA)	Convergent validity supported. No significant difference between females and males.	Factor structure invariant across gender.

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Swami et al., 2016a, sample b	BAS-2	.86		1 factor (CFA)	Convergent validity supported.	
Swami et al., 2016a, sample c	BAS-2	.89		1 factor (CFA)		
Swami et al., 2017, sample a	BAS-2	.93		1 factor (EFA)	Convergent and discriminant validity supported.	
Swami et al., 2017, sample b	BAS-2	.84		1 factor (EFA)	Convergent and discriminant validity supported.	
Swami et al., 2017, sample c	BAS-2	.89		1 factor (CFA)	Convergent and discriminant validity supported.	Factor structure not invariant across gender.
Swami et al., 2017, sample d	BAS-2		Females: ICC = .82; Males: ICC = .87			Time interval for test-retest: 3 weeks
Taylor et al., 2013	BAS	≥.83 (GBA); ≤.62 (BII)		2 factors (EFA): General Body Appreciation (GBA) and Body Image Investment (BII)	British-Polish participants scored significantly higher than Polish participants.	

Sample	Instrument version	Internal consistency (α)	Test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Tylka, 2013	BAS	Females: .94; Males: .92		1 factor (CFA)	Convergent validity supported among males.	Factor structure invariant across gender.
Tylka & Wood-Barcalow, 2015b, sample a	BAS-2	Females: .94; Males: .93	ICC = .90	1 factor (EFA)	Convergent, discriminant and incremental validity supported.	Content and criterion validity assessed.
Tylka & Wood-Barcalow, 2015b, sample b	BAS-2	Females: .96; Males: .96		1 factor (CFA)	Discriminant validity supported.	Factor structure invariant across gender and across university and community samples.
Tylka & Wood-Barcalow, 2015b, sample c	BAS-2	Females: .96; Males: .96		1 factor (CFA)	Discriminant validity supported.	
Tylka & Wood-Barcalow, 2015b, sample d	BAS-2	Females: .97; Males: .96		1 factor (CFA)	Convergent validity supported.	Altering item 8 did not change the factor structure.

*CFA = confirmatory factor analysis, EFA = exploratory factor analysis

Appendix 10

Body Esteem Scale for Adolescents and Adults (BESAA): Sample characteristics

Sample	Country	Language	Setting	*N* (female)	Age: Mean (*SD*)	Other characteristics
Confalonieri et al., 2008	Italy	Italian	School	674 (unclear)	13.33 (2.1)	
Cragun et al., 2013, sample a	USA	English	School	146 (0 %)	11.9 (0.54; females and males combined)	
Cragun et al., 2013, sample b	USA	English	School	153 (100 %)	11.9 (0.54; females and males combined)	
Franko et al., 2012	USA	English	University	173 (100 %)	19.8 (2.0)	Latina
Gallini, 2008	USA	English	School	196 (52 %)	9.8 (.78)	Children
Jónsdóttir et al., 2008	Iceland	Icelandic	School	316 (50 %)	12 – 14 years	
Mendelson et al., 2001	Canada	English	School	1334 (57 %)	16.8 (range 12-25)	

Body Esteem Scale for Adolescents and Adults (BESAA): Measurement properties by sample

Sample	Instrument version	Internal consistency (α)	Reliability/test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Confalonieri et al., 2008	Modified 14 items	.87 (AE: .80; WE: .87; A: .74)*		3 factors (EFA/CFA**): Appearance esteem (AE); Weight esteem (WE); Attribution (A)	Convergent validity supported. Males scored significantly higher than females on appearance and weight subscales, no difference in attribution.***	
Cragun et al., 2013, sample a	AE and WE subscales	.90 (AE); .90 (WE)		2 factors (CFA)	Convergent validity supported.	The factor structure did not display adequate fit. Content validity assessed and supported for AE and WE, not A.
Cragun et al., 2013, sample b	AE and WE subscales	.92 (AE); .93 (WE)		2 factors (CFA)	Convergent validity supported.	The factor structure did not display adequate fit.
Franko et al., 2012		Time 1: .91 (AE); .75 (WE); .94 (A). Time 2: .93 (AE); .68 (WE); .95 (A).	*r* =.93 (AE); *r* =.98 (WE); *r* =.89 (A)	3 factors (CFA)	Convergent validity supported. No significant differences in BESAA scores between Latina and Caucasian females.	Time interval for test-retest: 3-4 weeks

Gallini, 2008	Modified 24 items	.93 (AE: .90; WE: .90; A: .66)		3 factors (EFA)	Convergent validity supported.	Content validity assessed and resulted in modifications.
Jónsdóttir et al., 2008		.95 (AE: .92; WE: .92; A: .73)			Males scored significantly higher than females on appearance and weight subscales, no difference in attribution. Younger participants scored significantly higher than older participants on appearance and weight subscales, no difference in attribution.	Incremental validity assessed and supported. PCA resulted in 3 a factor solution.
Mendelson et al., 2001		.92 (AE); .94 (WE); .81 (A)	*r* =.89 (AE); *r* =.92 (WE); *r* =.83 (A)	3 factors (EFA)	Discriminant validity supported. Convergent validity supported for appearance subscale. Males scored significantly higher than females. Participants with higher weight scored significantly higher than participants with lower weight.	Time interval for test-retest: 3 months

*AE = appearance esteem subscale, WE = weight esteem subscale, A = attribution subscale; **CFA = confirmatory factor analysis, EFA = exploratory factor analysis, PCA = principal component analysis; ***Higher scores indicate higher body esteem

Appendix 11

Body shape questionnaire (BSQ): Sample characteristics

Sample	Country	Language	Setting	*N* (female)	Age: Mean (*SD*)	Other characteristics
Akdemir et al., 2012	Turkey	Turkish	School	665 (100%)	15.1 (.6)	
Conti et al., 2009	Brazil	Portuguese	School	386 (54%)	13.8 (2.1)	
Cooper et al., 1987, sample a	UK	English	General population	535 (100%)	Students: 20 (1.1); Occupational therapy students: 21.3 (3.2); Family planning clinic attenders: 23.8 (6.3)	
Cooper et al., 1987, sample b	UK	English	Medical	38 (100%)	22.2 (4.1)	Patients with bulimia nervosa
Di Pietro et al., 2009	Brazil	Brazilian Portuguese	University	164 (43%)	19.65 (1.5)	
Dowson & Henderson, 2001	UK	English	Medical	75 (100%)	24 (6.7)	Patients with psychogenic low weight and a history of full or partial anorexia nervosa

Evans & Dolan, 1993	UK	English	Medical	342 (100%)	27.1 (8.5)	Participants attending a family planning and well woman clinic
Franko et al., 2012	USA	English	University	173 (100%)	19.8 (2.0)	Latina
Ghaderi & Scott, 2004, sample a	Sweden	Swedish	General population	1157 (100%)	23.7 (3.7)	
Ghaderi & Scott, 2004, sample b	Sweden	Swedish	University	124 (81%)	28.8 (6.3)	
Ghaderi & Scott, 2004, sample c	Sweden	Swedish	Medical	90 (100%)	28.5 (9.6)	
Kapstad et al., 2015, sample a	Norway	Norwegian	School, University	690 (61%)	23.05 (8.67)	
Kapstad et al., 2015, sample b	Norway	Norwegian	Medical	49 (100%)	19.04 (3.06)	
Kim & Chee, 2018	Korea	Korean	General population	467 (79%)	27.6 (9.7)	
Lentillon-Kaestner et al., 2014, sample a	Switzerland	French	Medical	159 (100%)	48.40 (10.9)	Participants seeking dietetic counseling; 40% binge eating disorder

Lentillon-Kaestner et al., 2014, sample b	Switzerland	French	University	1169 (100%)	18.24 (2.82)	
Mazzeo, 1999, sample a	USA	English	University	302 (100%)	19.51 (1.31)	82% Caucasian
Mazzeo, 1999, sample b	USA	English	University	212 (100%)	19.59(1.17)	79% Caucasian
Mumford et al., 1991	UK	English	School	204 (100%)	15.1 (1.6)	South Asian British
Mumford et al., 1992	Pakistan	English	School	369 (100%)	14.3 (1)	
Pook et al., 2008, sample a	Germany	German	General population	1080 (100%)	50.3 (18.6)	
Pook et al., 2008 sample b	Germany	German	Medical	43 (100%)	Unclear	Patients with bulimia nervosa
Popkess-Vawter et al., 1992	USA	English	University	43 (100%)	30 (range 18-45)	
Probst et al., 2009	Belgium	Flemish	University	816 (48%)	Females: 17.3 (2.1); Males: 17.2 (2.0)	Caucasian
Reilly et al., 2014	USA	English	University	590 (60%)	Unclear	
Rosen et al., 1996, sample a	USA	English	Medical	155 (100%)	35.6 (11.4)	Patients with body image problems

Rosen et al., 1996, sample b	USA	English	Medical	83 (86%)	Females: 41.7 (11); Males: 46.7 (8.4)	Participants with obesity
Rosen et al., 1996, sample c	USA	English	University	163 (100%)	18.5 (1.9)	University students
Rosen et al., 1996, sample d	USA	English	University	89 (100%)	41.4 (10)	University staff
Silva et al., 2014	Brazil	Portuguese	University	739 (100%)	20.44 (2.45)	
Silva et al., 2016, sample a	Portugal	Portuguese	University	278 (100%)	20.9 (2.4)	
Silva et al., 2016, sample b	Portugal	Portuguese	University	248 (100%)	20.9 (2.3)	
Warren et al., 2008, sample a	USA	English	University	505 (100%)	19.34 (1.9)	Euro-American
Warren et al., 2008, sample b	USA	English	University	151 (100%)	19.62 (1.94)	Hispanic American
Warren et al., 2008, sample c	Spain	Spanish	University	445 (100%)	20.83 (3.49)	
Warren et al., 2008, sample d	Spain	Spanish	Medical	177 (100%)	20.42 (5.17)	Patients with eating disorders

Welch et al., 2012, sample a	Sweden	Swedish	University	182 (69%)	Unclear	Undergraduate students
Welch et al., 2012, sample b	Sweden	Swedish	General population	747 (100%)	23.9 (3.9)	

Body shape questionnaire (BSQ): Measurement properties by sample

Sample	Instrument version	Internal consistency (α)	Reliability/test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Akdemir et al., 2012		.96	*r* = .81	3 factors (EFA*; general body dissatisfaction; vomiting and laxative use; social avoidance)	Convergent validity supported.	Time interval for test-retest: 4 weeks
Conti et al., 2009		.96	*r* = .91		Convergent and discriminant validity supported.	Time interval for test-retest not described. Criterion validity assessed and supported.
Cooper et al., 1987, sample a					Convergent validity supported. Participants with BN** scored significantly higher than participants without BN.***	
Cooper et al., 1987, sample b					Convergent validity supported.	

Di Pietro et al., 2009		.97			Females scored significantly higher than males.	PCA resulted in a four-factor solution.
Dowson & Henderson, 2001	14-item	.93			Convergent validity supported.	
Evans & Dolan, 1993	Full; 16-item; 8-item	.97 (34 items); .93 - .96 (16 items); .87 - .92 (8 items)		1 factor (CFA)	Convergent validity supported for all three versions of BSQ.	
Franko et al., 2012	Full	Time 1: .82; Time 2: .88	ICC = .97	1 factor (CFA)	Convergent validity supported. No significant differences in BSQ scores between Latina and Caucasian females.	Time interval for test-retest: 3-4 weeks
Ghaderi & Scott, 2004, sample a	Full; 14-item	.94 - .97		1 factor (EFA; full version)	Convergent and discriminant validity supported. Participants with eating disorders scored significantly higher than participants without eating disorders.	PCA with 14-item version resulted in a 1-factor solution.

Ghaderi & Scott, 2004, sample b	Full; 14-item	.97	*r* = .90		Convergent and discriminant validity supported.	Time interval for test-retest: 2 weeks
Ghaderi & Scott, 2004, sample c	Full; 14-item	.94			Convergent and discriminant validity supported.	
Kapstad et al., 2015, sample a	Full; 14-item	.97	Females: *r* = .94; Males: *r* = .86. Scores were significantly lower at Time 2.		Convergent validity supported. Female patients scored significantly higher than female controls.	Time interval for test-retest: 1 week Criterion validity assessed and partly supported.
Kapstad et al., 2015, sample b	Full; 14-item	.94			Convergent validity supported.	
Kim & Chen, 2018		.97	*r* = .93	4 factors (EFA; feeling fat, shame and inferiority about one’s body shape, attitudes concerning body image perception, purging behavior)	Convergent validity supported. Females scored significantly higher than males.	Time interval for test-retest: 2 weeks

Lentillon-Kaestner et al., 2014, sample a	Full; 16-item; 14-item; 8-item	.68 - .96 (depending on BSQ version and BED/not BED)		2 factors for 34-item structure; 1 factor for short forms (EFA/CFA)	Convergent validity supported. Participants with BED** scored significantly higher than participants without BED.	
Lentillon-Kaestner et al., 2014, sample b	Full; 16-item; 14-item; 8-item	.81 - .96 (depending on BSQ versions)	*r* = .97 (all BSQ versions)	2 factors for 34-item structure; 1 or 2 factors for short forms (EFA/CFA)	Convergent validity supported. Participants seeking dietetic counseling scored significantly higher than students.	Time interval for test-retest: 3 weeks
Mazzeo, 1999, sample a	BSQ-R-10		*r* = .91	1 factor (EFA)		Time interval for test-retest: 3 weeks Criterion validity assessed and partly supported.
Mazzeo, 1999, sample b	BSQ-R-10	.96		1 factor (EFA)		Criterion validity assessed and partly supported.

Mumford et al., 1991				1 factor (EFA)	Convergent validity supported. No significant difference between Asian participants born in the UK, Asian participants born abroad, and Caucasian participants.	
Mumford et al., 1992				1 (EFA)	Convergent validity supported	
Pook et al., 2008, sample a	Full; 16-item; 14-item; 8-item	.97 (full); .88 - .95 (short versions)		1 (CFA)		Original 34-item structure not supported.
Pook et al., 2008, sample b	Full; 16-item; 14-item; 8-item					Responsiveness assessed and supported for 8-item version.
Popkess-Vawter et al., 1992		.96	*r* = .97			Time interval for test-retest: 2 weeks Criterion validity assessed.

Probst et al., 2009	.96 - .97		Convergent validity supported. Females scored significantly higher than males.	
Reilly et al., 2014	Females: .98; Males: .97			The 34-items of the BSQ were evaluated for gender based DIF****. One item evidenced clinically significant DIF.
Rosen et al., 1996, sample a			Convergent validity supported.	Criterion validity assessed.
Rosen et al., 1996, sample b			Convergent validity supported.	Criterion validity assessed and partly supported.
Rosen et al., 1996, sample c		*r* = .88	Convergent validity supported.	Time interval for test-retest: 3 weeks
				Criterion validity assessed and partly supported.

Rosen et al., 1996, sample d					Convergent validity supported.	Criterion validity assessed and partly supported.
Silva et al., 2014	Full; 16-item; 8-item	.97 (34 items); .93 (16 items); .88 (8 items)		1 factor (CFA)	Convergent validity supported for all three versions of BSQ.	Content validity assessed and partly supported.
Silva et al., 2016, sample a	Full; 8-item	.97 (refined 32-item BSQ); .87 (8-item version)		1 factor (CFA; refined 32-item BSQ; 8-item version)	Convergent validity supported.	Content validity assessed.
Silva et al., 2016, sample b	Full; 8-item	.97 (refined 32-item BSQ); .88 (8-item version)		1 factor (CFA; refined 32-item BSQ; 8-item version)	Convergent validity supported.	Content validity assessed.
Warren et al., 2008, sample a	Full; 16-item; 14-item; 8-item	.98 (34 items); .90 - .96 (short versions)		1 factor (CFA; all BSQ versions)		All factor structures displayed invariance across groups (Euro-American, Hispanic American, Spanish, clinical Spanish). A 10-item version displayed best fit.

Warren et al., 2008, sample b	Full; 16-item; 14-item; 8-item	.97 (34 items); .89 - .95 (short versions)		1 factor (CFA; all BSQ versions)		All factor structures displayed invariance across groups (Euro-American, Hispanic American, Spanish, clinical Spanish). A 10-item structure displayed best fit.
Warren et al., 2008, sample c	Full; 16-item; 14-item; 8-item	.97 (34 items); .87 - .95 (for short versions)		1 factor (CFA; all BSQ versions)		All factor structures displayed invariance across groups (Euro-American, Hispanic American, Spanish, clinical Spanish). 10-item structure displayed best fit.
Warren et al., 2008, sample d	Full; 16-item; 14-item; 8-item	.96 (34 items); .83 - .93 (short versions)		1 factor (CFA; all BSQ versions)		All factor structures displayed invariance across groups (Euro-American, Hispanic American, Spanish, clinical Spanish). 10-item structure displayed best fit.

Welch et al., 2012, sample a	BSQ-8C	Time 1: .92; Time 2: .93	*r* = .95			Time interval for test-retest: 15.1 days (*SD* = 4.3)
Welch et al., 2012, sample b	BSQ-8C	.94		1 factor (EFA/CFA)	Convergent validity supported.	

*CFA = confirmatory factor analysis, EFA = exploratory factor analysis, PCA = principal component analysis; **BN = bulimia nervosa, BED = binge eating disorder; ***Higher scores indicate higher body shape concerns; ****DIF=differential item functioning

Appendix 12

Centre for Appearance Research Valence Scale (CARVAL): Sample characteristics

Sample	Country	Language	Setting	*N* (female)	Age: Mean (*SD*)	Other characteristics
Moss & Rosser, 2012, sample a	Worldwide	English	General population	592 (81%)	25.1 (8.54)	Predominantly American (31%)
Moss & Rosser, 2012, sample b	UK	English	University	41 (83%)	21.2 (1.82)	
Moss et al., 2014	UK	English	Medical setting	1265 (67%)	47.32 (16.72)	Participants with visible difference

Centre for Appearance Research Valence Scale (CARVAL): Measurement properties by sample

Sample	Internal consistency (α)	Reliability/test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Cross-cultural validity/Translation process	Additional information
Moss & Rosser, 2012, sample a	.93			Convergent validity supported		Content validity assessed. Criterion validity assessed. PCA resulted in a 1-factor solution.
Moss & Rosser, 2012, sample b		*r* = .89				Time interval for test-retest: 1 month
Moss et al., 2014	.88	*r* = .69	1 factor (EFA)	Convergent and discriminant validity supported. Females scored significantly higher than males.**		Time interval for test-retest: 9 months

*EFA = exploratory factor analysis, PCA = principal component analysis; **Higher scores indicate a more negatively valenced evaluation of appearance

Appendix 13

Drive for Muscularity Scale (DMS): Sample characteristics

Sample	Country	Language	Setting	*N* (female)	Age: Mean (*SD*)	Other characteristics
Cafri & Thompson, 2004, sample a	USA	English	University	76 (0%)	21.12 (2.60)	
Cafri & Thompson, 2004, sample b	USA	English	University	103 (100%)	20.81 (2.48)	
Campana et al., 2013	Brazil	Brazilian Portuguese	Military, University, general population	878 (0%)	20.9 (4.74)	
Chaba et al., 2018, sample a	France/ Switzerland	French	General population	114 (0%)	23.35 (4.93)	Bodybuilding or strength training athletes
Chaba et al., 2018, sample b	France/ Switzerland	French	General population	129 (0%)	27.03 (7.81)	Bodybuilding or strength training athletes
Compte et al., 2015	Argentina	Spanish	University	423 (0%)	22.47 (5.21)	
DeBlaere et al., 2017	USA	English	General population	202 (0%)	28.80 (14.50)	Genderual minority males, 73 % White

Escoto et al., 2013	Mexico	Spanish	University	569 (0%)	20.89 (2.00)	
Keum et al., 2015	USA/Canada	English	Online communities	200 (0%)	27.9 (7.45)	90 % Asian-American, 10 % Asian-Canadian
McCreary & Sasse, 2000	Canada	English	High school	197 (51%)	18 (range 16-24)	
McCreary et al., 2004, sample a	Canada	English	High school, College	276 (0%)	17.5 (3.9; females and males combined)	
McCreary et al., 2004, sample b	Canada	English	High school, College	354 (100%)	17.5 (3.9; females and males combined)	
McPherson et al., 2010	UK (Scotland)	English	General population	594 (0%)	38.9 (9.8)	Males participating in an organized running event
Nerini et al., 2015, sample a	Italy	Italian	General population	212 (0%)	24.39 (4.25)	Heterogenderual males
Nerini et al., 2015, sample b	Italy	Italian	General population	143 (0%)	36.97 (10.31)	Gay males
Sepulveda et al., 2016	Spain	Spanish	School	212 (0%)	14.4 (1.5)	Adolescents

Swami et al., 2016b	Malaysia	Malay	General population	159 (0%)	28.78 (9.35)	
Swami et al., 2018	Romania	Romanian	University	343 (0%)	22.48 (6.02)	
Tod et al., 2012, sample a	UK	English	University	272 (0%)	20.3 (4.0)	
Tod et al., 2012, sample b	UK	English	University	54 (0%)	19.3 (2.2)	
Wojtowicz & von Ranson, 2006, sample a	Canada	English	University	51 (100%)	21.8 (3.9; females and males combined)	41% weightlifters
Wojtowicz & von Ranson, 2006, sample b	Canada	English	University	53 (0%)	21.8 (3.9; females and males combined)	51% weightlifters

Drive for Muscularity Scale (DMS): Measurement properties by sample

Sample	Instrument version	Internal consistency (α)	Reliability/test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Cafri & Thompson, 2004, sample a		.89 (Attitudes: .88; Behaviors: .86)	*r* =.93 (Attitudes: *r* = .84; Behaviors: *r* = .96)		Convergent validity supported. Males scored significantly higher than females.*	Time interval for test-retest: 7-10 days. Criterion validity assessed.
Cafri & Thompson, 2004, sample b		.81			Convergent validity supported.	Criterion validity assessed.
Campana et al., 2013		.87 (Attitudes); .86 (Behaviors)		2 factors (CFA**): Attitudes and Behaviors	Convergent and discriminant validity supported.	Content validity assessed but not supported.
Chaba et al., 2018, sample a		.87 (Attitudes); .85 (Behaviors)		2 factors (EFA): Attitudes and Behaviors		

Chaba et al., 2018, sample b		.87 (Attitudes: .87; Behaviors: .85)	*r* =.86 (Attitudes: *r* = .83; Behaviors: *r* = .86)	2 factors (CFA): Attitudes and Behaviors	Convergent validity supported.	Time interval for test-retest: 4 weeks
Compte et al., 2015		.89 (Attitudes: .91; Behaviors: .86)		2 factors (CFA): Attitudes and Behaviors	Convergent validity supported.	
DeBlaere et al., 2017		.93 (Attitudes: .93; Behaviors: .87)		2 factors (CFA): Attitudes and Behaviors	Convergent validity supported.	
Escoto et al., 2013		.86 (Attitudes: .87; Behaviors: .79; supplement consumption: .72; training adherence: .68)		3 factors (EFA/CFA): Attitudes, supplement consumption and training adherence		
Keum et al., 2015	12 item	.87 (Attitudes: .91; Behaviors: .82)		2 factors (CFA/EFA): Attitudes and Behaviors		

McCreary & Sasse, 2000	.84 (females: .78; males: .84)			Convergent and discriminant validity supported. Males scored significantly higher than females. Participants striving to gain weight scored significantly higher than participants not striving to gain weight.	
McCreary et al., 2004, sample a	.87 (Attitudes: .88; Behaviors: .81)		2 factors (EFA): Attitudes and Behaviors		2-factor structure for males
McCreary et al., 2004, sample b	.82		1 factor (EFA)		1-factor structure for females
McPherson et al., 2010	.91 (Attitudes: .92; Behaviors: .85)	*r* =.92 (Attitudes: *r* = .92; Behaviors: *r* = .86)	2 factors (EFA): Attitudes and Behaviors	Participants altering their food intake to gain muscle scored significantly higher than participants who did not.	Time interval for test-retest: 4 weeks

Nerini et al., 2015, sample a	.84 (Attitudes: .89; Behaviors: .81)		2 factors (CFA): Attitudes and Behaviors	Convergent validity supported. Gay participants scored significantly higher than heterogenderual participants.	
Nerini et al., 2015, sample b	.90 (Attitudes: .91; Behaviors: .85)		2 factors (CFA): Attitudes and Behaviors	Convergent validity supported.	
Sepulveda et al., 2016	.89 (Attitudes: .92; Behaviors: .87)		2 factors (CFA): Attitudes and Behaviors	Convergent and discriminant validity supported.	
Swami et al., 2016b	.91 (Attitudes); .90 (Behaviors)		2 factors (EFA): Attitudes and Behaviors	Convergent validity supported.	
Swami et al., 2018	.80 (Attitudes); .84 (Behaviors)		2 factors (EFA/CFA): Attitudes and Behaviors	Convergent validity supported.	
Tod et al., 2012, sample a	.91 (Attitudes); .89 (Behaviors)			Convergent and discriminant validity supported.	Time interval for test-retest: 7 days and 14 days

Tod et al., 2012, sample b		Time 1: .85 (Attitudes); .78 (Behaviors). Time 2: .87 (Attitudes); .85 (Behaviors). Time 3: 91 (Attitudes); .88 (Behaviors).	Time 1-Time 2: ICC = .82 (Attitudes); ICC = .81 (Behaviors). Time 1-Time 3: ICC = .70 (Attitudes); ICC = .89 (Behaviors).		
Wojtowicz & von Ranson, 2006, sample a		Weightlifters: .80; Non-weightlifters: .76			Convergent and discriminant validity supported. Males scored significantly higher than females. Weightlifters scored significantly higher than non-weightlifters.
Wojtowicz & von Ranson, 2006, sample b		Weightlifters: .84 (Attitudes: .78; Behaviors: .87); Non-weightlifters: .80 (Attitudes: .80; Behaviors: .73)			Convergent and discriminant validity supported. Weightlifters scored significantly higher than non-weightlifters on behaviors subscale.

*Higher scores indicate higher drive for muscularity; ** CFA = confirmatory factor analysis, EFA = exploratory factor analysis, PCA = principal component analysis

Appendix 14

Weight concerns subscale (WC) and Shape concerns subscale (SC) of the Eating Disorders Examination Questionnaire (EDE-Q): Sample characteristics

Sample	Country	Language	Setting	*N* (female)	Age: Mean (*SD*)	Other characteristics
Allen et al., 2011, sample a	Australia	English	Medical	228 (100%)	26.02 (9.09)	Patients with eating disorders
Allen et al., 2011, sample b	Australia	English	University	211 (100%)	21.03 (5.85)	
Bardone-Cone & Boyd, 2007, sample a	USA	English	University	97 (100%)	19.04 (1.59)	Black
Bardone-Cone & Boyd, 2007, sample b	USA	English	University	179 (100%)	18.58 (1.06)	White
Barnes et al., 2012	UK	English	University, Eating disorder charities	569 (92% of students, 96% of charities participants)	Unclear	Adults
Becker et al., 2010	Fiji	Fijian/English	School	523 (100%)	16.67 (1.09)	
Binford et al., 2005	USA	English	Medical	70 (96%)	15.79 (2.28)	Patients with eating disorders

Calugi et al., 2016	Italy	Italian	Medical	264 (97%)	22.2 (6.3)	Patients with eating disorders (mainly anorexia nervosa)
Carrard et al., 2015, sample a	Switzerland	French	General population, medical	116 (100%)	38.5 (11.4)	Participants with binge eating disorder symptoms
Carrard et al., 2015, sample b	Switzerland	French	General population	161 (100%)	28.1 (8.1)	
Chan & Leung, 2015	China (Hong Kong)	English	University	310 (54%)	20.75 (1.81)	
Darcy et al., 2013, sample a	USA	English	University	429 (100%)	21.01 (1.7)	
Darcy et al., 2013, sample b	USA	English	University	229 (0%)	20.90 (1.71)	
Darcy et al., 2013, sample c	USA	English	University	544 (100%)	20.63(1.48)	Competitive athletes
Darcy et al., 2013, sample d	USA	English	University	432 (0%)	21.03(1.77)	Competitive athletes
Elder & Grilo, 2007	USA	Spanish	General population	77 (100%)	41.5 (13.6)	Diverse backgrounds of Spanish speaking countries
Franko et al., 2012	USA	English	University	173 (100%)	19.8 (2.0)	Latina
Gideon et al., 2016	UK	English	Medical	489 (90%)	31.5 (11.5)	Patients with eating disorders

Giovazolias et al., 2013, sample a	Greece	Greek	University	500 (100%)	20.55 (3.27)	
Giovazolias et al., 2013, sample b	Greece	Greek	University	164 (100%)	20.90 (3.29)	
Grilo et al., 2013	USA	English	Medical	174 (75%)	42.9 (11.1)	Obese bariatric surgery candidates
Grilo et al., 2015	USA	English	University	801 (72%)	20 (2.5)	
Heiss et al., 2018, sample a	USA	English	General population	318 (82%)	31.76 (12.62)	Vegans
Heiss et al., 2018, sample b	USA	English	University	200 (63%)	18.86 (1.97)	Omnivores
Hilbert et al., 2012, sample a	Germany	German	General population	1354 (100%)	50.5 (18.59; females and males combined)	
Hilbert et al., 2012, sample b	Germany	German	General population	1166 (0%)	50.5 (18.59; females and males combined)	
Hrabosky et al., 2008	USA	English	Medical	337 (83%)	43.2 (10.5)	Obese bariatric surgery candidates
Isomaa et al., 2016, sample a	Finland	Finnish	School	242 (55%)	17.8 (range 15 - 24)	Adolescents
Isomaa et al., 2016, sample b	Finland	Finnish	Workplace	133 (51%)	46.1 (range 30 - 66)	Adults
Isomaa et al., 2016, sample c	Finland	Finnish	Medical	52 (96%)	27.8 (range 15 - 57)	Patients with eating disorders

Luce & Crowther, 1999	USA	English	University	139 (100%)	18.5 (2)	
Lydecker et al., 2016, sample a	USA	English	University	119 (83%)	45.34 (9.80)	Black participants with binge eating disorder
Lydecker et al., 2016, sample b	USA	English	University	119 (83%)	44.80 (10.55)	White participants with binge eating disorder
Machado et al., 2014, sample a	Portugal	Portuguese	School, university	4091 (100%)	School: 16.2 (1.33); University: 21.5 (2.75)	
Machado et al., 2014, sample b	Portugal	Portuguese	Medical	554 (100%)	AN: 22.0 (7.04); BN: 26.1 (7.61): BED: 30.6 (11.70); EDNOS: 19.5 (6.08); Obese: 41.6 (10.68)	Patients with eating disorders or obesity
Machado et al., 2018, sample a	Portugal	Portuguese	School, university	4117 (100%)	School: 16.2 (1.3); University: 21.7 (3.82)	
Machado et al., 2018, sample b	Portugal	Portuguese	Medical	609 (97%)	23.8 (9.16)	Patients with eating disorders
Mahmoodi et al., 2016	Iran	Persian	University	516 (100%)	23.71 (3.14)	
Mitsui et al., 2017, sample a	Japan	Japanese	University	1430 (72%)	19.4 (1.3)	
Mitsui et al., 2017, sample b	Japan	Japanese	University	558 (84%)	20.11 (2.52)	

Mitsui et al., 2017, sample c	Japan	Japanese	University	111 (100%)	18.52 (.77)	
Mitsui et al., 2017, sample d	Japan	Japanese	University	225 (100%)	19.6 (1.0)	
Mond et al., 2004a	Australia	English	General population	Unclear (100%); 802 participants enrolled	35.3 (8.5)	
Mond et al., 2004b	Australia	English	General population	495 (100%)	35.3 (8.5)	
Parker et al., 2015	Australia	English	Medical	108 (87%)	46 (12.2)	Post-bariatric surgical patients
Parker et al., 2016	Australia	English	Medical	405 (79%)	43.8 (11.6)	Bariatric surgery candidates
Peláez-Fernández et al., 2012	Spain	Spanish	School, university	1543 (59%)	15.73 (2.34)	
Penelo et al., 2012	Spain	Spanish	University	269 (0%)	23.3 (3.4)	
Penelo et al., 2013	Mexico	Spanish	School	2928 (53%)	15.1 (1.79)	
Peterson et al., 2007	USA	English	General population	203 (100%)	25.7 (8.9)	71% bulimia nervosa
Phillips et al., 2018	USA	English	Medical	169 (100%)	34.1 (13.7)	Patients with anorexia nervosa
Pretorius et al., 2009	UK	English	Medical	94 (unclear)	19.1 (1.6)	Participants with bulimia nervosa

Reas et al., 2006	USA	English	General population	86 (79%)	44.9 (8.9)	Participants with BMI > 27 and binge eating disorder diagnosis
Reas et al., 2012	Norway	Norwegian	School, university	250 (0%)	19.7 (2.3)	
Reilly et al., 2014	USA	English	University	1116 (67%)	Unclear	
Rø et al., 2010	Norway	Norwegian	University	670 (100%)	24.8 (6.9)	
Rose et al., 2013	USA	English	University	91 (48%)	19 (1.16)	
Unikel Santoncini et al., 2018, sample a	Mexico	Mexican Spanish	University	330 (100%)	19.3 (2.5)	
Unikel Santoncini et al., 2018, sample b	Mexico	Mexican Spanish	University	165 (100%)	22.0 (6.4)	Patients with eating disorders
Villarroel et al., 2011	Spain	Spanish	University	708 (100%)	22 (2.7)	
White et al., 2014, sample a	UK	English	School	458 (56%)	15.3 (1.18)	
White et al., 2014, sample b	UK	English	School	459 (58%)	15.2 (1.18)	
Yucel et al., 2011	Turkey	Turkish	School	925 (68%)	15.52 (1.88)	

Weight concerns subscale (WC) and Shape concerns subscale (SC) of the Eating Disorders Examination Questionnaire (EDE-Q): measurement properties by sample

Sample	Instrument version	Internal consistency (α)	Reliability/test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Allen et al., 2011, sample a	Brief 1-factor	.80 (WC); .88 (SC)		1 factor (CFA*, combined items from weight and shape concerns subscales)		Original factor structure not supported and not invariant across groups (eating disorder patients and controls). Criterion validity for 1-factor structure assessed and supported.
Allen et al., 2011, sample b	Brief 1-factor	.89 (WC); .93 (SC)		1 factor (CFA, combined items from weight and shape concerns subscales)		Original factor structure not supported and not invariant across groups (eating disorder patients and controls). Criterion validity for 1-factor structure assessed and supported.
Bardone-Cone & Boyd, 2007, sample a		.83 (WC); 89 (SC)	*r* = .81 (WC); *r* = .82 (SC)		Black participants scored significantly lower than white participants.**	Time interval for test-retest: 5.24 months

Bardone-Cone & Boyd, 2007, sample b	.84 (WC); .91 (SC)	*r* =.81 (WC); *r* = .80 (SC)		White participants scored significantly higher than black participants.	Time interval for test-retest: 5.32 months
Barnes et al., 2012	.94 (WC/SC)		1 factor (CFA; combined weight and shape concerns subscales, WC/SC)		
Becker et al., 2010	Fijian version: .66 (WC); .79 (SC). English version: .70 (WC); .84 (SC)	Fijian version: ICC = .56 (WC); ICC = .63 (SC). English version: ICC = .78 (WC); ICC = .70 (SC).	4 factors (EFA of all EDE-Q)	Convergent validity supported.	Original factor structure was not supported.
Binford et al., 2005					Criterion validity assessed and supported for all diagnostic groups (bulimia nervosa, partial syndrome bulimia nervosa, anorexia nervosa).

Calugi et al., 2016	Brief 3-factor	.80 (WC); .88 (SC)	ρ = .66 (WC); ρ = .80 (SC)	3 factors (CFA; dietary restraint, body dissatisfaction, and shape/weight overvaluation)	Participants with eating disorders scored significantly higher than controls.	Original factor structure was not supported.
Carrard et al., 2015, sample a	Brief 3-factor	.90 (shape/weight overvaluation); .71 (body dissatisfaction)		3 factors (CFA; dietary restraint, shape/weight overvaluation, body dissatisfaction)		Original factor structure was not supported. The 3-factor structure was invariant across binge eating disorder and control groups.
Carrard et al., 2015, sample b	Brief 3-factor	.95 (shape/weight overvaluation); .86 (body dissatisfaction)		3 factors (CFA; dietary restraint, shape/weight overvaluation, body dissatisfaction)		Original factor structure was not supported.
Chan & Leung, 2015	Brief 1-factor	.94 (WC/SC)		1 factor (CFA, combined items from weight and shape concerns subscales)	Convergent validity supported.	Factor structure not supported among males

Darcy et al., 2013, sample a		3 factors (EFA of all EDE-Q)	Original factor structure was not supported. Tendency of WC and SC to load onto same factors.
Darcy et al., 2013, sample b		2 factors (EFA of all EDE-Q)	Original factor structure was not supported. WC and SC loaded onto same factor.
Darcy et al., 2013, sample c		3 factors (EFA of all EDE-Q)	Original factor structure was not supported. WC and SC loaded onto same factor.
Darcy et al., 2013, sample d		3 factors (EFA of all EDE-Q)	Original factor structure was not supported. WC and SC loaded onto same factor.
Elder & Grilo, 2007	ρ = .73 (WC); ρ = .81(SC)		Time interval for test-retest: 5-14 days. Criterion validity assessed and supported.

Franko et al., 2012	Time 1: .83 (WC); .91 (SC). Time 2: .86 (WC); .94 (SC).	ICC = .95 (WC); ICC = .97 (SC)	2 factors (CFA, WC and SC)	Convergent validity supported. Latina participants scored significantly higher than Caucasian participants.	Time interval for test-retest: 3-4 weeks
Gideon et al., 2016	.70 (WC); .80 (SC)				PCA and Raschs analysis resulted in a 5-factor solution (based on all EDE-Q items). Content validity assessed.
Giovazolias et al., 2013, sample a	.91 (WC/SC)		1 factor (CFA, combined WC and SC)		
Giovazolias et al., 2013, sample b				Convergent and discriminant validity supported.	

Grilo et al., 2013	Brief 3-factor	.60 (WC); .83 (SC); .96 (shape/weight overvaluation); .69 (body dissatisfaction)		3 factors (CFA; dietary restraint, shape/weight overvaluation, body dissatisfaction)	Convergent validity supported for WC and SC, as well as shape/weight overvaluation and body dissatisfaction.	Original factor structure was not supported.
Grilo et al., 2015	Brief 3-factor	.86 (WC); .91 (SC)		3 factors (CFA; dietary restraint, shape/weight overvaluation, body dissatisfaction)	Convergent and discriminant validity supported.	Original factor structure was not supported.
Heiss et al., 2018, sample a		.85 (WC); .90 (SC)		CFA		Original factor structure was not supported. Four other structures tested and not any was supported.
Heiss et al., 2018, sample b		.85 (WC); .91 (SC)		CFA		Original factor structure was not supported. Four other structures tested and not any was supported.
Hilbert et al., 2012, sample a		.80 (WC); .90 (SC)			Females scored significantly higher than males.	PCA resulted in a 3-factor solution based on all EDE-Q items.

Hilbert et al., 2012, sample b	.72 (WC); .86 (SC)			Males scored significantly lower than females.	PCA resulted in a 3-factor solution based on all EDE-Q items.
Hrabosky et al., 2008	.61 (WC); .78 (SC); .95 (shape/weight overvaluation); .83 (appearance concern)		4 factors (EFA/CFA of all EDE-Q; eating disturbance, appearance concern, dietary restraint, shape/weight overvaluation)	Convergent validity supported.	Original factor structure was not supported.
Isomaa et al., 2016, sample a	.89 (WC); .95 (SC)			Females scored significantly higher than males. Eating disorder patient group scored significantly higher than adolescent group.	

Isomaa et al., 2016, sample b	.81 (WC); .89 (SC)			Females scored significantly higher than males. Eating disorder patient group scored significantly higher than adult group.	
Isomaa et al., 2016, sample c	.69 (WC); .82 (SC)			Eating disorder patient group scored significantly higher than adolescent and adult groups.	
Luce & Crowther, 1999	Time 1: .89 (WC); .93 (SC). Time 2: .89 (WC); .92 (SC).	*r* = .92 (WC); *r* = .94 (SC)			Time interval for test-retest: 2 weeks
Lydecker et al., 2016, sample a	.51 (WC); .71 (SC)				Alphas reported for combined Black and White sample.
					Criterion validity assessed and supported.

Lydecker et al., 2016, sample b		.51 (WC); .71 (SC)		Alphas reported for combined Black and White sample. Criterion validity assessed and supported.
Machado et al., 2014, sample a		School: .80 (WC); .90 (SC). University: .84 (WC); .93 (SC).		
Machado et al., 2014, sample b				PCA resulted in a 3-factor solution based on all EDE-Q items. Criterion validity assessed and supported.
Machado et al., 2018, sample a	Brief 3-factor	.90 (shape/weight overvaluation); .90 (body dissatisfaction)	3 factors (CFA)	Original factor structure was not supported.

Machado et al., 2018, sample b	Brief 3-factor	.91 (shape/weight overvaluation); .89 (body dissatisfaction)		3 factors (CFA)		Original factor structure was not supported. 3-factor structure invariant across eating disorder group and control group, and across different eating disorder diagnoses groups Criterion validity assessed and supported.
Mahmoodi et al., 2016		.58 (WC); .81 (SC)			Convergent and discriminant validity supported.	
Mitsui et al., 2017, sample a		.91 (fear of obesity); .82 (self-esteem based on shape and weight)		2 (EFA; fear of obesity, self-esteem based on shape and weight)		Original factor structure was not supported.
Mitsui et al., 2017, sample b					Convergent validity supported for fear of obesity and self-esteem based on shape and weight.	

Mitsui et al., 2017, sample c				Convergent validity supported for fear of obesity and self-esteem based on shape and weight.	
Mitsui et al., 2017, sample d				Convergent validity supported for fear of obesity and self-esteem based on shape and weight.	
Mond et al., 2004a	.83 (SC)	*r* = .73 (WC); *r* = .75 (SC)			Time interval for test-retest: 303.2 (*SD* = 57.4) days
Mond et al., 2004b					Criterion validity assessed and supported.
Parker et al., 2015	.98 (shape/weight overvaluation); .91 (appearance concern)		4 factors (EFA of all EDE-Q; dietary restraint, eating concern, shape/weight overvaluation, appearance concern)	Convergent and discriminant validity supported.	Original factor structure was not supported.

Parker et al., 2016	.56 (WC); .71 (SC); .93 (shape/weight overvaluation); .80 (appearance concern)		4 factors (EFA of all EDE-Q; dietary restraint, eating concern, shape/weight overvaluation, appearance concern)	Convergent validity supported.	Original factor structure was not supported. Criterion validity assessed and partly supported.
Peláez-Fernández et al., 2012	.74 (WC); .93 (SC)			Convergent validity supported.	Criterion validity assessed and supported.
Penelo et al., 2012	.65 (WC); .87 (SC)		2 factors (CFA; WC and SC)	Convergent validity supported. Spanish males scored significantly lower than Spanish females and American males.	
Penelo et al., 2013	.92 (eating-shape-weight concern)	ICC = .88 (eating-shape-weight concern)	2 factors (CFA of all EDE-Q; restraint and eating-shape-weight concern)	Convergent validity supported.	Original factor structure was not supported. Two-factor structure invariant across gender and area of residence.
Peterson et al., 2007	.72 (WC); .83 (SC)		3 factors (EFA of all EDE-Q)		Original factor structure was not supported.

Phillips et al., 2018	.84 (WC); .90 (SC)		4 factors (EFA of all EDE-Q)		Original factor structure was not supported. WC and SC primarily loaded onto the same factor.
Pretorius et al., 2009					Criterion validity assessed and supported. Criterion validity was higher among bulimia nervosa patients than among EDNOS-BN*** participant
Reas et al., 2006		ρ = .71 (WC); ρ = .66 (SC)			Time interval for test-retest: 1-14 days
Reas et al., 2012	.67 (WC); .84 (SC)			Males scored significantly lower than females.	
Reilly et al., 2014	Females: .93; Males: .89 (combined WC/SC subscale)			Males scored significantly lower than females. No evidence of gender-related DIF****.	
Rø et al., 2010	.81 (WC); .90 (SC)	ρ = .86 (WC); ρ = .91(SC)			

Rose et al., 2013		Time 1: .82 (WC); .87 (SC). Time 2: .87 (WC), .92 (SC).	*r* ranged from .87 to .94 across groups (full sample, males, females) and subscales (WC, SC)			Time interval for test-retest: 1 week
Unikel Santoncini et al., 2018, sample a	Brief 3-factor	.86 (WC); .92 (SC)		3 factors (CFA)	Student group scored significantly lower than eating disorder group.	The brief 3-factor model was supported
Unikel Santoncini et al., 2018, sample b	Brief 3-factor	.82 (WC); .91 (SC)		3 factors (CFA)	Eating disorder group scored significantly higher than student group.	The brief 3-factor model was supported
Villarroel et al., 2011		.83 (WC); .92 (SC)		2 factors (CFA; WC and SC)	Convergent validity supported.	
White et al., 2014, sample a				4 factors (CFA of all EDE-Q)	Females scored significantly higher than males.	Original factor structure was not supported.
White et al., 2014, sample b				1 factor (EFA; combined WC/SC)	Females scored significantly higher than males.	

Yucel et al., 2011	.78 (WC); .86 (SC)	*r* = .89 (WC); *r* = .89 (SC)	Convergent validity supported.

*CFA = confirmatory factor analysis, EFA = exploratory factor analysis, PCA = principal component analysis; **Higher scores indicate higher weight and shape concerns; *** EDNOS-BN = Eating disorders not otherwise specified – bulimia nervosa; ****DIF=differential item functioning

Appendix 15

Body dissatisfaction subscale (BD) of the Eating Disorder Inventory 3 (EDI-3): Sample characteristics

Sample	Country	Language	Setting	*N* (female)	Age: Mean (*SD*)	Other characteristics
Belon et al., 2015	USA	English	University	688 (100%)	20.4 (3.5)	56 % Hispanic, 44 % Caucasian
Clausen et al., 2011, sample a	Denmark	Danish	Medical	561 (100%)	24.8 (5.7)	Patients with eating disorders
Clausen et al., 2011, sample b	Denmark	Danish	General population	878 (100%)	25.8 (3.6)	
Cordero et al., 2013	USA	English	University	248 (100%)	20.3 (4.5)	Latina
Dadgostar et al., 2017	Iran	Persian	University	452 (66%)	Males: 22.31 (3.30); females: 22.43 (4.41)	
Elosua & Hermosilla, 2013, sample a	Spain	Spanish	School, University	1616 (0%)	15.53 (1.26)	
Elosua & Hermosilla, 2013, sample b	Spain	Spanish	School, University	1429 (100%)	15.42 (1.23)	

Elosua & López-Jáuregui, 2012	Spain	Spanish	Medical	394 (100%)	20.8 (6.61)	Patients with eating disorders
Kashubeck-West et al., 2013	USA	English	University	278 (100%)	29.04 (9.35)	African American
Lehmann et al., 2013	Netherlands	Dutch	Medical	514 (98%)	25.3 (7.2)/25.7 (6.6)	Patients with eating disorders
Nyman-Carlsson et al., 2015, sample a	Sweden	Swedish	Medical	292 (100%)	20.6 (2.23)	Patients with eating disorders
Nyman-Carlsson et al., 2015, sample b	Sweden	Swedish	Medical	140 (100%)	20.6 (2.23)	Psychiatric outpatients
Nyman-Carlsson et al., 2015, sample c	Sweden	Swedish	General population	648 (100%)	19.8 (4.53)	
Rothstein et al., 2017, sample a	USA	English	General population	197 (100%)	27.30 (9.82)	European American
Rothstein et al., 2017, sample b	USA	English	General population	104 (100%)	29.03 (11.37)	African American
Stein et al., 2015	USA	English	University	477 (100%)	19.8 (2.4)	Mexican American

Body dissatisfaction subscale (BD) of the Eating Disorder Inventory-3 (EDI-3): Measurement properties by sample

Sample	Internal consistency (α)	Reliability/test -retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Belon et al., 2015	.91		1 factor (CFA*)		Factor structure not invariant across Hispanic and Caucasian participants.
Clausen et al., 2011, sample a	.90		1 factor (CFA)	Discriminant validity supported.	Criterion validity assessed and supported
Clausen et al., 2011, sample b	.93		1 factor (CFA)	Discriminant validity supported.	
Cordero et al., 2013	.87		1 factor (EFA)		Original factor structure was partly supported
Dadgostar et al., 2017	Females: .8; Males: .6	Females: ICC = .67; Males: ICC = .69			Time interval for test-retest: 2 weeks
					Content validity assessed and supported.

Elosua & Hermosilla, 2013, sample a	.80		2 factors (CFA; BD and method factor)		Original factor structure was not supported. Factor structure partial invariant across females and males.
Elosua & Hermosilla, 2013, sample b	.87		2 factors (CFA; BD and method factor)		Original factor structure was not supported. Factor structure partial invariant across females and males.
Elosua & López-Jáuregui, 2012	.92	*r* = .96	CFA's performed on EDI-3 composites (not subscales)		Time interval for test-retest: 15 days
Kashubeck-West et al., 2013	.88 (Stomach size: .87; Thighs/Hips/Butt: .87)		2 factors (EFA; Stomach size and Thighs/Hips/Butt)	Convergent and discriminant validity supported.	Original factor structure was not supported.
Lehmann et al., 2013	.88		1 factor (CFA)		Criterion validity assessed and supported.

Nyman-Carlsson et al., 2015, sample a	.91		Eating disorder patients scored significantly higher than psychiatric outpatients and controls. Swedish eating disorder patients scored overall lower than Danish and international clinical samples.**	Criterion validity assessed and supported.
Nyman-Carlsson et al., 2015, sample b	.93		Psychiatric outpatients scored significantly higher than controls, and significantly lower than eating disorder patients.	
Nyman-Carlsson et al., 2015, sample c	.92		Controls scored significantly lower than eating disorder patients and psychiatric outpatients. Swedish controls scored significantly higher than Danish controls.	

Rothstein et al., 2017, sample a	.95 (for entire Eating Disorder Risk composite)		1 factor (CFA)		
Rothstein et al., 2017, sample b	.89 (for entire Eating Disorder Risk composite)		2 factors (EFA; body satisfaction, body dissatisfaction)	Convergent and discriminant validity supported for body satisfaction and body dissatisfaction factors	Original factor structure was not supported
Stein et al., 2015			2 factors (CFA; "overall body shape and stomach", "hips, thighs and buttock")	Convergent validity supported.	Criterion validity assessed and partly supported for "overall body shape and stomach" and "hips, thighs and buttock".

*CFA = confirmatory factor analysis, EFA = exploratory factor analysis, PCA = principal component analysis; **Higher scores indicate higher body dissatisfaction

Appendix 16

Appearance Evaluation (AE) and Body Areas Satisfaction Scale (BASS) of the Multidimensional Body Relations Questionnaire (MBSRQ): Sample characteristics

Sample	Country	Language	Setting	*N* (female)	Age: Mean (*SD*)	Other characteristics
Argyrides & Kkeli, 2013	Greece (Cyprus)	Greek	School	1312 (65%)	16.1 (.89)	Adolescents
Brytek-Matera & Rogoz, 2015	Poland	Polish	University	341 (100%)	23.23 (3.27)	
Cruzat-Mandich et al., 2019	Chile	Chilean Spanish	School, University	451 (56%)	19.57 (2.57)	
Kashubeck-West et al., 2013	USA	English	University	278 (100%)	29.04 (9.35)	African American
Kelly et al., 2012, sample a	USA	English	University	1467 (100%)	19.7 (3.8)	White
Kelly et al., 2012, sample b	USA	English	University	741 (100%)	19.7 (3.8)	Black
Marco et al., 2017	Spain	Spanish	School	355 (53%)	13.15 (.84)	Early adolescents
Naqvi & Kamal, 2017, sample a	Pakistan	Urdu/English	University	200 (50%)	19.55 (1.41)	

Naqvi & Kamal, 2017, sample b	Pakistan	Urdu	University	350 (61%)	19.12 (1.86)	
Naqvi & Kamal, 2017, sample c	Pakistan	Urdu	University	500 (55%)	17.19 (3.45)	
Nevill et al., 2015	UK	English	University	99 (57%)	20.4 (3.1)	94% Caucasian
Roncero et al., 2015	Spain	Spanish	School, general population	1041 (67%)	22.23 (3.07)	
Rusticus & Hubley, 2006	Canada	English	General population	1262 (67%)	39.7 (19.1)	75% White
Sabiston et al., 2010, sample a	Canada	English	General population	469 (100%)	57.1 (7.9)	Breast cancer survivors
Sabiston et al., 2010, sample b	Canada	English	General population	385 (100%)	55.4 (13.5)	
Smith & Davenport, 2012	USA	English	University	85 (100%)	20.33 (1.29)	Hispanic
Thoma et al., 2005	Canada	English	Medical setting	49 (100 %)	38 (range 20 - 68)	Patients waiting for reduction mammoplasty
Untas et al., 2009	France	French	University, medical setting	765 (76%)	Females: 33.3 (13.4); males: 31 (13.3)	

Vossbeck-Elsebusch et al., 2014	Germany	German	University, medical setting, general population	523 (100%)	26.43 (6.65)	44% diagnosed with eating disorder

Appearance Evaluation (AE) and Body Areas Satisfaction Scale (BASS) of the Multidimensional Body Relations Questionnaire (MBSRQ): Measurement properties by sample

Sample	Internal consistency (α)	Reliability/test-retest (ICC, *r*, *ρ*)	Structural validity	Hypotheses testing	Additional information
Argyrides & Kkeli, 2013	.82 (AE); .86 (BASS)	*r* = .87 (AE); *r* = .75 (BASS)	1 factor for AE (EFA*)	Convergent validity supported. Males scored significantly higher than females. Underweight participants scored significantly higher than normal weight participants. Normal weight participants scored significantly higher than overweight participants.**	Time interval for test-retest: 1 month BASS not included in factor analysis Criterion validity assessed.
Brytek-Matera & Rogoz, 2015	McDonald's *ω* = .91 (combined AE and BASS)		1 factor including both AE and BASS (EFA)		Original factor structure was not supported
Cruzat-Mandich et al., 2019	From .70 to .92. Factor "Evaluation of appearance" = .91		7 new factors based on full MBSRQ (EFA)		Original factor structure was not supported

Kashubeck-West et al., 2013	.79 (AE); .85 (BASS)		3 new factors based on full MBSRQ-Appearance Scales (EFA)	Convergent and discriminant validity supported	Original factor structure was not supported
Kelly et al., 2012, sample a	.90 (AE)		1 factor for AE (CFA)	Convergent validity supported for AE	BASS not included in analyses. AE was invariant across White and Black sample.
Kelly et al., 2012, sample b	.88 (AE)		1 factor for AE (CFA)	Convergent validity supported for AE.	BASS not included in analyses. AE was invariant across White and Black sample
Marco et al., 2017	.84 (AE); .84 (BASS)		1 factor for AE and 1 factor for BASS (CFA)	Convergent validity supported in females. Males scored significantly higher than females.	

Naqvi & Kamal, 2017, sample a		Urdu-Urdu: *r* = .89; Urdu-English: *r* = .85; English-Urdu: *r* = .82; English-English: *r* = .80		Time interval for test-retest: 15 days
Naqvi & Kamal, 2017, sample b	.75 (AE); .80 (BASS)		1 factor for AE and 1 factor for BASS (EFA)	Two items excluded for AE
Naqvi & Kamal, 2017, sample c			1 factor for AE and 1 factor for BASS (CFA)	Two items excluded for AE
Nevill et al., 2015		Non-parametric approach. BASS showed reasonable stability, AE did not.		Time interval for test-retest: 2 weeks

Roncero et al., 2015	.87 (AE); .78 (BASS)	1 factor for AE and 1 factor for BASS (CFA)	Convergent validity supported. Males scored significantly higher than females. Middle adolescences scored significantly higher than other age groups.	
Rusticus & Hubley, 2006		1 factor for AE and 1 factor for BASS (CFA)	Young adult women scored significantly higher on AE than older adult women did.	Invariance of the factor structure was not supported across gender and age group.
Sabiston et al., 2010, sample a	.85 (AE); .77 (BASS)	1 factor for AE and 1 factor for BASS (CFA)	Breast cancer survivors scored significantly higher on AE than controls.	Invariance of the factor structure was supported for AE across breast cancer survivors and controls (not supported for BASS).
Sabiston et al., 2010, sample b	.88 (AE); .82 (BASS)	1 factor for AE and 1 factor for BASS (CFA)	Breast cancer survivors scored significantly higher on AE than controls.	Invariance of the factor structure was supported for AE across breast cancer survivors and controls (not supported for BASS).

Smith & Davenport, 2012	.88 (AE); .73 (BASS)			Convergent validity supported for BASS. Participants scored significantly lower on AE than participants in a previously published study.	
Thoma et al., 2005		ICC = .85 (total MBSRQ-AS)		Convergent validity supported for total MBSRQ-AS***	Time interval for test-retest: 1 week Responsiveness assessed and supported for total MBSRQ-AS
Untas et al., 2009	.88 (AE); .66 (BASS)	*r* = .80 (AE); *r* = .86 (BASS)	1 factor for AE (EFA)	Convergent validity supported. Males scored significantly higher than females. Participants with lower BMI scored significantly higher than participants with higher BMI.	Time interval for test-retest: 1 month BASS not included in the factor analysis
Vossbeck-Elsebusch et al., 2014	.90 (AE); .85 (BASS)	*r* = .75 (AE); *r* = .79 (BASS)	1 factor for AE and 1 factor for BASS (CFA)	Convergent and discriminant validity supported. Eating disorder patients scored significantly lower than controls. No significant difference between different eating disorder groups.	Time interval for test-retest: 6 weeks

*CFA = confirmatory factor analysis, EFA = exploratory factor analysis, PCA = principal component analysis **Higher scores indicate higher body satisfaction ***MBSRQ-AS = MBSRQ Appearance Scales (including AE and BASS)