Running head: TEST ADAPTATION

Translation and Validation of Body Image Instruments: Challenges, Good Practice Guidelines, and Reporting Recommendations for Test Adaptation

Viren Swami1-2 & David Barron2

1School of Psychology and Sports Sciences, Anglia Ruskin University, Cambridge, United Kingdom

2Centre for Psychological Medicine, Perdana University, Serdang, Malaysia

Address for correspondence: Prof. Viren Swami, School of Psychology and Sports Sciences, Anglia Ruskin University, East Road, Cambridge, Cambridgeshire CB1 1PT, United Kingdom. Email: viren.swami@anglia.ac.uk.

**Abstract**

Body image research has grown rapidly to include new cultural and linguistic populations, but this gives rise to a need for measurement instruments that are sensitive to local contextual variations while remaining equivalent across groups. Test adaptation, or the translation and validation of a source instrument for use in a new cultural group, is an important part of this process. Here, we offer an operational framework for conducting effective test adaptation. We cover good-practice guidelines for instrument translation and suggest effective strategies for achieving semantic equivalence of translated instruments. We also focus on measurement invariance and provide good-practice and reporting guidelines for conducting exploratory and confirmatory factor analyses. Finally, we suggest good-practice guidelines for demonstrating that scores on translated measures have good reliability and validity. It is our hope that the availability of this article will assist body image scholars seeking to conduct robust test adaptations of existing measurement tools.

**Keywords:** Test adaptation; Translation; Exploratory factor analysis; Confirmatory factor analysis; Equivalence

**Introduction**

When *Body Image* was launched more than a decade ago, Cash (2004, p. 2) hoped the journal would become a home for “ideas and evidence about the psychosocial significance of embodied human experience.” An important plank for this mission was the consideration and publication of articles dealing with “physical appearance and body image in diverse cultural contexts” (Cash, 2004, p. 3). In the period since the launch of *Body Image*, the sub-disciplinary study of physical appearance and body image from cultural and cross-cultural perspectives has experienced rapid growth and now represents a thriving intellectual enterprise in its own right (e.g., see Anderson-Fye, 2009, 2011; Swami, 2015, 2018a; Swami & Barron, 2017a). A useful indicator of these developments is the significant increase in the cross-cultural diversity of published authors and data samples reported in *Body Image* between 2004 and 2015 (Swami & Barron, 2017b, cited in Cash, 2017).

While such trends are undoubtedly important, they also highlight a number of concerns inherent in body image research and testing practices. Perhaps the most important of these concerns relates to the issue of cross-cultural *test adaptation*, or the translation and validation of a source instrument1 for use in a new cultural group that is different from the one in which it was originally developed. Research on different cultural populations requires instruments that are sensitive to local contextual variations while remaining equivalent across groups. To put it differently, when a known body image instrument has been translated and used in a new language, readers need to be certain that the translated measure is equivalent to the original measure and that the instrument is capturing the same constructs across different groups. Ensuring that this is the case is not easy and a cursory glance at test adaptation papers that have been published in this journal suggests that the quality of the processes used for test adaptation procedures, and the reporting thereof, varies widely.

In this article, we provide an overview of suggested good practice when conducting test adaptation studies. The first part of the paper focuses on methods used to ensure semantic equivalence through translational techniques, while the second part of the paper focuses on issues of measurement equivalence. Where appropriate, we highlight challenges faced by body image scholars working on test adaptation and provide reporting recommendations for future studies. Importantly, while our focus is on test adaptation for establishing cross-cultural equivalence, many of the issues raised here will also be of interest to scholars more broadly (e.g., scholars interested in scale construction and factor analytic methods). Further, while our primary focus is on test adaptation across cultural and national groups, many of the issues highlighted here should also be of interest to researchers interested in test adaptation across other social identity groups (e.g., across gender, ethnic/racial, or age groups within the same population). The overall aim of this paper is to provide body image scholars with clear guidelines for achieving equivalence in test adaptation studies.

**Choosing an Appropriate Instrument**

When conducting research in new populations, body image scholars generally have three options for choosing an appropriate study instrument (Cha, Kim, & Erlen, 2007; He & van de Vijver, 2012; Khalaila, 2013): (a) develop a new instrument; (b) use an instrument that was originally developed for the cultural or linguistic group under investigation; or (c) use an instrument that was previously developed in a different cultural or linguistic population than the target population. The first of these options is sometimes preferable, particularly where the development of the measure occurs in distinct national and linguistic groups at the same time (Solano-Flores, Trumbull, & Nelson-Barber, 2002; Sprangers et al., 1993). In this scenario, scholars generate a common set of items that are conceptually relevant for different cultural populations (the parallel approach) or develop culture-specific items at the same time in different cultures (the simultaneous approach) (Anderson, Aaronson, Bullinger, & McBee, 1996). However, these approaches are rarely utilised in body image research, possibly because it requires a great deal of effort, time, and cross-cultural knowledge. Scholars seeking good-practice guidelines and empirical-based rules for the parallel or simultaneous approaches are encouraged to consult the seminal volume *Cross-Cultural Research Methods* (Brislin, Lonner, & Thorndike, 1973).

 A second approach is to use instruments that were originally developed for the particular cultural or linguistic group under investigation (the *emic approach*; Brislin et al., 1973). For example, a scholar wishing to operationalise body dissatisfaction in a particular cultural group may use a measure that has previously been developed for use with that cultural group, with item-content relevant to that singular culture. However, scholars seeking to adopt this approach may experience two related difficulties. The first is that the emic approach typically suffers from a lack of generalisability: although it provides a thorough understanding of concepts relevant to one culture, these concepts are rarely comparable to other cultural groups (Canino, Lewis-Fernández, & Bravo, 1997). The second difficulty is that emic-focused measures are often poorly constructed, with little evidence of adequate psychometric properties (Khalaila, 2013). In the context of Malaysia, for example, Swami (2018b) reported on a range of measures of attitudinal body image that are widely used but lack indices of validity and reliability.

 The third, and most common, approach utilised by body image scholars is the *sequential approach* (Anderson et al., 1996), where a measure that is available in an original source language (often English) is translated into another language. This approach is sometimes preferred because it can save time and effort (Cha et al., 2007), but the decision to use a measure developed in a different cultural context raises difficulties of its own. For example, one common assumption in body image research is that culture has minimal impact on the construct being measured (the *absolutist approach*; Herdman, Fox-Rushby, & Badia, 1997); that is, it is assumed that the way a construct is defined and measured in one culture can be directly applied to another culture. Sometimes, this leads scholars to create literal translations that are linguistically identical to the original measure, with insufficient attention paid to the translation process and reporting thereof. On other occasions, it can lead scholars toward assuming that the meaning of scores on a measure will be identical across groups and that said scores can be compared across distinct groups. For example, Swami (2018a) cites a number of cases where researchers have used translated versions of the Body Appreciation Scale (BAS; Avalos, Tylka, & Wood-Barcalow, 2005) in new cultural groups without adequate attention paid to the test adaptation process. This is problematic because scores on a test are only interpretable in light of evidence that the meaning and dimensionality of the construct being measured, as well as the items comprising the measurement instrument, and are equivalent across groups (van de Vijver & Leung, 1997).

This hints at a different perspective, namely the *universalist approach* (Herdman et al., 1997), which does not assume *a priori* that a particular construct will be the same in all cultural groups. To return to the example of body appreciation, it is possible that the construct exists in some form across all cultural groups, but culture may play a role in variations of expression (Swami & Chamorro-Premuzic, 2008). In this view, it is assumed that an instrument developed in one cultural context to measure the construct, such as the BAS, will need to go through a rigorous process of culture-specific test adaptation. The goal of test adaptation in this case is to strive for equivalence in terms of keeping the translated version as similar as possible to the original version, while aiming for a conceptually equivalent version that measures the same constructs and in which constructs retain the same meaning (van Widenfelt, Treffers, de Beurs, Siebelink, & Koudijs, 2005; Ziegler & Bensch, 2013). This, in turn, requires a culturally sensitive approach to test adaptation (García Coll, Akerman, & Cicchetti, 2000), which can be promoted through rigorous test adaptation procedures that help to ensure equivalence (Hambleton, Merenda, & Spielberger, 2005; Iliescu, 2017; Ziegler & Bensch, 2013).2

The issue of equivalence has been widely discussed in cross-cultural psychology (e.g., Poortinga, 1995; van de Vijver & Leung, 2000; van de Vijver & Tanzer, 1997; Ægisdóttir, Gerstein, & Çinarbas, 2008) and a full discussion is beyond the scope of this paper. However, based on the universalist approach, it is possible to identify five broad types of equivalence that scholars should pay attention to (Herdman, Fox-Rushby, & Badia, 1998). The first of these relates to *conceptual equivalence*, which refers to the extent to which the construct tapped by the original measure is relevant and pertinent to the new cultural context for which the measure is being adapted (Herdman et al., 1998). Conceptual equivalence is closely related to *item equivalence*, which refers to the utility of individual items in a measure for tapping the construct being evaluated in the new context (Herdman et al., 1998). Conceptual and item equivalence should ideally be determined through a thorough reading of the available literature (including publications in the culture of the original measure and the target population, as well as other adaptation studies involving the measure), discussions between subject specialists, and discussions with representatives of the target population (Reichenheim & Moraes, 2007). The latter, in particular, is an invaluable source of *a priori* judgement-driven information on item and conceptual equivalence (van de Vijver, 2011) and can be achieved through various qualitative methods (e.g., cognitive interviews; see Collins, 2003; Willis, 2005) that provide informal tests of the suitability of instruments (He & van de Vijver, 2012; van de Vijver & Tanzer, 2004). The goal in this sense is to determine whether it is feasible to construct a version of the instrument in the target language and culture and to determine whether the instrument would be too biased following adaptation. When bias is deemed to be a limiting issue, researchers should either adapt only parts of the instrument or develop a new instrument (van de Vijver & Hambleton, 1996).

*Semantic equivalence* refers to the extent to which the meaning of concepts tapped in the original measure are retained in translated versions of the measure. This form of equivalence is most directly relevant to the translation process itself, which is discussed in more detail below (see Ensuring Semantic Equivalence). *Operational equivalence* refers to the characteristics of using an instrument in the target and source populations (e.g., the layout and format of the instrument, the application setting, the use of appropriate response categories, and translation of instructions to participants). This is important because instructions can influence how a measure is completed, whereas imprecise translation of response categories can result in skewed responses (Guillemin, Bombardier, & Beaton, 1991; Hambleton & Patsula, 1999; He & van de Vijver, 2012; van de Vijver & Leung, 2011). The latter in particular can have an important influence of the response characteristics of items (van de Vijver & Poortinga, 1997; van Widenfelt et al., 2005). A fifth form of equivalence is broadly referred to as *measurement equivalence*, which is determined through an investigation of the psychometric properties of scores derived from the translated measure. Given the centrality of measurement invariance in the test adaptation process, this issue is discussed in detail later in this article (see Measurement Equivalence).

**Ensuring Semantic Equivalence**

 The process of translation is an important part of test adaptation generally and ensuring semantic equivalence specifically (Geisinger, 1994; Gudmunsson, 2009; Sechrest, Fay, & Zaidi, 1972). In body image research, including papers published in this journal (e.g., Swami & Chamorro-Premuzic, 2008), Brislin’s (1970) back-translation is both popular and widely used. This method uses an iterative process of independent translation and back-translation by independent bilingual translators (Behling & Law, 2000; Triandis & Brislin, 1984; Yu, Lee, & Woo, 2004). A bilingual translator blindly translates (i.e., forward-translates) the source measure (along with instructions and response categories) to the target language; a second bilingual translator independently back-translates the instrument from the target language to the original language. Next, the two versions of the instrument (i.e., the original language and back-translated versions) are compared for (conceptual, item, semantic, and operational) equivalence and, where discrepancies are found, another translator attempts to retranslate the problematic item(s). This process should continue until a team of bilingual translators agree that the two versions of the instrument are identical in conceptual meaning.

 However, there is increasing recognition that exclusive use of back-translation methods may not be sufficient and can result in poor translations (Hambleton et al., 2005; Perneger, Leplège, & Etter, 1999; van de Vijver & Tanzer, 1997). One major limitation of the method is that researchers cannot estimate *a priori* how many independent bilingual translators will be needed to obtain equivalence between the source and translated versions of the measure. In practice, many body image scholars rely on two bilingual translators, who resolve any discrepancies between translations through consensus (e.g., Swami, Barron, Lau, & Jaafar, 2016; Vossbeck-Elsebusch et al., 2014) or discussion (e.g., Alleva, Martijn, Veldhuis, & Tylka, 2016; Rodgers et al., 2016). The difficulty with this approach is that it may amplify problems in the translated measures when the original and target languages have different vocabularies, grammar or syntax, or the use of idioms, or when objects and experiences that are familiar to members of one cultural group are alien to members of another (Hui & Triandis, 1985; Sechrest et al., 1972). Related problems occur when the forward- and back-translations are conducted by translators who are not sufficiently knowledgeable about the requirements of cross-cultural translations or the specific content of the instrument being translated (Hambleton & de Jong, 2003; Sperber, Devellis, & Boehlecke, 1994; for practical solutions, see Tsai, Luck, Jefferies, & Wilkes, 2018) and when forward- and back-translators share demographic backgrounds that do not adequately represent the target population (Wang, Lee & Fetzer, 2006).

 In his seminal work, Brislin (1970) offered three other techniques for ensuring the quality of a translation. The first is the bilingual technique in which the instrument is administered in both the original and target languages to bilingual participants. The aim of the bilingual technique is to detect items that yield discrepant responses across the two versions of the instrument. Where discrepancies are detected, researchers should consider potential reasons for these and seek ways to resolve the discrepancies. One difficulty with the bilingual technique is that it requires that respondents are sufficiently knowledgeable about, and fluent in, the target language. However, this in itself may give rise to discrepant responses compared to monolinguals (McDermott & Palchanes, 1992). For example, discrepant responses on the original and target languages of a measure may arise because bilinguals are more acculturated to the culture in which the measure was originally developed (Sperber et al., 1994).

 A second method suggested by Brislin (1970) is the committee approach, in which a group of bilingual individuals forward-translate the original measure to a target language. The main benefit of this method is that it is more effective at detecting mistakes made by one translator and at revealing inconsistencies (Munet-Vilaró & Egan, 1990), although it does require more than three bilingual translators (Cha et al., 2007). The final method used for translation is the pre-test procedure, which involves pilot-testing a translated measure in study conditions that mirror the final study and in a population that resembles the sample that will ultimately be recruited. In this procedure, the pilot group is probed about understanding of randomly-selected items from a scale (e.g., “What does this item mean?”), and responses that are unfitting are scrutinised (Brislin et al., 1973). This method may also be used to identify words, items, or sentences that are grammatically awkward or unusual in the target language (Philips, de Hernandez, & de Ardon, 1994); more sophisticated pilot testing can involve cognitive interviewing3 (Iliescu, 2017; Lee, 2014). Alternatively, the pilot group is asked to rate each translated item for understanding on a predetermined scale (Brislin et al., 1973) and some scholars recommend that pre-testing continues until a pre-established percentage of understanding (e.g., ≥ 90%) is achieved for all items (Reichenheim & Moraes, 2007).

 There is now general consensus that each of the methods above, when used in isolation, does not produce true equivalence (Maneesriwongul & Dixon, 2004). Instead, scholars – including the International Test Commission’s (ITC) Guidelines for Test Adaptation (Hambleton et al., 2005) – recommend combined translation techniques that include back-translation alongside additional techniques proposed by Brislin (1970). One framework that body image scholars may find useful is the five-stage test adaptation procedure recommended by Beaton, Bombardier, Guillemin, and Ferraz (2000). In brief, this involves: (a) an informed and an uninformed bilingual translator (whose first language is the target language) independently forward-translating the measure, instructions, and response categories from the original to the target language; (b) the resolution of any discrepancies between the two translations through written reports from the two translators and a recording observer, resulting in the production of a synthesised translation through a consensus approach; (c) the back-translation of the synthesised measure by two independent bilingual translators whose first language is the source language and who are naïve to the measure; (d) a review and comparison of the forward- and back-translations by an expert committee comprising methodologists, language professionals, and the translators (and, where feasible, the original scale developers), with a view to producing a pre-final version of the measure; and (e) pre-testing with a minimal sample of 30-40 respondents (whose first, or only, language is the target language), who are probed for their understanding, acceptability, and emotional impact of the items. Following pre-test, final semantic adjustments are made so that the translated measure can then be used in testing.

**Recommendations for Body Image Scholars**

It is difficult to prescribe a set of guidelines aimed at producing an ideal translation, because recommended procedures vary widely (Epstein, Santo, & Guillemin, 2015) and because there is no good empirical data to suggest that one method will suit all practicalities (van Widenfelt et al., 2005). However, one general recommendation that is consistent with current best-practice is to avoid using back-translation as the sole translational method (Brislin et al., 1973; Hambleton et al., 2005). In particular, there appears to be little value in using the back-translation method merely to demonstrate to the original author or to potential reviewers that a literal translation was conducted. Indeed, the notion that absolute measurement equivalence is possible is debatable (Erkut, Alarcón, García Coll, Tropp, & Vásquex García, 1999) and researchers who rigidly apply back-translation methods are likely to produce poor translations (van Widenfelt et al., 2005).

Instead, best practice indicates that a combination of translation procedures should be adopted: while Beaton and colleagues (2000) provide a framework that some body image scholars will find useful (for applied examples in the body image literature, see Argyrides, Kkeli, Kendeou, 2014; González-Martí, Bustos, Contreras, & Mayville, 2012; Swami et al., 2011) it is also time-demanding and effortful. Some scholars may find useful alternative, and less taxing, combined translation techniques. For example, some researchers have introduced a combined approach that uses both the back-translation and bilingual technique (Jones, Lee, Phillips, Zhang, & Jaceldo, 2001), whereas others have used a combination of the back-translation, committee, and pre-test procedures (Cha et al., 2007). More broadly, body image scholars should consult the ITC’s Guidelines for Test Adaptation (Hambleton et al., 2005), alongside seminal works in cross-cultural methodology (Brislin et al., 1973), choose the most appropriate combination of methods for their particular cultural context, and explicate the methods used in their article. More detailed procedures in preparing semantically equivalent translations are described elsewhere (e.g., Beaton et al., 2000; Behling & Law, 2000; Cha et al., 2007; Eremenco, Cella, & Arnold, 2005; Gudmunsson, 2009; Guillemin et al., 1991; Hambleton & Zenisky, 2011; Herdman et al., 1998; Iliescu, 2017; Jones et al., 2001; Krach, McCreery, & Guerard, 2016; Maneesriwongul & Dixon, 2004; Perneger et al., 1999; Sperber, 2004; Sumathipala & Murray, 2006; van Widenfelt et al., 2005) and interested readers are advised to consult these sources prior to conducting test adaptation work. Nevertheless, some specific recommendations for body image scholars are worth highlighting.

**Cultural appropriateness**. One aspect of conceptual equivalence that is particularly pertinent to body image research is *cultural appropriateness*, which refers to the degree to which a measure (or items of a measure) are deemed inoffensive, relevant, and normative for the target population (Iliescu, 2017). An example of poor cultural appropriateness is the use of figural rating scales that depict White morphological traits in non-White populations (see Pulvers et al., 2012), a limitation that could be dealt with through the use of standardised figural rating scales (e.g., Swami, Salem, Furnham, & Tovée, 2008). Even where standardised figural rating scale are used, care should be taken to ensure that figures are not perceived as offensive or in contravention of local norms, such as in the depiction of clothing or nudity. The importance of ascertaining cultural appropriateness is not only limited to figural rating scales but is also vital when considering attitudinal statements. For example, certain behaviours or concepts (e.g., smiling) may vary in meaning across cultures and may, therefore, not be suitable for cross-cultural comparisons (Lonner, 1985; Malpass & Poortinga, 1986).

**Vocabulary equivalence**. Vocabulary equivalence can be compromised when a word does not exist in the target language or when a particular word has multiple or different meanings in the target language. Becker, Gilman, and Burwell (2013) provide an example of difficulty translating a weight-related attitudes measure from English to Nadroga because of vocabulary differences. In such cases, the problem can usually be solved by using a comparable word or a group of words in place of the original formulation (Sechrest et al., 1972). In many cases, it will be possible to find a word (or words) that is similarly referential or denotative in meaning to the original, thus allowing for a literal translation. Where this is not possible, it may still be possible to adapt an item so that it retains its general or connotative meaning (Herdman et al., 1998). In doing so, however, researchers should pay careful attention to issues that transcend the literalness of words, such as their emotional or affective meaning in the target culture (Harkness, 1998; Reichenheim & Moraes, 2007).

**Idiomatic equivalence**. Some body image measures use idioms that are not easily translated. To take one example, Item #8 of the Body Appreciation Scale-2 includes the exemplar “I walk holding my head high” (Tylka & Wood-Barcalow, 2015), which in English means that one is confident and proud. To be able to effectively translate this item, translators must first be aware that it is in fact an idiomatic expression; where there is a lack of familiarity with the English language, it may result in a literal translation that lacks meaning in the target language. This speaks to the importance of including translators whose native language and culture are those of the place for which the translation is being done, but who are also fluent in the source language (Guillemin et al., 1993; Perneger et al., 1999). Sechrest and colleagues (1972) provide recommendations for dealing with idioms and maintaining idiomatic equivalence, the most direct of which is the removal of idiomatic expressions in translations.

**Experiential equivalence**. A violation of experiential equivalence occurs when two cultures differ markedly in the nature of their understanding of objects or overall way of life, or when objects and experiences that are familiar to members of one culture are alien to another (Sechrest et al., 1972). One example from the Drive for Muscularity Scale (DMS; McCreary & Sasse, 2000) serves to highlight this issue. This relates to Item #9 (“I think that I would look better if I gained 10 pounds in bulk”) which includes an imperial weight measurement that may not be commonly used in some national contexts. Although this may seem trivial, retaining the imperial measurement in national contexts where metric measurements are the norm would serve little purpose and would detract from experiential equivalence. The most straightforward way to deal with this issue is revise the statement to refer to a metric equivalent (“I think that I would look better if I gained 5 kilogrammes in bulk”; Swami et al., 2016; Swami, Vintila, Tudorel, Goian, & Barron, 2018; see also Tao & Zhong, 2010), thus ensuring experiential equivalence.

**Other adaptation issues**. Occasionally, researchers may be tempted to generate new items for an existing measure because of a perceived omission of relevant content or constructs. It is important to note, however, that the addition of novel items may alter the conceptual equivalence of the translated measure. In most cases, it is desirable for translated measures to remain conceptually similar to the original version, both in terms of item content and the number of items. Maintaining this similarity will also help to ensure that test scores are comparable across studies. Where it is deemed that new items are essential, these should be designed in collaboration with the original scale author(s) where possible. Finally, deleting part of an item or entire items should be avoided. Although some deletions may be well-intentioned, doing so may compromise the quality of information generated by the scale in the target culture (Nunnally & Berstein, 1995). Where deletion is deemed necessary, it should normally be based on psychometric evaluations of scores on the translated measure (see Measurement Equivalence below).

**Recommended reporting guidelines**. Where feasible, authors should report their translational procedures in detail either in the manuscript itself or in supplementary materials. Authors may sometimes be hesitant about doing so, believing that primacy should be given instead to data analyses (i.e., measurement invariance) or erroneously perceiving that drawing attention to adaptations to a source measure mean that the translation is compromised in some way. Authors should be encouraged to view the translation process as just important as measurement issues, which in turn means that it should be given due coverage. In particular, authors should be encouraged to report in full all test translation methods (if necessary, as supplementary materials; Greiff & Iliescu, 2017) and highlight where translational difficulties were encountered, where item alterations were made, and how translational discrepancies or issues were resolved. Rigorous translation procedures, and the reporting thereof, will help readers be confident that a translated measure retains its conceptual, item, semantic, and operational equivalence in the target culture.

**Measurement Invariance**

 Even where the translation process has been rigorous, many adaptation problems may still go undetected until the measure is field-tested (Byrne & van de Vijver, 2010; Ercikan & Lyons-Thomas, 2013; Hambleton & Patsula, 1999). It is, therefore, vital that researchers conduct empirical analyses with a dataset generated from a translated instrument, with a view to establishing measurement equivalence (Schmitt & Kuljanin, 2008). In this sense, a thorough evaluation of psychometric properties of scores derived from a translated instrument should ideally follow best-practice guidelines for any novel measure (Reichenheim & Moraes, 2007). More specifically, it is useful to divide the psychometric evaluation process into two stages that are focused on: (a) examining the latent dimensionality or factor structure of scores on the translated measure; and (b) evaluating the reliability and validity of scores on the translated measure. General guidelines for achieving rigorous psychometric evaluations and reporting recommendations for body image scholars are provided below.

**Latent Dimensionality**

 Examining the latent dimensionality of scores on a newly-translated measure is an essential first step in the psychometric evaluation process. Occasionally, researchers may conduct a rigorous translation of a source measure, but then assume that scores derived from the translation should be scored in an identical manner to the parent study conducted in a different culture. That is, they adopt an absolutist approach in which culture has minimal or no impact on the constructs that are being measured and in which the dimensionality of scores does not change (Herdman et al., 1998). Likewise, researchers sometimes assume that, because scores on a translated measure are internally consistent (as determined, more often than not, by Cronbach’s coefficient alpha), scores on the measure *ipso facto* retain its parent factor structure. For example, Swami (2018a) cites several studies where the one-factor structure of translated versions of the BAS has been assumed because the one-dimensional score demonstrated adequate internal consistency.

 Such practices are inherently problematic from a universalist approach, which suggests that scholars should not assume *a priori* that latent dimensionality of scores in a new cultural context will mirror that of the source measure. It is again useful to examine the BAS in view of the universalist approach: while studies using English versions of the scale in populations from the United States have consistently documented a one-factor solution (Avalos et al., 2005; Tylka, 2013), studies in some other cultural contexts using translated versions of the BAS have reported that scores reduce to two distinct dimensions (for a review, see Swami, 2018a). Moreover, studies within the same linguistic group have sometimes reported difficulty confirming earlier-reported factor structures of scores on the BAS (e.g., Bakalim & Tasdelen-Karçkay, 2016; Ferreira, Neves, & Tavares, 2014). This brief overview of translated versions of the BAS highlights the ways in which social and cultural identities can impact on score dimensionality (Swami, 2018a; Swami & Chamorro-Premuzic, 2008) and why, from a practical point-of-view, making *a priori* assumptions about the dimensionality of BAS scores should be avoided. Indeed, best-practice guidelines recommend that the factor structure of scores on any given measure are (re-)examined any time the measure is used in a new context or with different population groups (American Education Research Association, American Psychological Associations, & National Council on Measurement in Education, 2014).

Once the decision has been made to examine the factor structure of scores on a translated measure, researchers are faced with having to choose between a number of statistical methods that could be applied. Here, best-practice guidelines suggest that researchers should begin their investigation of the dimensionality of scores using exploratory factor analysis (EFA; Comrey & Lee, 1992; Costello & Osborne, 2005; Pett, Lackey, & Sullivan, 2003). This remains true where a particular factor structure has been proposed in the source study or where previous test adaptation studies have reported on factor structure(s) in other cultural or social identity groups (Worthington & Whittaker, 2006). One of the benefits of beginning the empirical analysis of scores on a translated measure using EFA is that it allows researchers to explore the latent dimensionality of scores without any *a priori* limitations in terms of modelling (Costello & Osborne, 2005; Gorsuch, 1983; Kline, 1994; Rummel, 1988). That is, EFA allows researchers to effectively assess the underlying factor structure in a dataset and to determine the best factorial solution for that dataset.

Examining the dimensionality of scores from a translated measure using EFA is only a first step to establishing measurement invariance (Henson & Roberts, 2006). Ideally, an EFA should be followed-up with a second step utilising confirmatory factor analysis (CFA) on another sample. Not all scholars recommend using both EFA and CFA (e.g., Kline [2005] recommended using either EFA or CFA, but not necessarily both), but there is emerging consensus that a two-step analytic strategy consisting of EFA followed by CFA on a different sample offers the most robust approach to test adaptation and validation (Worthington & Whittaker, 2006). An EFA-to-CFA strategy is effective at exploring the extent to which the source factor structure exists in a new population, but also allows for an in-depth exploration of item behaviour with regards to multiple hypothesised models (Loehlin, 2004; Maruyama, 1998). When using this two-step strategy, it is important to conduct EFA and CFA on separate, adequately-sized samples; that is, EFA and CFA should not be performed on the same sample (Fokkema & Greiff, 2017).

Researchers may sometimes be tempted to progress immediately to CFA in the absence of exploratory analysis of data derived from a translated scale, but this should be avoided (Worthington & Whittaker, 2006). In CFA, researchers must first develop clear, theory-driven hypotheses about what factors are underlying scores on the measures used and impose constraints on the model based on their *a priori* hypotheses. Because of these constraints, CFA is very useful for testing whether data fit *a priori* hypothesised measurements, but less useful at detecting whether alternative models that were not hypothesised fit the data better (Thompson, 2004). That is, CFA is too restrictive to be used in an exploratory fashion, which is the case in translations of measures. In terms of test adaptation, CFA is best used after an EFA (Browne, 2001; Schmitt, 2011). A related limitation of CFA is the requirement of zero loadings (i.e., for indicators not supposed to load on a certain factor), which may be overly strict and unrealistic (Marsh, Lüdtke, Nagengast, Morin, & von Davier, 2013; Marsh, Morin, Parker, & Kaur, 2014; Marsh et al., 2009) and can be contrasted with EFA where all loadings are free to vary.

One alternative to CFA is exploratory structural equation modelling (ESEM), which specifies hypotheses about the relation between observed indicators and their primary latent factors while allowing for estimation of loadings with other latent factors as well (Asparouhov & Muthén, 2009). ESEM estimates of factor correlations are typically more accurate than CFA estimates (Marsh et al., 2014) and, although ESEM is primarily a confirmatory tool, it can be used with care as an exploratory analytic method in a manner that has many advantages over EFA and CFA (Marsh et al., 2009, 2013). ESEM is increasingly used by body image researchers (for examples in the body image literature, see Morin & Maïano, 2011; Sánchez-Carracedo et al., 2012), but its use is perhaps most appropriate where clear *a priori* hypothesised models can be specified or where a translational model is being revisited (Marsh et al., 2013). Regardless of whether a researcher decides to use EFA, CFA, or ESEM, the rationale for doing should always be justified on empirical grounds and all analyses should be clearly reported. Given that most researchers working on test adaptations are likely to use an EFA-to-CFA approach, some general reporting guidelines for EFA and CFA are provided below.

**Conducting and Reporting Exploratory and Confirmatory Factor Analyses**

For both EFA and CFA, the following reporting recommendations should be considered:

1. Report the name and version of the statistical software and packages used for all analyses.
2. Report the amount of missing data, the extent to which missing data were missing at random or missing completely at random (Little, 1988; Little & Rubin, 1987), and how missing data were handled. Problems of missing data are often magnified in CFA, which makes missing data imputation especially important in these cases. Best-practice guidelines for handling missing data are provided by Parent (2013) and, in most cases, researchers should consider using multiple imputation (see Schafer & Olsen, 1998) or model estimation with full-information maximum likelihood (Rose & Fraser, 2008; Sterne et al., 2009).
3. Researchers should pay careful attention to sampling and carefully consider the extent to which method non-equivalence is an issue in test adaptations (van der Vijver & Leung, 2011). For example, *sample bias* can be caused by a lack of comparability in sample characteristics between the source and target samples (e.g., when the parent study validated a measure in a community sample and the adaptation is validated in a college sample). Relatedly, where a sample is split to accommodate an EFA-to-CFA strategy, researchers should report how the data were split and test for subsample differences in terms of key demographics. Where different sampling strategies were used to recruit participants for EFA and CFA, respectively, conduct standard analyses to test for group differences in key demographics. Ideally, the two (sub-)samples used for EFA and CFA, respectively, should not differ significantly in terms of key demographics.

For EFA specifically:

1. Consider whether an adequately-sized sample has been recruited. Various recommendations have been proposed to determine an adequate sample for EFA based either on minimal sample size (e.g., Comrey & Lee, 1992; Gorsuch, 1983; Kline, 1979) or participant-to-item ratios (e.g., see Arrindell & van der Ende, 1985; Nunnally & Bernstein, 1995). The most commonly-used guideline is that researchers should ensure a participant-to-item ratio of 10:1 (Nunnally & Berstein, 1995; Tabachnick & Fidell, 2013), though a ratio of 20:1 should be aimed for (Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005). However, these guidelines have been criticised for failing to consider item communality, the degree of overestimation of factors, and the size of loadings (MacCallum, Widaman, Zhang, & Hong, 1999; Preacher & MacCallum, 2002). Instead, it is recommended that researchers recruit as a large sample as practical because sample adequacy cannot be determined until after the data have been analysed (Henson & Roberts, 2006). Worthington and Whittaker (2006) suggest that, if item communalities are ≥ .50 or there are 10:1 items per factor with factor loadings of about .40, then a sample size of 150-200 may be adequate; if communalities are ≥ .60 or there is a minimum of 4:1 items per factor with factor loadings above .60, then smaller samples may be adequate.
2. If an adequate sample size has been recruited, the next decision to be made relates to the factor extraction method, which can be broadly distinguished between exploratory (or common) factor analysis and principal components analysis (PCA). Researchers sometimes mistake EFA and PCA as the same extraction technique. However, while both sets of analyses have similarities (e.g., both are variable reduction techniques and share common requirements for use; Child, 1990) and may sometimes provide similar results (e.g., where communalities are large; Preacher & MacCallum, 2003; Truxillo, 2003), they have different purposes. PCA is a statistical method that reduces the number of observed variables to a smaller number of “principal components” that account for most of the variance of the observed variables. EFA, on the other hand, identifies the latent constructs and the underlying factor structure of a set of variables. Furthermore, PCA tends to overestimate factor loadings relative to EFA (Snook & Gorsuch, 1989) and underestimate the strength of correlations between factors (Widaman, 1993). For these reasons, it is almost always appropriate to begin an analysis of the dimensionality of scores on a translated measure using EFA rather than PCA (Fabrigar & Wegener, 2012; Pett et al., 2003). Principal-axis factoring is the most commonly-used fitting procedure to estimate the factor loadings and unique variances of a model, especially as it can be used when the assumption of normality is violated (Fabrigar, Wegener, MacCallum, & Strahan, 1999). An alternative to principal-axis factoring is maximum likelihood estimation (Jöreskog, 1977), which provides estimations of standard errors for factor loadings and factor correlations (Cudek & O’Dell, 1994), as well as an χ2 statistic that can be used to evaluate fit of the extracted factor solution (Fabrigar & Wegener, 2012; Tabachnick & Fidell, 2013). Maximum likelihood estimation requires that data meet assumptions of multivariate normality and is infrequently used in the body image literature, so the points below are focused on principal-axis EFA.4
3. Researchers should report whether their data meet assumptions for EFA based on item distribution, average correlation with other items, and item-total correlations. Clark and Watson (1995) advocate average inter-item correlations of .15-.50 across constructs and .40-.50 for narrowly-defined constructs, although some scholars recommend more conservative average communalities (e.g., inter-item communalities of ≥ .50; Bearden & Netemeyer, 1999; MacCallum et al., 1999). In addition, researchers should report the degree to which their data are *factorable*, that is, the degree to which there are at least some correlations among variables so that coherent factors can be identified. Factorability can be assessed through inter-item correlations, the Kaiser-Meyer-Olkin measure of sampling adequacy (which should be ≥ .60 and ideally ≥ .80; Kaiser, 1974) and Bartlett’s test of sphericity (which should be significant).
4. Researchers should report and justify the rotation method that was used in EFA. Orthogonal rotation methods (e.g., quartimax, varimax) assume that the factors in the EFA are uncorrelated, whereas oblique rotation methods (e.g., direct oblimin, promax) assume that the factors are correlated (Gorsuch, 1983). In most cases, the decision to use an orthogonal or oblique rotation should be based on the source study. Where this is not possible (e.g., is a source paper has not reported the rotation method that was used), it may be helpful to begin with an oblique rotation to assess factor inter-correlations (which should be reported), before deciding on a final rotation solution (Henson & Roberts, 2006; Tabachnick & Fidell, 2013; Worthington & Whittaker, 2006). Tabacknick and Fidell (2013) recommend that, if factor correlations ≥ .32, then there is enough variance overlap to warrant a final oblique rotation. Whatever rotation a researcher decides upon, it should be fully justified (Fabrigar et al., 1999; Pedhazur & Schmelkin, 1990).
5. Authors should clearly describe the criteria used to retain factors. Two commonly-used methods are the Kaiser or mineigen greater than 1 criterion (K1), which retains factors with eigenvalues greater than 1 (Kaiser, 1970), and Cattell’s (1966) scree test, which involves examination of a plot of the eigenvalues for breaks or discontinuities. However, both of these methods have been shown to result in factor over-retention (Horn, 1965; Hayton, Allen, & Scarpello, 2004). In addition, the K1 criterion has been argued to be arbitrary (Fabrigar et al., 1999), while the ambiguity and subjectivity of the scree test may cause low inter-rater reliabilities (Crawford & Koopman, 1979). Conversely, parallel analysis (Horn, 1965) is one of the most accurate methods for determining the number of factors to retain (Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). Parallel analysis involves the construction of correlation matrices of random variables based on the same sample size and number of variables in the real dataset. Factors in the real dataset should only be retained if their eigenvalues are greater the eigenvalues from the random data; factors in the real dataset with eigenvalues less than or equal to the parallel average random eigenvalues are considered to be due to sampling error and should be discarded (Glorfeld, 1995; Fabrigar et al., 1999; Montanelli & Humphreys, 1976; Zwick & Velicer, 1986). Hayton and colleagues (2004) provide a step-by-step guide for conducting parallel analysis. However, it should be noted that there are other factor-retention methods that could be applied, including Velicer’s (1976) minimum average partial method (which calculates the average of squatted partial correlations after each component is partialled out; no further factors are extracted when the minimum average squared partial correlation is reached) and Bartlett’s (1950) chi-square test (which tests the hypothesis that remaining eigenvalues are equal; factors are retained only when the test returns a significant result). Also worth highlighting is Tabachnick and Fidell’s (2013) recommendation of excluding any factors with fewer than three items and Guadagnoli and Velicer’s (1988) recommendation of excluding any factor that explains less than 5% of the variance, as these factors are likely to be uninterpretable.
6. Once factors have been retained, authors should describe item-retention methods for each factor. Some scholars recommend retaining only those items that have minimally “fair” factor loadings (i.e., ≥ .33) and deleting items that cross-load with values ≥ .33 on at least two factors (Comrey & Lee, 1992; Tabachnick & Fidell, 2013). On the other hand, Guadognoli and Velicer (1988) suggested that factors are only interpretable if they have four or more loadings of at least .60; Stevens (1992) suggests using a more liberal cut-off of .40. Hair, Black, Babin, and Anderson (2010) provide an alternative recommendation for assessing the practical significance of standardised factor loadings as denoted by the fact matrix (in a one-dimensional model) or the pattern matrix (in a correlated, multiple-factor model) based on sample size. Whichever method is used should be justified and it is also good reporting practice to provide the whole factor pattern (or structure matrix), including all items in the analysis (Henson & Roberts, 2006) in the original and translated language.
7. It is possible to examine the extent to which two or more groups (e.g., across gender, culture) share the same factor dimensionality for scores on a measure using EFA, although this method does not appear to have been utilised by body image scholars. The degree of factor similarity is assessed using Tucker’s (1951) congruence coefficient and is usually computed after one of the factor loading matrices has been transformed to fit another loading matrix in the least squares sense by a Procrustes rotation (see Lorenzo-Seva & ten Berge, 2006). Various thresholds have been proposed for Tucker’s congruence coefficient (e.g., Bentler & Bonett, 1980; Horn, Wanberg, & Appel, 1973; Mulaik, 1972), but simulations by Lorenzo-Seva and ten Berge (2006) suggest that values between .85 and .94 correspond to fair similarity across groups, whereas values ≥ .95 suggest that factor structures can be considered equal across groups.

Erring on the side of over-reporting EFAs is a useful recommendation (Henson & Robert, 2006), particularly as these analyses are iterative and subjective. Independent researchers should be able to evaluate the analytic decisions of authors in test adaptation studies in order to be able to replicate those findings in a new study or to employ CFA. For CFA specifically, the following recommendations should be considered:

1. Authors should carefully specify the hypothesised models that are tested using CFA. This should include the model identified through earlier an EFA and the source model in the parent study if the latter is different. It is also useful to examine the fit of other competing models that have been proposed in the translational literature and consider whether all proposed paths are theoretically sound or whether there are paths that should be included based on relevant theory (Jackson, Gillaspy, & Purc-Stephenson, 2009; Nunkoo, Ramkissoon, & Gursov, 2013). However, there should be clear justifications on theoretical and statistical grounds for doing so. Taking the example of the DMS, for instance, it is possible to identify various three-factor models of DMS scores that have been proposed in the translational literature, but these have either been found to have poor confirmatory fit to data or consist of factors with less-than-adequate internal consistencies (for a recent review, see Swami et al, 2018). In these cases, there may be no clear justification for testing the fit of these hypothesised models in new test adaptation studies of the DMS. For each model that is tested, it is important to provide a full description of the *a priori* parameter specification (fixed parameters, free parameters, and constrained parameters). Including a figure of each hypothesised model being tested is often useful and various authors have provided graphical conventions that should be adhered to (Kline, 2005; Ullman, 2006a).
2. Consider whether an adequately-sized sample has been recruited. Common recommendations for determining an adequate *N* for CFA include a minimum sample size of 100 or 200 (Boomsma, 1982, 1985; Jackson, 2001) and 5 to 20 observations per estimated parameter (Bentler & Chou, 1987; Bollen, 1989; Tanaka, 1987). However, these recommendations have been shown to have limited practical value (Marsh, Hau, Balla, & Grayson, 1998; Myers, Ahn, & Jin, 2011) because model characteristics (e.g., size of the model, distribution of variables, strength of the relationships among variables, the amount of missing data) can all affect accuracy of parameter estimates and model fit statistics (MacCallum et al., 1999; Wolf, Harrington, Clark, & Miller, 2013). Most scholars now agree that there is no single guideline that will be applicable to all datasets (Myers et al., 2011; Wolf et al., 2013). Three improved approaches to evaluating sample size requirements for CFA are: (a) the Satorra and Satis (1985) method, which estimates power based on the amount of model misspecification; (b) the MacCallum, Browne, and Sugawara (1996) method, which is based on the power of the model to obtain a root mean square error of approximation (RMSEA) value that is consistent with good model fit; and (c) the Monte Carlo simulation method (Muthén & Asparouhov, 2002; Muthén & Muthén, 2002). A useful strategy for researchers working on test adaptation is to use a proactive Monte Carlo approach (Marcoulides & Chin, 2013; Marcoulides & Saunders, 2006), which involves conducting simulation studies of a hypothesised model with relationships among the variables specified based on the results of an earlier EFA or on the available research literature. Where this is not possible, researchers should consider a reactive Monte Carlo approach (Marcoulides, 1990; Marcoulides & Chin, 2013), which involves analysis existing data after a study has been completed to evaluate the hypothesised model(s).
3. Authors should assess whether data used for CFA are within critical normality limits using both univariate and multivariate indices. Univariate distributions are typically assessed through skewness and kurtosis, and Weston and Gore (2006) recommended that skewness should be < |3| and kurtosis < |10|. Multivariate distributions are typically evaluated using Mardia’s (1970) coefficients (which assesses multivariate kurtosis and skewness), with normalised coefficients > |5| indicative of non-normality (Byrne, 2010). Where univariate or multivariate non-normality is detected, it is advisable to use an estimation method that addresses non-normality (see point 14).
4. Authors should describe how an appropriate estimation technique was selected. When data are univariate and multivariate normal, the maximum likelihood and generalised least squares estimation methods are the most widely-recommended (Ullman, 2006a) and are likely to be the most suitable for body image scholars working with medium-to-large datasets (*N* ≥ 120; see Bentler & Yuan, 1999; Hu, Bentler, & Kano, 1992). The asymptotically distribution free method has no distributional assumptions but should be avoided in the absence of very large samples (*N* ≥ 2500; Fouladi, 2001; Ullman, 2006a). When data are univariate or multivariate non-normal, adjustments for non-normality should be applied (Bentler, 1990, 2001; Satorra, 2001); the Satorra-Bentler scaled χ2 is a correction to the χ2 that is often used along with robust standard errors for parameter estimates and robust goodness-of-fit indices (Satorra & Bentler, 2001), but other corrections are available (Bentler & Dijkstra, 1985).
5. Once a model has been estimated, the fit of the model to data should be assessed. The clearest index for making judgements on the acceptability of model fit is the χ2 (with a nonsignificant result indicative of good fit; Ropovik, 2015), but this statistic is sensitive to sample size (larger samples produce a significant result even when discrepancies between implied and obtained covariance matrices are small) and the size of factor loadings, and thus increases the risk of Type II error (Bentler & Bonnet, 1980; Hancock & Mueller, 2011; McNeish, An, & Hancock, 2018; Saris, Satorra, & van der Veld, 2009). As such, scholars recommend routinely reporting χ2 along with other absolute fit indices (e.g., the χ2/df ratio or normed chi-squared, the standardised root mean square residual [SRMR], the goodness-of-fit index [GFI], the absolute goodness-of-fit index [AGFI]), relative fit indices (e.g., the Tucker-Lewis Index [TLI], the Bentler-Bonett Normed Fit Index [NFI], Bollen’s Incremental Fit Index [IFI or BL89]), or non-centrality parameters (e.g., RMSEA and its confidence intervals, Bentler’s Comparative Fit Index [CFI], the McDonald-Marsh Relative Noncentrality Index [RNI]) (Crowley & Fan, 1997; Jackson et al., 2009; McDonald & Ringo Ho, 2002; Rapovik, 2015; Ullman, 2006b). There are no clear guidelines about which fit indices to present: Hu and Bentler (1998, 1999) recommended a two-index format (consisting of SRMR along with TLI, RMSEA, CFI, or BL89), whereas Kline (2005) advocated reporting the χ2, RMSEA, CFI, and SRMR. BL89 is relatively unaffected by sample size and is worth reporting alongside other indices (Bollen, 1990; Hu & Bentler, 1995, 1998). Note also that models with small degrees-of-freedom and low *N* can have artificially large RMSEA values, so this statistic should not be relied upon in these cases (Kenny, Kaniskan, & McCoach, 2015; though some scholars disagree; see Sivo, Fan, Witta, & Willse, 2006). Where competing hypothesised models are tested, it may also be useful to report a measure of comparative fit, such as Akaike’s Information Criterion (AIC), and a parsimony-corrected fit index (e.g., the Parsimony Goodness-of-Fit Index [PGFI], Parsimony Normed Fit Index [PNFI]) (Hooper, Couglan, & Mullen, 2008). There is some debate as to whether parsimony adjustments are useful (Marsh & Hau, 1996), but when used in combination with other fit indices, they assist researchers in making decisions about models *vis-á-vis* parsimony, while not penalising models for having more parameters (Mulaik et al., 1989). It is bad statistical practice to select fit indices that are favourable to one’s hypotheses, while discounting those that indicate poor fit (Jackson et al., 2009).
6. It is important to assess model fit based on *a priori* values, which should be reported in full (Jackson et al., 2009). Hu and Bentler (1999) proposed fit values for all commonly-used indices assuming maximum-likelihood estimation; these include values ≤ 3.00 for χ²/df, close to .08 for SRMR, and close to .95 for CFI, GFI, TLI, BL89, and RNI, and close to .90 for AGFI as indicative of good fit. Recommendations for RMSEA vary, but generally values close to .06 are considered to be indicative of good fit and values of about .07-.08 indicative of adequate fit (Hu & Bentler, 1999; Kline, 2015; Steiger, 2007). No thresholds have been recommended for PGFI and PNFI, but Mulaik and colleagues (1989) suggested that PGFI values should be in the region of .50-.90. It should be noted, however, that Hu and Bentler (1999) explicitly warned against over-generalisation of these indices to all scenarios. Indeed, there is considerable controversy about using these values as cut-offs or “golden rules” to judge model fit (Fan & Sivo, 2005, 2007; Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; Hogarty, Hines, Kromrey, Ferron, & Mumford, 2007; Hopwood & Donnellan, 2009; Marsh, Hau, & Wen, 2004; Perry, Nicholls, Clough, & Crust, 2015), primarily because goodness-of-fit (as indicated by fit indices) are influenced by a range of “nuisance” factors (e.g., sample size, size of factor loadings, the number of items per latent factor; Chen, Curran, Bollen, Kirby, & Paxton, 2008; Fan & Sivo, 2007; Garrido, Abad, & Ponsoda, 2016; Greiff & Heene, 2017; Heene, Hilbert, Freudenthaler, & Bühner, 2012; McNeish et al., 2018; Saris et al., 2009). Some have called for the abandoning of fit indices altogether (e.g., Barrett, 2007), though the broad consensus is that fit indices retain their value. Best-practice is to report multiple fit indices and use the Hu and Bentler (1999) values as a guide for assessing model fit, while acknowledging that these values should not be interpreted rigidly (Markland, 2007; Marsh et al., 2004). Where alternative (e.g., MacCallum et al., 1996) or contextualised fit indices (Hair, Black, Babin, & Anderson, 2010) are used, these should be justified. It is also recommended that authors keep up-to-date with ongoing discussions over the use of fit indices.
7. Sometimes, modifications may be made to a hypothesised model to improve fit. It should be noted that, when this is conducted, the analysis changes from being confirmatory to exploratory in nature and, as a result, any conclusions drawn from the modified model should be presented with caution (Ullman, 2006a). One common method for improving model fit in test adaptation studies is to examine modification indices to locate potential areas of misspecification and model improvement (Schumacker & Lomax, 2004). For example, it is sometimes reasonable to estimate a model that allows for covariances between the variance in each item that is not already accounted for a given construct (Ullman, 2006a). Modification indices values ≥ 3.84 usually have a statistically significant effect on the model’s χ2, but some authors recommend a more conservative convention of using values ≥ 5.00 (Byrne, Shavelson, & Muthén, 1989). When improving model fit, modifications should be made to a model based on assumed correlations among like items from the same factor (i.e., model improvement should be based on theoretical and substantive meaning, such as items that include similar words, phrases, or meanings, and should not be arbitrary; Shumacker & Lomax, 2004; Whittaker, 2012). In addition, to ensure that improved fit of a model is not due to chance alone, modifications should be kept to a minimum (MacCallum, Roznowski, & Necowitz, 1992; Steenkamp & Baumgartner, 1998). If extensive modifications are necessary to achieve good fit, researchers should consider evaluating the revised model in a cross-validation sample, as modifications may be unstable across different samples of respondents (e.g., they may be based on unique response patterns that are not generalisable). Once modifications have been made, a chi-square difference test (or the Satorra-Bentler scaled chi-square difference test when data are non-normal) should be used to determine whether modifications improved model, with a significant result indicating improvement (Brown, 2015). However, the chi-square difference test is sensitive to sample size and improvements may be difficult to detect when sample sizes are small (Ullman, 2006a). In these cases, an alternative test – such as the Lagrange multiplier or Wald test (Ullman, 2006b) – may be more appropriate. When reporting factor loadings and other standardised parameter estimates for a final, fitted model, researchers should also report the unstandardised estimates and standard errors (Brown, 2015; Jackson et al., 2009).
8. Just because a model has been demonstrated to have good fit does not mean that researchers should conduct between-group tests of scale means. Any comparison of means (e.g., between women and men, or between different cultural groups) presupposes that the measure functions similarly across groups (Byrne et al., 1989; Cheung & Rensvold, 1999, 2002; Iliescu, 2017; Millsap, 2011; Nagengast & Marsh, 2013; Vandenberg & Lance, 2000). To investigate the degree to which a measure is invariant across groups (i.e., that the response to individual items can be explained by the same latent factors), it is recommended that researchers first perform multi-group CFA (Chen, 2008). Multi-group invariance should, at a minimum, be examined consecutively at the configural, metric, and scalar levels (Chen, 2007; Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012; Marsh et al., 2009). *Configural invariance* implies that the number of latent variables and the pattern of loadings of latent variables on indicators are similar across groups (i.e., the unconstrained latent model should fit the data well in all groups; Horn & McArdle, 1992). If configural invariance is not supported, it suggests there are fundamental differences in the dimensionality of constructs across groups, which in turn means that between-group comparisons should be avoided (Marsh & Grayson, 1994). *Metric invariance* (also called *weak factorial invariance*; Widaman & Reise, 1997) implies that the magnitude of the loadings is similar across groups and is tested by comparing two nested models consisting of a baseline model and an invariance model (Horn & McArdle, 1992). The baseline model allows factor loadings to be freely estimated across groups, whereas the invariance model constrains factor loadings to be equivalent across groups. A nonsignificant chi-square difference test (Δ*χ*²) indicates that the invariance model is a better representation of the data (because it has better parsimony than the baseline model) and is indicative of metric invariance (Chen, 2007). Finally, *scalar invariance* (sometimes called *strong factorial invariance*; Meredith, 1993) implies that both the item loadings and item intercepts are similar across groups and is examined using the same nested-model comparison strategy as with metric invariance (Chen, 2007). However, one issue is that the Δ*χ*² statistic is sensitive to sample size and may be an overly stringent criterion of scalar invariance (Chen, Sousa, & West, 2005; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Simulation studies have suggested that ΔCFI may be better alternative, as it is independent of model complexity and less sensitive to sample size. A ΔCFI value of less than .01 indicates metric invariance (Cheung & Rensvold, 2002) and some scholars suggest that, where ΔCFI indicates invariance and *N* ≥ 200 but Δ*χ*² is significant, metric invariance is still supported as differences between groups are likely trivial (Meade et al., 2008). For scalar invariance, Chen (2007) suggested that invariance is supported when ΔCFI < .01 *and* ΔRMSEA < .015 *or* ΔSRMR < .030, although other scholars suggest that ΔCFI < .01 is sufficient (Cheung & Rensvold, 2002). General consensus among scholars indicates that latent factor means should not be compared in the absence of scalar invariance (Chen, 2007); that is, scalar invariance is a precondition of conducting between-group comparisons of mean scores on a measure.
9. Some scholars also require that *strict invariance* is achieved (Wu, Li, & Zumbo, 2007), which implies that invariance is met at the configural, metric, and scalar, as well as equality of variables and factor residual variance levels across groups (Meredith, 1993; Widaman & Reise, 1997). However, strict invariance is likely to be too stringent to hold in practice and most scholars agree that scalar invariance is sufficient in order to compare the means of latent variables across different groups (Chen, 2007, 2008; van de Schoot, Schmidt, & DeBeuckelaer, 2015; Vandenburg & Lance, 2000). In fact, scalar invariance may also be an unrealistic goal, particularly where the number of factors, groups, or indicators is large (Davidov, Meuleman, Cieciuch, Schmidt, & Billet, 2014; Zercher, Schmidt, Cieciuch, & Davidov, 2015; see also Svetina & Rutkowski, 2018). For this reason, some scholars allow for partial measurement invariance when true metric or scalar invariance is not achieved (Byrne et al., 1989). Partial measurement invariance involves the application of forward stepwise selection procedures to free constraints that are not equivalent between groups at any of the stages mentioned in our point 18. Vandenberg and Lance (2000) recommended that partial invariance should be examined only on a small proportion of indicators and relaxation of parameters should be based on strong theoretical and practical justifications. Comparison of relationships between the latent construct and other constructs can be conducted where partial metric invariance is supported, but comparison of means across groups should only be conducted when partial scalar invariance is found (Davidov et al., 2012). Where partial scalar invariance is not obtained, Vandenburg and Lance (2000, Figure 2) present a number of additional options that may be considered. However, the use of forward stepwise selection procedures to determine partial invariance has been strongly criticised (e.g., Harrell, 2001) and two alternatives to partial invariance testing are the alignment method (Asparouhov & Muthén, 2014; Marsh et al., 2018) and equivalence testing (Yuan & Chan, 2016; Yuan, Chan, Marcoulides, & Bentler, 2016), which interested readers may wish to consider.

For scholars using an EFA-to-CFA strategy, following the guidelines above should assist in determining the extent to which scores on translated measures are truly invariant across populations of interest. Establishing measurement invariance is an important step that should be accomplished before researchers compare latent scores on a particular construct across groups. However, there may be some situations where the steps above have been followed, yet researchers still fail to find measurement invariance across groups. In these cases, differential item functioning (DIF) may be useful to help identify whether the lack of measurement invariance is being caused by anomalous items. DIF indicates the degree to which an item is more discriminating, more difficult to endorse, or more extreme in a target group as compared to a reference group when the different groups are matched on the latent trait under investigation (Camilli & Shepard, 1994; Zumbo, 1999, 2003). From a psychometric point-of-view, the presence of item bias and DIF reduces the validity of a measure.

He and van de Vijver (2012) provide the following three-step recommendation for simple estimation of DIF, where multi-group invariance CFA suggests that scores on a measure for two or more groups lack invariance. First, a total score on a one-dimensional scale should be calculated (for multi-dimensional scales, these steps should be conducted with subscale scores), irrespective of the groups being investigated. Next, the total score is divided into multiple levels based on the range and, finally, an analysis of variance (ANOVA) is computed, with the target groups and score level as independent variables and item scores as the dependent variable. Item bias is likely when the ANOVA returns a significant effect of culture and a significant interaction between culture and score level. Where this is the case, a closer inspection of the items may highlight problems with the translation, lack of conceptual equivalence, or a possible cross-group difference that requires further attention. More sophisticated DIF methods are also available and should be applied where necessary (see Osterlind & Everson, 2009; Teresi, 2006; Teresi, Ramirez, Lai, & Silver, 2008). For an application of DIF in the body image literature, see Jewett and colleagues (2017).

**Assessing Reliability and Validity**

 Once factorial validity has been established through EFA and CFA, it is important to assess the reliability of factor scores.5 The most frequently used index of reliability is coefficient alpha (Zumbo & Rupp, 2004), with values of ≥ .80 considered adequate for basic research tools (Nunnally, 1976; see Lance, Butts, & Michels, 2006). However, recent work has suggested that, when using Likert-type scales – as many body image researchers do – coefficient alpha may be a negatively-biased estimate of reliability (Sijtsma, 2009). Instead, emerging consensus suggests that researchers should use and report ordinal coefficient alpha in factor analysis models (Zumbo, Gadermann, & Zeisser, 2007). Ordinal coefficient alpha has been found to be an adequate estimate of theoretical reliability, irrespective of the number of scale points and the skewness of scale point distributions (Zumbo et al., 2007). Two alternatives to ordinal coefficient alpha are the greatest lower bound (glb) value (the lowest possible value that a scale score’s reliability can have; Sijtsma, 2009) and omega (Dunn, Baguley, & Brunsden, 2014; Revelle & Zinbarg, 2009). It is recommended that, where possible, researchers report one of ordinal coefficient alpha, glb, or omega in place of coefficient alpha when using Likert-type scales. We recommend the use of the statistical software package *R* (*R* Development Core Team, 2011) for computing these estimations (see Gaderman, Guhn, & Zumbo, 2012, for a practical guide).

Where feasible, scholars should also assess test-retest reliability, which refers to reproducibility of scores on a measure over time in the same population (Aaronson et al., 2002). A number of different statistical methods exist to assess test-retest reliability. The most popular of these methods is to compute the correlation coefficient between scores across time (Pearson’s *r* or Spearman’s *rho* if the statistical assumptions required for *r* are not met; Rousson, Gasser, & Seifert, 2002). However, many scholars discourage the use of correlation coefficients to assess test-retest reliability because systematic error cannot be detected (Baumgartner, 2000; Bédard, Martin, Krueger, & Brazil, 2000). In addition, there are no clear guidelines for judging the minimum acceptable value for reproducibility estimates based on correlation coefficients (Crocker & Algina, 1986). Instead, Weir (2005) recommended reporting the intraclass correlation coefficient (Shrout & Fleiss, 1979), with values ranging from |0| to |1|. Higher ICCs are indicative of good test-retest reliability, although the use of universal ICC cut-offs should be avoided given that ICCs will vary depending on the version used and data variability (Charter & Feldt, 2001; Shrout, 1998). Computing paired-samples *t*-tests is also useful, with a nonsignificant result indicative of good test-retest reliability.

Finally, it is important to assess scores on translated measures for construct validity, which refers to the extent that scores on an instrument what it purports to be measuring (Cronbach & Meehl, 1955). Messick (1995, 1998) highlights different forms of construct validity, of which the most relevant to the present discussion is external validity. This refers to the extent that scores on a test have *convergent* (the extent to which scores on a test are related to scores on other theoretically-similar constructs), *concurrent* (the extent to which scores on a test are related to scores on other theoretically-similar constructs), *discriminant* (the extent to which scores on a test are unrelated to other theoretically unrelated constructs), and (strict) *predictive validity* (the extent to which scores on a test predict some criterion measure, where scores on the latter are measures at a different point in time to the test measure). Because of the difficulty measuring strict predictive validity, it is infrequently assessed in test adaptation studies. One alternative to assessing strict predictive validity is to assess incremental validity, which refers to the extent to which scores on a test predict some outcome measure above-and-beyond the variance accounted for by scores on theoretically-related constructs (for examples in the body image literature, see Alleva, Tylka & Kroon van Diest, 2017; Tylka & Iannantuono, 2016; Tylka & Wood-Barcalow, 2015).

A common source of error in test adaptation studies occurs when researchers pay insufficient attention to additional measures included in a questionnaire that are intended to provide evidence of construct validity. Consider, for example, a research seeking to establish measurement invariance for a translated version of a measure of positive body image. Care has been taken to translate the measure and establish its factorial validity, but to demonstrate construct validity the researcher includes additional measures that have not been validated for use in the target population. Sometimes this occurs because insufficient thought has gone into the design of a validation study; other times this may occur because there are few available measures for use in the target population. In either case, it important that researchers include measures to establish construct validity that have good psychometric properties in the target population. Where there are few such measures, it may be possible to concurrently validate two or more measures in the same study (e.g., see Swami et al., 2011).

Alternatively, researchers should consider other methods of establishing construct validity, such as the known-groups technique, which involves administering the target measure to groups that are theoretically expected to differ on latent scores due to known characteristics (Polit & Beck, 2012). Note also that it is possible to provide evidence of convergent and discriminant validity through CFA using the Fornell-Larcker criterion (Fornell & Larcker, 1981). Convergent validity is assessed by calculating the average variance extracted (AVE) for each factor of scale (values ≥ .50 are considered adequate), whereas discriminant validity is assessed by using the squared correlation (*r*2) between factors and values of AVEs (when the AVEs for each pair of correlated factors > *r*2, discriminant validity is supported) (Hair et al., 2010; Malhotra & Dash, 2011). An alternative to the Fornell-Larcker criterion that can be used to assess discriminant validity is the heterotrait-monotrait ratio (Henseler, Ringle, & Sarstedt, 2014).

**Conclusion**

 Body image research has rapidly expanded to include new cultural and linguistic populations, and scholars are increasingly interested in drawing conclusions about similarities or differences in test scores across such populations. Being able to do so, however, is heavily dependent on the quality of the data that is generated. An important precondition for generating that high-quality data are generated is the process of test adaptation, which helps to ensure that measures are sensitive to local contextual variations while remaining equivalent across groups. In many cases, however, body image scholars have not paid sufficient attention to issues of test adaptation or the deficiencies of using data derived from measures that have not been through a rigorous process of test adaptation. If the aim of body image scholars is to be able to claim with confidence that the measures they are using in new populations demonstrate reliable and valid scores, and if scholars have an interest in comparing results from different cultural and linguistic groups, then it is vital that greater consideration is given to issues of test adaptation.

**Footnotes**

1 We use the terms “instrument,” “test,” and “measure” interchangeably in this article. Additionally, we use the term “test adaptation” here, but it is elsewhere used interchangeably with “test localisation” or “test indigenisation.”

2 Not all scholars agree that test adaptation of existing measures is appropriate. Berry (1989, p. 721), for example, viewed this method as “imposed etics,” arguing that fails to capture culture-specific elements of a construct and wrongly assumes that a construct will, by definition, function similarly across groups. To overcome this limitation, some scholars have proposed that researchers begin with an existing instrument that has been developed in one culture and translate it for use in a target culture. To this translation, novel items should be included (developed by persons knowledgeable about the culture or drawn from relevant literature) that are reflective of, and unique to, the target culture (Brislin, 1976; see also Triandis, 1975, 1976). The novel items have the potential to provide greater depth of understanding of a construct in a target culture, but assessments of equivalence should only be conducted on items that are shared across groups (Brislin et al., 1973).

3 In cognitive interviewing, translated items are presented to members of the target population in order to assess their cognitive responses in terms of comprehension, retrieval/recall, estimation/judgement, and response (Drennan, 2003; Willis, 2005). A trained facilitator uses verbal probing techniques and guided “thinking aloud” to evaluate how members of the target population understand about questionnaire items (item-by-item or even word-by-word), what they elicit in terms of their recall of information, judgement processes, and responses processes. This technique provides maximum pretesting information and can be used to improve the quality of the translation (Lee, 2014).

4 Where a maximum likelihood extraction method is used, researchers should compute χ2, which represents how well the factor analytic model fits the observed data. A significant χ2 test indicates that the model does not fit the data perfectly and increasing the number of factors retained will result in lower χ2 values (i.e., better fit). Researchers could, therefore, extract solutions with an increasing number of factors retained until the χ2 test return a nonsignificant result (Fabrigar & Wegener, 2012). However, some authors caution against using this method because its null hypothesis (a perfectly-fitting model) is unrealistic and because the χ2 statistics is sensitive to sample size (Brown, 2015). Nevertheless, computing the χ2 statistic also facilitates computation of other model fit indices, which could be used to help make factor-extraction decisions and should be interpreted following *a priori* criteria (see our point 16).

5 It is important to note that instruments are not inherently reliable or valid (i.e., reliability and validity are not properties of the scale itself); rather, instruments yield reliable and valid scores with certain samples. Conceptualising instruments as inherently “reliable” and/or “valid”, and neglecting to consider the sample this evidence is based on, is bad practice that could lead researchers unfamiliar with concept to assume that a particular instrument can be used in a new sample without appropriate consideration of the issues raised in this article.

**References**

Aaronson, N., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D. L., Perrin, E., & Stein, R. E. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research*, *11*, 193-205. doi:10.1023/A:1015291021312.

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. doi: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 evaluation in U.S. community women and men. *Body Image*, *23*, 28-44. doi:10.1016/j.bodyim.2017.07.008

American Education Research Association, American Psychological Associations, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.

Anderson, R. T., Aaronson, N. K., Bullinger, M., & McBee, W. L. (1996). A review of the progress towards developing health-related quality-of-life instruments for international clinical studies and outcomes research. *PharmacoEconomics*, *10*, 336-355. doi:10.2165/00019053-199610040-00004

Anderson-Fye, E. (2009). Cross-cultural issues in body image among children and adolescents. In L. Smolak & J. K. Thompson (Eds.), *Body image, eating disorders, and obesity in youth: Assessment, prevention, and treatment* (2nd ed., pp. 113-133). Washington, DC: American Psychological Association.

Anderson-Fye, E. (2011). Body images in non-Western cultures. In T. F. Cash & L. Smolak (Eds.), *Body image: A handbook of science, practice, and prevention* (2nd ed., pp. 244-252). London: Guildford.

Argyrides, M., Kkeli, N., & Kendeou, P. (2014). Validation of the factor structure of the Greek adaptation of the Sociocultural Attitudes Towards Appearance Questionnaire (SATAQ-3). *Body Image*, *11*, 201-205. doi:10.1016/j.bodyim.2014.02.003

Arrindell, W. A., & van der Ende. J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, *9*, 165-178. doi:10.1177/014662168500900205

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397-438. doi:10.1080/10705510903008204

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, *21*, 495-508. doi:10.1080/10705511.2014.919210

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

Bakalim, O., & Tasleden-Karçkay, A. (2016). Body Appreciation Scale: Evaluation of the factor structure and psychometric properties among male and female Turkish university students. *Mersin Üniversitesi Egitim Fakültesi Dergisi*, *12*, 410-422. doi:10.17860/efd.38032

Barrett, P. (2007). Structural equation modelling: adjudging model fit. *Personality and Individual Differences*, *42*, 815-824. doi:10.1016/j.paid.2006.09.018

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology*, *3*, 77-85. doi:10.1111/j.2044-8317.1950.tb00285.x

Baumgartner, T. A. (2000). Estimating the stability reliability of a score. *Measurement in Physical Education and Exercise Science*, *4*, 175-178. doi:10.1207/S15327841Mpee0403\_3

Bearden, W. O., & Netemeyer, R. G. (1999). *Handbook of marketing scales: Multi-item measures for marketing and consumer behaviour research* (2nd ed.). Thousand Oaks, CA: Sage.

Beaton, D., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*, 3186-31919.

Becker, A. E., Gilman, S. E., & Burwell, R. A. (2013). Changes in prevalence of overweight and in body image among Fijian women between 1989 and 1998. *Obesity*, *13*, 110-117. doi:10.1038/oby.2005.14

Bédard, M., Martin, N. J., Krueger, P., & Brazil, K. (2000). Assessing and reproducibility of data obtained with instruments based on continuous measurements. *Experimental Aging Research*, *26*, 353-365. doi:10.1080/036107300750015741

Behling, O., & Law, K. O. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. Thousand Oaks, CA: Sage.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 256–259. doi:10.1037/0033-2909.107.2.238

Bentler, P. M. (2001). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606. doi:10.1037/0033-2909.88.3.588

Bentler, P. M., & Chou, C. H. (1987). Practical issues in structural modeling. *Sociological Methods and Research*, *16*, 78-117. doi:10.1177/0049124187016001004

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis 6* (pp. 9-42). Amsterdam: North Holland.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*, 181-197. doi:10.1207/S15327906Mb340203

Berry, J. W. (1989). Imposed etics-emics-derived etics: The operationalization of a compelling idea. *International Journal of Psychology*, *26*, 721-735. doi:10.1080/00207598908247841

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bollen, K.A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107*, 256-259. doi:10.1037/0033-2909.107.2.256

Bollen, K. A., & Long, J. S. (Eds.) (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Boomsma, A. (1982). Robustness of LISREL against small sample sizes in factor analysis models. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirection observation: Causality, structure, prediction* (Part I, pp. 149-173). Amsterdam: North Holland.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50*, 229-242. doi:10.1007/BF02294248

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185-216. doi:10.1177/135910457000100301

Brislin, R. W. (1976). Comparative research methodology: Cross cultural studies. *International Journal of Psychology*, *11*, 213-229. doi:10.1080/00207597608247359

Brislin, R. W., Lonner, W., & Thorndike, R. (1973). *Cross-cultural research methods*. New York, NY: Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111-150. doi:10.1207/S15327906MBR3601\_05

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., & Cheung, F. M. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology*, *3*, 94-105. doi:10.1037/a0014516

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105, 456-466. doi:10.1037/0033-2909.105.3.456

Byrne, B. M., & van De Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*, 107-132. doi:10.1080/1530505100363730

Camilli, G., & Shephard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Canino, G., Lewis, Lewis-Fernández, R., & Bravo, M. (1997). Methodological challenges in cross-cultural health research. *Transcultural Psychiatry*, *34*, 163-184. doi:10.1177/13646159703400201

Cash, T. F. (2004). Body image: Past, present, and future. *Body Image*, *1*, 1-5. doi: 10.1016/S1740-1445(03)00011-1

Cash, T. F. (2017). *Body Image*: A joyous journey. *Body Image*, *23*, A1-A2. doi:10.1016/j.bodyim.2017.11.001

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276. doi:10.1207/s15327906mbr0102\_10

Cha, E.-S., Kim, K. H., & Erlen, J. A. (2007). Translation of scales in cross-cultural research: Issues and techniques. *Journal of Advanced Nursing*, *58*, 386-385. doi:10.1111/j.1365-2648-2007.04242.x

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology*, *23*, 530-537. doi:10.1076/jcen.23.4.530.1227

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504. doi:10.1207/s15328007sem1203\_7

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*, 1005. doi:10.1037/a0013193

Chen, F. F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research*, *36*, 462-494. doi:10.1177/0049124108314720

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471-492. doi:10.1207/s15328007sem1203\_7

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27. doi:10.1177/014920639902500101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. doi:10.1207/s15328007sem0902\_5

Child, D. (1990). *The essentials of factor analysis* (2nd ed.). London: Cassel Educational Limited.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309-319. doi:10.1037/1040-3590.7.3.309

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Collins, D. (2003). Pre-testing survey instruments: An over-view of cognitive methods. *Quality of Life Research*, *12*, 229-238. doi:10.10123/A:1023254226592

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Costello, A., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your data. *Practical Assessment, Research, and Evaluation*, *10*, 1-9.

Crawford, C. B., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills*, *49*, 223-226. doi:10.2466/pms.1979.49.1.223

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302. doi:10.1037/h0040957

Crowley, S. L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment*, *68*, 508-531. doi:10.1207/s15327752jpa6803\_4

Cudek, R., & O’Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin*, *115*, 475-487. doi:10.1037/0033-2909.115.3.475

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*, 558-575. doi:10.1177/0022022112438397

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40,* 55-75. doi:10.1146/annurev-soc-071913-043137

Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing*, *23*, 57-63. doi:10.1046/j.1365-2648.2003.02579.x

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399-412. doi:10.1111/bjop.12046

Epstein, J., Santo, R. M., & Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaire could not bring out a consensus. *Journal of Clinical Epidemiology*, *68*, 435-441. doi:10.1016/j.jclinepi.2014.11.021

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), APA handbooks in psychology. APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education (pp. 545-569). Washington, DC: American Psychological Association. doi:10.1037/14049-029

Eremenco, S. L., Cella, D., & Arnold, B. J. (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Evaluation and the Health Professions*, *28*, 212-232. doi:10.1177/0163278705275342

Erkut, S., Alarcón. C., García Coll, L. R., Tropp, H. A., & Vásquez García, H. A. (1999). The dual focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology*, *30*, 206-218. doi:10.1177/0022022199030002004

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis.* New York, NY: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299. doi:10.1037/1082-989X.4.3.272

Fan, X., & Sivo, S.A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*, 343–367. doi:10.1207/s15328007sem1203\_1

Fan, X., & Sivo, S.A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*, 509-529. doi:10.1080/0027370701382864

Ferreira, L., Neves, A. N., & Tavares, M. C. G. C. F. (2014). Validity of body image scales for Brazilian older adults. *Motriz: Revista de Educação Física*, *20*, 4. doi:10.1590/S1980-65742014000400002

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it. *European Journal of Psychological Assessment*, *33*, 399-402. doi:10.1027/1015-5759/a000460

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39-50. doi:10.2307/3151312

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, *7*, 356-410. doi:10.1207/S15328007SEM0703\_2

Gaderman, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation*, *17*, 1-13. Available online at: https://pareonline.net/pdf/v17n3.pdf

García-Coll, C., Akerman, A., & Cicchetti, D. (2000). Cultural influences on developmental processes and outcomes: Implications for the study of development and psychopathology. *Development and Psychopathology*, *12*, 333-356. doi:10.1017/S0954579400003059

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*, 93-111. doi:10.1037/met0000064

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6, 304-312. doi:10.1037/1040-3590.6.3.304

Gelin, M. N., Beasley, T. M., & Zumbo, B. D. (2003, April). What is the impact on scale reliability and exploratory factor analysis of a Pearson correlation matrix when some respondents are not able to follow the rating scale? Paper presented at the annual meeting of the American Educational Research Association (AERA) in Chicago, Illinois.

Glorfeld, L. W. (1995). An improvement on Horn’s parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*, 377-393. doi:10.1177/0013164495055003002

González-Martí, I., Bustos, J. G. F., Contreras, O. R. C., & Mayville, S. B. (2012). Validation of a Spanish version of the Muscle Appearance Satisfaction Scale: Escale de Satisfacción Muscular. *Body Image*, *9*, 517-523. doi:10.1016/j.bodyim.2012.05.002

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*, 313-317. doi:10.1027/1015-5759/a000450

Greiff, S., & Iliescu, D. (2017). A test is much more than just the test itself: Some thoughts on adaptation and equivalence. *European Journal of Psychological Assessment*, *33*, 145-148. doi:10.1027/1015-5759/a000428

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103,* 265-275. doi:10.1037/0033-2909.103.2.265

Gudmundsson, E. (2009). Guidelines for translating and adapting psychological instruments. *Nordic Psychology*, *61*, 29-45. doi:10.1027/1901-2276.61.2.29

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology*, *46*, 1417-1432. doi:10.1016/0895-4356(93)90142-N

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th edition). Upper Saddle River, NJ: Prentice Hall.

Hambleton, R. K., & de Jong, J. H. A. L. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing*, *20*, 127-134. doi:10.1191/0265532203lt247xx

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1*, 1-30.

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Culture and psychology: Cross-cultural research methods in psychology* (pp. 46-74). New York, NY: Cambridge University Press.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*, 306–324. doi:10.1177/0013164410384856

Harkness, J. (Ed.). (1998). *Cross-cultural equivalence*. Mannheim: ZUMA.

Harrell, F. E. (2001). *Regression modeling strategies*. New York, NY: Springer-Verlag.

Hayduk, L., Cummings, G. G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! Testing! One, two three: Testing the theory in structural equation models! *Personality and Individual Differences*, *42*, 841-50. doi:10.1016/j.paid.2006.10.001

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191-205. doi:10.1177/1094428104263675

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, *2*, 8. doi:10.9707/2307-0919.1111

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*, 319-336. doi:10.1037/a0024917

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling*, *19*, 36-50. doi:10.1080/10705511.012.634710

Henseler, J., Ringle, C. M., & Sarstedt, M. (2014). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*, 115-135. doi:10.1007/s11747-014-0403-8

Henson, R., & Roberts, J. (2006). Use of exploratory factor analysis in published research: Common errors and some content on improved practice. *Educational and Psychological Measurement*, *66*, 393-416. doi:10.1177/0013164405282485

Herdman, M., Fox-Rushby, J., & Badia, X. (1997). “Equivalence” and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research*, *6*, 237-247. doi:10.1023/A:1026410721664

Herdman, M., Fox-Rushby, J., & Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: The universalist approach. *Quality of Life Research*, *7*, 323-335. doi:10.1023/A:1024985930536

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*, 202-226. doi:10.1177/0013164404267287

Hooper, D., Couglan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*, 53-60.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332-346. doi:10.1177/1088868310361240

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *32*, 179-185. doi:10.1007/BF02289447

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117-144. doi:10.1080/03610739208253916

Horn, J. L., Wanberg, K. W., & Appel, M. (1973). On the internal structure of the MMPI. *Multivariate Behavioral Research*, *8*, 131-172. doi:10.1207/s15327906mbr0802\_1

Hu, L.-T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). London: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparmaterized model misspecification. *Psychological Methods*, *3*, 424-425.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi:10.1080/10705519909540118

Hu, L.-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351-362. doi:10.1037/0033-2909.112.2.351

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*, 131-152. doi:10.1177/0022002185016002001

Iliescu, D. (2017). *Adapting tests in linguistic and cultural contexts*. Cambridge: Cambridge University Press.

Jackson, D. L. (2003). Revisiting sample size and the number of parameter estimates: Some support for the *N*:q hypothesis. *Structural Equation Modeling*, *10,* 128-141. doi:10.1207/S15328007SEM1001\_6

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6-23. doi:10.1037/a0014694

Jewett, L. R., Kwakkenbos, L., Hudson, M., Baron, M., Thombs, B. D., & The Canadian Scleroderma Research Group. (2017). Assessment of English-French differential item functioning of the Satisfaction with Appearance Scale (SWAP) in systemic sclerosis. *Body Image*, *22*, 97-102. doi:10.1016/j.bodyim.2017.06.005

Jones, P., Lee, J. W., Phillips, L. R., Zhang, X. E., & Jaceldo, K. B. (2001). An adaptation of Brislin’s translation model for cross-cultural research. *Nursing Research*, *50*, 300-304.

Jöreskog, K. G. (1977). Factor analysis by least-squares and maximum-likelihood methods. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 125-153). New York, NY: Wiley.

Kaiser, H. F. (1970). A second-generation Little Jiffy. *Psychometrika*, *35*, 401-415. doi:10.1007/BF02291817

Kaiser, H. (1974). An index of factor simplicity. *Psychometrika*, *39*, 31-36. doi:10.1007/BF02291575

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. Sociological Methods and Research, 44, 486-507. doi:10.1177/0049124114543236

Khalaila, R. (2013). Translation of questionnaires into Arabic in cross-cultural research: Techniques and equivalence issues. *Journal of Transcultural Nursing*, *24*, 363-370. doi:10.1177/1043659613493440

Kline, P. (1979). *Psychometrics and psychology*. London: Academic Press.

Kline, R. B. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.

Kline, R. B. (2005). *Principles and practices of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Krach, S. K., McCreery, M. P., & Guerard, J. (2016). Cultural-linguistic test adaptations: Guidelines for selection, alteration, use, and review. *School Psychology International*, *38*, 3-21. doi:10.1177/0143034316684672

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202-220. doi:10.1177/1094428105284919

Lee, J. (2014). Conducting cognitive interviews in cross-national settings. *Assessment*, *21*, 227-240. doi:10.1177/1073191112436671

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

Lonner, W. J. (1985). Issues in testing and assessment in cross-cultural counseling. *The Counseling Psychologist*, *13*, 599-614. doi:10.1177/0011000085134004

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker’s congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57-64. doi:10.1027/1614-1881.2.2.57

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structural modeling. *Psychological Methods*, *1*, 130-149. doi:10.1037/1082-989X.1.2.130

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504. doi:10.1037/0033-2909.111.3.490

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong S. (1999). Sample size in factor analysis. Psychological Methods, 4, 84-99.

Malhotra, N. K., & Dash, S. (2011). *Marketing research: An applied orientation*. London: Pearson.

Malpass, R. S., & Poortinga, Y. H. (1986). Strategies for design and analysis. In J. W. Berry & J. W. Lonner (Eds.), *Cross-cultural research and methodology series: Vol. 8. Field methods in cross-cultural research* (pp. 47-83). Beverly Hills, CA: Sage.

Maneesriwongul, W., & Dixon, J. K. (2004). Instrument translation process: A methods review. *Journal of Advanced Nursing*, *48*, 175-186. doi:10.1111/j.1365-2648.2004.03185.x

Marcoulides, G. A. (1990). Evaluation of confirmatory factor analytic and structural equation models using goodness-of-fit indices. *Psychological Reports*, *67*, 669-670. doi:10.2466/pr0.1990.67.2.669

Marcoulides, G. A., & Chin, W. (2013). You write, but others read: Common methodological misunderstandings in PLS and related methods. In H. Abdi, W. Chin, V. E. Vinzi, G. Russolillo, & L. Trinchera L. (2013). *New perspectives in partial least squares and related methods* (pp. 31-64). Berlin: Springer-Verlag. doi:10.1007/978-1-4614-8283-3\_2

Marcoulides, G. A., & Saunders, C. (2006). Editor’s comments: PLS: A silver bullet? *MIS Quarterly*, *30*, iii–ix.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519-530. doi:10.1093/biomet/57/3/519

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett’s recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences*, *42*, 851-858. doi:10.1016/j.paid.2006.09.023

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling*, *1*, 317-359. doi:10.1080/10705519409539984

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*. Advance online publication. doi:10.1037/met0000113

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education*, *64*, 364-390. doi:10.1080/00220973.1996.10806604

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181-220. doi:10.1207/s15327906mbr3302\_1

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler’s (1999) findings. *Structural Equation Modeling*, *11*, 320-341. doi:10.1207/s15328007sem1103\_2

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right – camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*, 257-284. doi:10.1037/a0032773

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85-110. doi:10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students’ evaluations of university teaching. *Structural Equation Modeling, 16,* 439-476. doi:10.1080/10705510903008220

Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

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

McDermott, M. A., & Palchanes, K. (1992). A process for translating and testing a quantitative measure for cross-cultural nursing research. *Journal of the New York State Nurses’ Association*, *23*, 12-15.

McDonald, R. P., & Ringo Ho, M. H. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64-82. doi:10.1037//1082-989X.7.1.64

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, *100*, 43-52. doi:10.1080/00223891.2017.1281286

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*,*93*, 568-592. doi:10.1037/0021-9010.93.3.568

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543. doi:10.1007/BFO2294825

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement*, *14*, 5-8. doi:10.1111/j.1745-3992.1995.tb00881.x

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*, 35-44. doi:10.1023/a:1006964925094

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Montanelli, R. G., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonals: A Monte Carlo study. *Psychometrika*, *41*, 341-348. doi:10.1007/BF02293559

Morin, A. J. S., & Maïano, C. (2011). Cross-validation of the short form of the Physical Self-Inventory (PSI-S) using exploratory structural equation modeling (ESEM). *Psychology of Sport and Exercise*, *12*, 540-554. doi:10.1016/j.psychsport.2011.04.003

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York, NY: McGraw-Hill.

Mulaik, S. A., James, L. R., van Alstine, J., Bennet, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*, 430-445. doi:10.1037/0033-2909.105.3.430

Munet-Vilaró, F., & Egan, M. (1990). Reliability issues of the Family Environment Scale for cross-cultural research. *Nursing Research*, *39*, 244-247.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 588-620. doi:10.1207/S15328007SEM0904\_8

Myers, N. D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: A Monte Carlo approach. *Research Quarterly for Exercise and Sport*, *82*, 412-423. doi:10.1080/02701367.2011.10599773

Nagengast, B., & Marsh, H. W. (2013). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup SEMs across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and data analysis* (pp. 317-344). Boca Raton, FL: Chapman & Hall/CRC Press.

Nunkoo, R., Ramkissoon, H., & Gursoy, D. (2013). Use of structural equation modeling in tourism research: Past, present, and future. *Journal of Travel Research*, *52*, 759-771.
doi:10.1177/0047287513478503

Nunnally, J. C. (1976). *Psychometric theory* (2nd ed.). New York. NY: McGraw-Hill.

Nunnally, J. C. J., & Bernstein, I. (1995). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Osterlind, S. J., & Everson, H. T. (Eds.). (2009). *Differential item functioning.* Thousand Oaks, CA: Sage.

Parent, M. C. (2013). Handling item-level missing data: Simpler is just as good. *Counseling Psychologist*, *41*, 568-600. doi:10.1177/0011000012445176

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Perneger, T., V., Leplège, A., & Etter, J.-F. (1999). Cross-cultural adaptation of a psychometric instrument: Two methods compared. *Journal of Clinical Epidemiology*, *52*, 1037-1406. doi:10.1016/S0895-4365(99)00088-8

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science*, *19*, 12-21. doi:10.1080/1091367X.2014.952370

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage.

Philips, L. R., de Hernandez, L. L., & de Ardon, E. T. (1994). Focus on psychometrics: Steps for achieving cultural equivalence. *Research in Nursing and Health*, *17*, 149-154. doi:10.1002/nur.477017210

Polit, D. F., & Beck, C. T. (2012). *Nursing research: Generating and assessing evidence for nursing practice* (9th ed.). Philadelphia, PA: Wolters Klower Health, Lippincott Williams, and Wilkins.

Poortinga, Y. (1995). Use of tests across culture. In T. Oakland & R. K. Hambleton (Eds.), *Interpersonal perspectives on academic assessment* (pp. 187-206). Norwell, MA: Kluwer.

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics*, *32*, 153-161. doi:10.1013/A:1015210025234

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift’s electric factor analysis machine. *Understanding Statistics, 2*, 13-43. doi:10.1207/S15328031US0201\_02

Pulvers, K. M., Lee, R. E., Kaur, H., Mayo, M. S., Fitzgibbon, M. L., Jeffries, S. K., … & Ahluwalia, J. S. (2012). Development of a culturally relevant body image instrument among urban African Americans. *Obesity*, *12*, 1641-1651. doi:10.1038/oby.2004.204

*R* Development Core Team (2011). *R: A language and environment for statistical computing.* Vienna: *R* Foundation for Statistical Computing.

Reichenheim, M. E., & Moraes, C. L. (2007). Operationalizing the cross-cultural adaptation of epidemiological measurement instruments. *Revista de Saúde Pública*, *41*, 1-9. doi:10.1590/S0034-891200600500035

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145-154. doi:10.1007/s11336-008-9102-z

Rodgers, R., Schaefer, L. M., Thompson, K., Girard, M., Bertrand, M., & Chabrol, H. (2016). Psychometric properties of the Sociocultural Attitudes Towards Appearance Questionnaire-4 (SATAQ-4) in French women and men. *Body Image*, *17*, 143-151. doi:10.1016/j.bodyim.2016.003.02

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, 1715. doi:10.3389/fpsyg.2015.01715

Rose, R., & Fraser, M. (2008). A simplified framework for using multiple imputation in social work research. *Social Work Research*, *32*, 171-178.

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater, and test-retest reliability of continuous measurements. *Statistics in Medicine*, *21*, 3431-3446. doi:10.1002/sim.1253

Rummel, R. J. (1988). *Applied factor analysis* (4th ed.). Evanston, IL: Northwest University Press.

Sánchez-Carracedo, D., Barrasa, J. R., López-Guimerà, G., Fauquet, J., Almenara, C. A., & Trepat, E. (2012). Analysis of the factor structure of the Sociocultural Attitudes Towards Appearance Questionnaire (SATAQ-3) in Spanish secondary-school students through exploratory structural equation modeling. *Body Image*, *9*, 163-172. doi:10.1016/j.bodyim.2011.10.002

Saris, W. E., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561-582. doi:10.1080/2F10705510903203433

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In D. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A festschrift for Heinz Neudecker* (pp. 233-247). Dordrecht, The Netherlands: Kluwer Academic.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514. doi:10.1007/BF02296192

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83-90. doi:10.1007/BF02294150

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst’s perspective. *Multivariate Behavioral Research*, *33*, 545-571. doi:10.1207/s15327906mbr3304\_5

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*, 210-222. doi:10.1016/j.hrmr. 2008.03.003

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*, 304-321. doi:10.1177/0734282911406653

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner’s guide to structural equation modeling* (3rd ed.). New York, NY: Taylor and Francis.

Sechrest, L., Fay, T. L., & Zaidi, H. (1972). Problems of translation in cross-cultural research. *Journal of Cross-Cultural Psychology*, *3*, 41-56. doi:10.1177/002202217200300103

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*, 301-317. doi:10.1177/096228029800700306

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428. doi:10.1037/0033-2909.86.2.420

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s Alpha. *Psychometrika*, *74*, 107-120. doi:10.1007/s11336-008-9101-0

Sivo, S. A, Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for “optimal” cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education*, *74*,267-288. doi:10.3200/JEXE.74.3.267-288

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, *106*, 148-154. doi:10.1037/0033-2909.106.1.148

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, *2*, 107-129. doi:10.1207/S15237574IJT0202\_2

Sperber, A. D. (2004). Translation and validation of study instruments for cross-cultural research. *Gastroenterology*, *126*, S124-S128. doi:10.1053/j.gastro.2003.10.016

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation: Methodology and validation. *Journal of Cross-Cultural Psychology*, *25*, 501-524. doi:10.1177/00220022194254006

Sprangers, M. A. G., Cull, A., Bjordal, K., Groenvold, M., Aaronson, N. K., & EORTC Study Group on Quality of Life. (1993). The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: Guidelines for developing questionnaire modules. *Quality of Life Research*, 2, 287-295. doi:10.1007/BF00434800

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78-90. doi:10.1086/209528

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*, 893-98. doi:10.1016/j.paid.2006.09.017

Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., … & Carpenter, J. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, *339*, 157-160. doi:10.1136/bmj.b2393

Stevens, J. P. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Sumathipala, A., & Murray, J. (2006). New approach to translating instruments for cross-cultural research: A combined qualitative and quantitative approach for translation and consensus generation. *International Journal of Methods in Psychiatric Research*, *9*, 87-95. doi:10.1002/mpr.83

Svetina, D., & Rutkowski, L. (2018). Multidimensional measurement invariance in an international context: Fit measure performance with many groups. *Journal of Cross-Cultural Psychology*, *48*, 991-1008. doi:10.1177/0022022117717028

Swami, V. (2015). Cultural influences on body size ideals: Unpacking the impact of Westernization and modernization. *European Psychologist*, *20*, 44-51. doi:10.1027/1016-9040/a000150

Swami, V. (2018a). Considering positive body image through the lens of culture and minority social identities. In C. Markey, E. Daniels, & M. Gillen (Eds.), *The body positive: Understanding and improving body image in science and practice* (in press). Cambridge: Cambridge University Press.

Swami, V. (2018b). Measures of body image and disordered eating for use with Malaysian populations: A critical review and methodologic critique of the recent literature. In J. L. Jaafar & D. Barron (Eds.), *Psychology in Southeast Asia: Voices from the non-western communities* (in press)*.* London: Routledge.

Swami, V., & Barron, D. (2017a). Recommendations to improve body image research in an increasingly globalised world. *Malaysian Journal of Nutrition*, *23*, 3-10.

Swami, V., & Barron, D. (2017b). *International publication trends in Body Image, 2004-2015: The signs are encouraging, but more can still be done*. Unpublished manuscript.

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

Swami, V., Campana, A. N. N. B., Ferreira, L., Barrett, S., Harris, A. S., & Tavares, M. C. G. C. F. (2011). The Acceptance of Cosmetic Surgery Scale: Initial examination of its factor structure and correlates among Brazilian adults. *Body Image*, *8*, 179-185. doi:10.1016/j.bodyim.2011.01.001

Swami, V., & Chamorro-Premuzic, T. (2008). Factor structure of the Body Appreciation Scale among Malaysian women. *Body Image*, *5*, 409-413. doi:10.1016/j.bodyim.2008.004.005

Swami, V., Salem, N., Furnham, A., & Tovée, M. J. (2008). Initial examination of the validity and reliability of the female Photographic Figure Rating Scale for body image assessment. *Personality and Individual Differences*, *44*, 1752-1761. doi:10.1016/j.paid.2008.02.002

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. doi:10.1016/j.bodyim.2018.02.004

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). New York, NY: Pearson.

Tanaka, J. S. (1987). How big is big enough? Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*, 134-146. doi:10.2307/1130296

Tao, Z., & Zhong, W. (2010). Eating attitudes and weight concern among Chinese middle-age women: A comparison between different age and BMI groups. *European Journal of Psychiatry*, *24*, 146-157. doi:10.4321/S0213-61632010000300003

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages, and some neglected topics. *Medical Care*, *44*, S152-S170. doi:10.1097/01.mlr.0000245142.74628.ab

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly*, *50*, 538-612.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Triandis, H. C. (1975). Social psychology and cultural analysis. *Journal of the Theory of Social Behaviour*, *5*, 81-106. doi:10.1111/j.1468-5915.1975.tb00344.x

Triandis, H. C. (1976). Approaches toward minimizing translation. In R. Brislin (Ed.), *Translation: Applications and research* (pp. 229-243). New York, NY: Wiley/Halstead.

Triandis, H. C., & Brislin, R. W. (1984). Cross-cultural psychology. *American Psychologist*, *39*, 1006-1016. doi:10.1037/0003-066X.39.9.1006

Truxillo, C. (2003). *Multivariate statistical methods: Practical research applications course notes*. Cary, NC: SAS Institute.

Tsai, T.-I., Luck, L., Jefferies, D., & Wilkes, L. (2018). Challenges in adapting a survey: Ensuring cross-cultural equivalence. *Nurse Researcher*, *26*, 28-32. doi:10.7748/nr.2018.e1581

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

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. doi:10.1016/j.bodyim.2013.02.006

Tylka, T. L., & Iannantuono, A. C. (2016). Perceiving beauty in all women: Psychometric evaluation of the Broad Conceptualization of Beauty Scale. *Body Image*, *17*, 67-81. doi:10.1016/j.bodyim.2016.02.005

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

Ullman, J. B. (2006a). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*, 35-50. doi:10.1207/s15327752jpa8701\_03

Ullman, J. B. (2006b). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (5th ed., pp. 653-771). Boston, MA: Allyn & Bacon.

van de Schoot, R., Schmidt, P., & DeBeuckelaer, A. (2015). *Measurement invariance*. Lausanne: Frontiers Media SA.

van de Vijver, F. J. R. (2011). Bias and real differences in cross-cultural differences: Neither friends not foes. In F. J. R. van de Vijver, A. Chasiotis, & S. M. Breugelmans (Eds.), *Fundamental questions in cross-cultural psychology* (pp. 235-258). New York, NY: Cambridge University Press. doi:10.1017/cbo9780511974090.011

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*, 89-99. doi:10.1027/1016-9040.1.2.89

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Research*, *31*, 33-51. doi:10.1177/0022022100031001004

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. R. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17-45). New York, NY: Cambridge University Press. doi:10.1017/cbo9780511779381.003

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 29-37. doi:10.1027/1015-5759.13.1.29

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*, 263-279. doi:10.1016/j.erap.2003.12.004

van Widenfelt, B. M., Treffers, P. D. A., de Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical Child and Family Psychology Review*, *8*, 135-147. doi:10.1007/s10567-005-4752-1

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-70. doi:10.1177/109442810031002

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321-327. doi:10.1007/BF02293557

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 47-71). Norwell, MA: Kluwer Academic.

Vossbeck-Elsebusch, A. N., Waldorf, M., Legenbauer, T., Bauer, A., Corde, 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. doi:10.1016/j.bodyim.2014.02.002

Wang, W.-L., Lee, H.-L., & Fetzer, S. J. (2006). Challenges and strategies of instrument translation. *Western Journal of Nursing Research*, *28*, 310-321. doi:10.1177/0193945905284712

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlational coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*, 231-240.

Weston, R., & Gore, P. A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist, 34,* 719-751. doi:10.1177/0011000006286345

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *Journal of Experimental Education, 80*, 26-44. doi:10.1080/00220973.2010.531299

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient reported outcome (PRO) measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*, 94-104. doi:10.1111/j.1524.4733-2005.04054.x

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *76*, 913-934. doi:10.1177/0013164413495237

Worthington, R., & Whittaker, T. (2006). Scale development research: A content analysis and recommendations for best practice. *Counseling Psychologist*, *34*, 806-838. doi:10.1177/0011000006288127

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation*, *12*, 1-26.

Yu, D. S. F., Lee, D. T. F., & Woo, J. (2004). Issues and challenges of instrument translation. *Western Journal of Nursing Research*, *26*, 307-320. doi:10.1177.01939459003260554

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*, 405-426. doi:10.1037/met0000080

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indices. *Structural Equation Modeling*, *23*, 319-330. doi:10.1080/10705511.2015.1065414

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact vs. approximate measurement invariance. *Frontiers in Psychology*, *6*,733. doi:10.3389/fpsyg.2015.00733

Ziegler, M., & Bensch, D. (2013). Lost in translation. Thoughts regarding the translation of existing psychological measures into other languages. *European Journal of Psychological Assessment*, *29*, 81-83. doi:10.1027/1015-5759/a000167

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-Type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*, 136-147. doi:10.1191/02655322203lt248oa

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods* *6*, 4. doi:10.22237/jmasm/1177992180

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modelling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442. doi:10.1207/s15327906mbr1702\_5

Ægisdóttir, S., Gerstein, L. H., & Çinarbas, D. C. (2008). Methodological issues in cross-cultural counselling research: Equivalence, bias, and translations. *The Counseling Psychologist*, *36*, 188-219. doi:10.1177/0011000007305384