Translation and Validation of Body Image Instruments: An Addendum to Swami and Barron (2019) in the Form of Frequently Asked Questions

Viren Swami[1-2], Jennifer Todd[1-2], & David Barron[2]

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

[2]Centre for Psychological Medicine, Perdana University, Kuala Lumpur, 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@aru.ac.uk.

Abstract

Test adaptation – the translation and validation of source instruments for use in new social identity groups – plays a vital role in body image research. Previously, Swami and Barron (2019) developed a set of good practice recommendations and reporting guidelines for the test adaptation of body image instruments. However, a number of issues in that article were not covered in depth and new issues have emerged as a result of developments in theory and practice. Here, we offer an addendum to Swami and Barron in the form of frequently asked questions. Issues discussed in this article include various methods for achieving good translations, the appropriateness of revising instrument components prior to empirical analyses, determining the number of factors to extract in exploratory factor analyses (EFA), and the usefulness of EFA versus confirmatory factor analyses (CFA) in determining factorial validity. We also cover methods of analyses that have been infrequently utilised by body image scholars, including exploratory structural equation modelling (ESEM), bifactor model analyses, and various methods for establish measurement invariance. When read as an addendum to Swami and Barron, we hope this article helps to clarify issues of importance for body image researchers interested in conducting test adaptation work.

Keywords: Test adaptation; Translation; Exploratory factor analysis; Confirmatory factor analysis; Exploratory structural equation modelling; Bifactor model

1. Introduction

Almost two decades have now passed since *Body Image* was first launched as a home for studies on "physical appearance and body image in diverse cultural contexts" (Cash, 2004, p. 3). In that time, the journal has played a crucial role in supporting and deepening research on body image in historically neglected cultural, national, and linguistic groups (Cash, 2017), but also in facilitating stronger scholarly connections between researchers internationally (Andersen & Swami, 2021; McCabe et al., 2019). An important component of these developments has been the publication of studies on *test adaptation* (i.e., studies reporting on the translation and validation of source instruments for use in new social identity groups that are different from the one in which it was originally developed). Indeed, unlike many other journals, *Body Image* regularly publishes test adaptation studies – an important service for the community of body image scholars and practitioners globally (Tylka et al., 2020). Importantly, this support for test adaptation studies ensures that the journal continues to provide a voice for communities and groups who have been historically marginalised from psychological research.

As a contribution to that body of work, we previously developed a set of good practice recommendations and reporting guidelines for the test adaptation of body image instruments (Swami & Barron, 2019). In brief, in that article we discussed methods used to ensure semantic equivalence through translation techniques and methods of ensuring measurement equivalence through data treatment and analyses. We have been heartened to see our advice considered and applied by body image scholars, which we believe has – along with improvements in practice more generally – helped produce higher quality work. We are also encouraged to see body image scholars considering, discussing, and evaluating what remain important issues for the field as a whole. However, it has also become apparent to us that there were a number of questions in Swami and Barron (2019) that were not dealt with in

sufficient depth, new areas of enquiry that require commentary, and frequently-encountered issues that could do with some clarification.

Our objective in this article is to answer some of those questions and provide better coverage of test adaptation topics that we hope will be of interest and practical use to body image scholars. In that sense, we strongly recommend that this article is read as an addendum to Swami and Barron (2019), rather than as a standalone article. That is, while Swami and Barron (2019) covered most of the important issues that will be of interest to readers interested in test adaptation, the present paper considers particular issues in more detail, adds context that was previously omitted for the sake of brevity, or answers questions that are often posed to us by other researchers. In order to facilitate reading, we have structured the present article in the form of responses to frequently asked questions. These represent some of the most common questions we have encountered since the publication of Swami and Barron (2019); while we have tried to cover topics and issues that we believe are important, these questions may not be exhaustive.

2. Frequently Asked Questions and Our Responses

2.1. What is the best way to translate an instrument?

This question assumes that there is a singular or ideal way for instruments to be translated into a new language. However, as were at pains to point out in Swami and Barron (2019), there are multiple ways in which an instrument could be translated – most of which we reviewed previously – and no one method will suit all practicalities (van Widenfelt et al., 2005). Instead, the method that a scholar chooses should take into consideration issues such as feasibility of comprehensive versus minimalist translation methods, the monetary and time cost of different methods, and the instrument itself. As examples, a more comprehensive translation strategy will be important for instruments consisting of multiple statement-based

items (particularly items that contain idiomatic expressions or that may pose concerns around cultural appropriateness), but may be of less value when translating simple instructions and response options for figural rating scales. Given that there are no "gold standards" for translating instruments (Wild et al., 2005, 2008), we reiterate the need for scholars to make translation steps transparent through detailed reporting and to justify the design of the translation process. We further encourage researchers to follow and report compliance with checklists for test adaptation, such as the criterion checklist (Hernández et al., 2020) based on the International Test Commission's (2017) guidelines.

Unfortunately, we continue to see some scholars relying on back-translation alone as their sole translational method. There is strong consensus that the application of back-translation alone is likely to produce poor translations (Maneesriwongul & Dixon, 2004; van Widenfelt et al., 2005), often because, on its own, back-translation does not address issues of conceptual equivalence, respondent comprehension, and contextualised meaning (Colina et al., 2017; Douglas & Craig, 2007). Moreover, when used in the absence of iterative processes, when translations are produced by a single translator, or when instruments include linguistic ambiguities, back-translation alone often results in inconsistencies in the detection of flaws, resulting in many translational problems remaining hidden (Behr, 2017; Ozolins et al., 2020). To put it bluntly, researchers should not infer that bilingual translators alone will be able to produce adequate translations, even if they have expertise on the content of the instrument (Solano-Flores, 2008).

When researchers rely on back-translation alone, there is a risk that multiple forms of translational equivalence will be violated. Instead, we repeat our recommendation that scholars should aim at combining multiple techniques in a multi-stage approach when translating instruments (Brislin et al., 1973), particularly as moderately resource-intensive methods are often sufficient to produce sound translations (Perneger et al., 1999). In

particular, the adoption of a committee approach (see Section 2.2.) is especially useful in minimising threats to validity (Solano-Flores et al., 2009), though we recognise that this approach can be time-consuming and resource-intensive. If a minimum recommendation is sought, then the combination of the back-translation method and pre-testing (assessing understanding, acceptability, and/or emotional impact of the items in a pilot study) may suffice (Maneesriwongul & Dixon, 2004), although we stress this is a minimum standard that often can and should be surpassed.

2.2. If a committee approach is used for translating instruments, who should be included on the committee?

When a multi-stage translation methodology is used, researchers often include a committee approach, wherein a team of researchers scrutinise forward- and back-translations that have been produced in earlier steps (Cha et al., 2007). In Beaton and colleague's (2000) widely-used 5-stage procedure, for instance, the final stage consists of a review of forward- and back-translations produced in earlier stages, with a view to consolidating all versions of the instrument and, through consensus, developing a prefinal version of the instrument that can be field-tested. This committee plays an important role in resolving discrepancies between translations, as well as in achieving semantic, idiomatic, experiential, and conceptual equivalence between the source and target versions of the instruments. Given this important role, our recommendation is that any such committee should be multidisciplinary, with representation from translators involved in earlier stages, content specialists (i.e., researchers with knowledge of the instruments and/or research topic), linguists or language professionals, and methodologists or psychometricians. If possible, the original developers of the instrument should also be involved in this committee.

Where a committee approach is utilised, there may be added value in adopting a stance that identifies errors in test translation as inevitable (i.e., the *theory of test translation*

error or TTTE; Solano-Flores et al., 2009). This approach suggests that errors are inevitable because languages encode meaning in different ways and because instruments typically impose severe restrictions in the way that meaning can be conveyed. Consider the following item, taken from the Appearance Schemas Inventory-Revised (ASI-R; Cash et al., 2004):

> I spend little time on my physical appearance.

In English, the item is likely intended to convey the equivalent meaning as "I do not spend much time on my physical appearance". However, multiple translations are possible and, depending on the target language and linguistic constraints, could result in altered meanings, such as:

> I spend a little time on my physical appearance.
> I spend some time on my physical appearance.
> I do not spend time on my physical appearance.
> I do not spend much time worrying about my physical appearance.
> I do not waste much time on my physical appearance.
> I do not care about my physical appearance.
> My physical appearance is a little time-consuming.

In each of these cases, the intended meaning of the item has been altered to a lesser or greater degree from the original. Such discrepancies can occur because words or terms that are used or that are familiar in one language may not be in another, or because grammatical differences across languages make it difficult to convey the item's original meaning. In such cases, the point of TTTE-based methods is to find disconfirming evidence that items on an

instrument have been translated correctly (i.e., identifying test translation errors), rather than – as in conventional methods – finding evidence that a translation is adequate (Solano-Flores et al., 2009). In the case of the ASI-R translations above, for example, a TTTE-based approach would problematise each of the item translations, which in turn would highlight the need to consider alternative approaches to convey the intended meaning of the item.

While the rationale of TTTE is that translation errors cannot be avoided, this is not to say they cannot be minimised (Solano-Flores et al., 2009), especially when due attention is given in a consensus-based approach to item design, language, and content. In this view, the most useful translation will be the one that minimises inevitable translation errors. In fact, there is some evidence that a TTTE-based approach is able to detect translation errors with a high degree of precision (Solano-Flores et al., 2013), although typologies or dimensions of errors will need to be sensitive to local issues (Zhao et al., 2018). In addition, given the importance of consensus in this form of committee approach, it is vital that issues of power and status associated with committee members (e.g., seniority, age, gender, race) do not impede discussions. Zhao and Solano-Flores (2021) have suggested ways in which the impact of power and status can be minimised, the most important of which is the inclusion of a facilitator.

2.3. Is it acceptable to alter response anchors when preparing a translation?

When preparing a translation, an important issue that researchers have to deal with is the translation of response options or anchors. This is important because research has shown that familiarity with the category labels used in response scales can affect the manner of responding (Weijters et al., 2013). To take a simple example: an instrument may use a *strongly disagree–strongly agree* response scale in English because these anchors are commonly used in everyday speech and thus will likely be familiar to respondents. When a literal translation is produced, however, there is no guarantee that the translated anchors will

be equally familiar or contain the same semantic meaning. For instance, in Bahasa Malaysia, "strongly agree" is literally translated as "*sangat bersetuju*" ("*in strong agreement*"), but this translation is overly formal compared to the more colloquial "*sangat setuju*" (closer to "*agree strongly*"). When response scales are less familiar, respondents are less likely to fully endorse scale endpoints, which can in turn result in artefactual biases in latent responses (Baumgartner & Weijters, 2015). In these cases, we recommend translating scale responses so that they are familiar to respondents, even if they are not literal equivalents.

A related question is whether it is acceptable to alter the number of response options that are presented to participants (e.g., from a 5- to a 7-point scale). Researchers may sometimes believe that there are good reasons to do so. For instance, they may be concerned about possible *response biases* (i.e., a tendency to disproportionately select a subset of response options; Baumgartner & Steenkamp, 2001) in a particular national or linguistic group. These may include acquiescence (i.e., disproportionate use of positive response options), dis-acquiescence (disproportionate use of negative response options), and extreme response styles (Weijters et al., 2010), all of which can be affected by language and culture (Baumgartner & Weijters, 2015). It should be noted, however, that these concerns can usually be mitigated through effective translations that remove ambiguities or uncertainty in the meaning of items (Baumgartner & Steenkamp, 2006). Where responses biases are deemed to be a real concern, there are alternative strategies for dealing with such biases that do not require alteration of the number of response options (for a review of such strategies, see Baumgartner & Weijters, 2015).

In other situations, there may be legitimate concerns about whether a response scale is able to capture subtle degrees of measure that participants from a particular cultural or linguistic group want to express (Hamamura et al., 2008). For instance, some research has suggested that 5-point response scales can sometimes be too coarse a method for allowing

participants to accurately express their true intended responses (Russell & Bobko, 1992). Instead, response scales that approximate a more continuous distribution are often favoured (Cummins & Gullone, 2000; Nunnally, 1978), with 7-point scales usually balancing requirements for sensitivity and reliability (Cox, 1980; Diefenbach et al., 1993; Preston & Colman, 2000). In this situation, researchers may be tempted to alter response options, such as from a 4- or 5-point scale to a 7-point scale. Although well-intentioned, doing so also introduces complexities in drawing comparisons of latent means across studies. Here, we recommend that scholars think carefully about the need for altering response options, compare responding using different response options where possible (e.g., see Diefenbach et al., 1993), and always seek permission from the instrument developers before making any alterations.

2.4. Is it acceptable to truncate an instrument prior to data collection?

In some cases, researchers may be tempted to truncate (i.e., remove items) from an instrument either before or after translation, but prior to data collection. There may be a number of reasons for believing that item truncation is necessary, some legitimate and others less so. In the first instance, there may be difficulties translating items, particularly items that are idiomatic or where equivalent translations are not possible in the new language (see Section 2.2.). In these cases, we encourage researchers to work collaboratively, if possible, with the instrument's developers to design revised or new items that capture the conceptual meaning of items or to utilise a committee approach to translation (Brislin et al., 1973). In most cases, *accommodations* – changes to the item without altering the underlying meaning of the item or the construct being measured – should be possible. Item elimination should only be considered in extreme cases and, where they deemed necessary, should only be undertaken with the permission of the instrument developers and the number of eliminated items should be kept to a minimum.

In other instances, researchers may decide *a priori* that an instrument includes too many items (such that response fatigue becomes an issue of concern), includes items that would (presumably) be excluded anyway following data reduction, or that are problematic for other (often unspecified) reasons. We suggest that these are not adequate reasons for item elimination. Elimination of items without a strong rationale that is grounded in theory or related to translational difficulties can result in construct under-representation, that is, when content that makes up a construct is not represented in the instrument (Messick, 1994). Given that scores on a translated instrument should authentically represent the targeted construct, item elimination prior to analyses can seriously affect the validity of derived scores. In these cases, we strongly recommend retaining all items and making instrument truncation decisions based on empirical analyses, rather than *a priori* assumptions.

2.5. Is it acceptable to add items to an instrument prior to data collection?

Occasionally, researchers may determine there is a need to add new items to an existing instrument prior or subsequent to its translation, but prior to data collection. This may be because they have identified a clear conceptual gap in the instrument (i.e., the instrument does not fully capture the meaning of a construct in a particular cultural or linguistic context) or for purely exploratory reasons (i.e., to see whether new items are retained in analyses). We would discourage such revisions to an instrument, as improperly designed items can often affect the reliability and validity of scores on an instrument, introduce multidimensionality where there was none previously, or change the conceptual meaning of constructs being measured. In cases where researchers have clearly identified omissions in an instrument, we strongly recommend discussing any revisions with the instrument's developers before making any decisions about new additions. Where additions are included in an instrument, it may be useful to include these after the original items to minimise possible contamination or semantic priming effects (e.g., Dignard & Jarry, 2019).

A different form of this issue is more problematic. Scholars occasionally combine instruments that have been developed in other cultural contexts for use in local settings. The idea here appears to be that, by combining multiple instruments and treating scores as having been derived from a "single" instrument, it will be possible to produce "new" multidimensional instruments for local use. We discourage such methods in the strongest possible terms. Doing so would represent a significant departure from the intended purpose of the original instruments and would likely result in conceptually muddled constructs. It would also be unethical should researchers fail to first obtain permission to engage in such conduct. Should researchers feel there is a need to construct a novel measure of body image for use in local settings, we would instead encourage them to utilise appropriate scale construction methods, some of which we discussed previously in Swami and Barron (2019).

2.6. Could two or more instruments be validated concurrently?

There is no reason why multiple instruments could not be translated concurrently, especially if identical methods and procedures are applied consistently across instruments. Likewise, there is no reason why the validation of multiple instruments could not be reported in the same manuscript (for an example, see Swami et al., 2015), particularly if said instruments share some commonalities (e.g., they are all measures of a specific facet of body image). One benefit of this approach is that it allows researchers to simultaneously validate multiple instruments; that is, scores on each instrument produces an index of validity for every other instrument included in the analysis. However, a common pitfall of this approach is that scores on one or more instrument provide less-than-adequate indices of validity. In such cases, establishing the validity of multiple measures simultaneously can be problematic. One way to mitigate against this is to include additional survey instruments that have been previously validated, so as to maximise the likelihood of being able to report something meaningful.

In terms of the latter, we strongly encourage authors to pay careful attention to the nomological net of scores on an instrument, which refers to assumed relationships with other constructs (Cronbach & Meehl, 1955). We continue to see some authors utilising excellent methods for translating an instrument, only to be let down by neglecting to adequately attend to other design-related issues *vis-à-vis* nomological nets (e.g., not including additional variables to establish construct validity, establishing construct validity through instruments that have not been previously validated or that are not concurrently validated). Understanding the nomological net is vital in test adaptation studies because it can help scholars interpret the usefulness of an instrument in a new cultural or linguistic setting and because it provides information about the internal structure of scores on an instrument (Loevinger, 1957). In addition, understanding the nomological net is important in terms of developing appropriate hypotheses that guides the research (Campbell & Fiske, 1959), particularly in terms of establishing support for multiple forms of validity.

2.7. Should fit indices be used to determine the number of factors to be extracted in exploratory factor analysis?

A central question in exploratory factor analysis (EFA) is: how many factors should be extracted (for a review, see Preacher & MacCallum, 2003)? Traditional approaches to answering this question have relied on eigenvalues of the observed or reduced correlation matrix, which provide approximate information about matrix dimensionality (i.e., they provide an indication of how much information can be explained by subsequent components). Two popular approaches are the Kaiser or eigenvalue > 1 (or K1) rule (i.e., factors with eigenvalues > 1 are retained) and Cattell's scree test (i.e., eigenvalues are plotted and researchers retain the number of factors based on the last "substantive" visual drop-off). However, as we indicated in Swami and Barron (2019), both of these methods have been shown to be deeply problematic, resulting in either factor over- or under-retention depending

on the circumstances. Instead, we advocated for more consistent use of parallel analysis, also based on eigenvalues but with much stronger empirical support (Dinno, 2009).

More recently, however, it has become possible to use fit indices commonly used in confirmatory factor analysis (CFA) to solve the "number of factors" problem in EFA (Clark & Bowles, 2018). Here, the number of factors to be retained is determined by selecting the model that falls below commonly-used fit cut-offs (Preacher et al., 2013), with some combination of the root mean square errors of approximation (RMSEA), the comparative fit index (CFI), the standardised root mean squared residual (SRMR), and the Tucker-Lewis Index (TLI) currently popular. Although the application of these fit index recommendations has become increasingly popular for factor retention decisions in EFA, there is some concern that model fit indices perform too unpredictably to warrant their use for categorical (Clark & Bowles, 2018; Garrido et al., 2016) and continuous data (Montoya & Edwards, 2021). For instance, recent simulation work has suggested that – with the possible exception of SRMR – fit indices are overly sensitive to correlated residuals and non-specific error, which increases the likelihood of factor over-retention (Montoya & Edwards, 2021; but see Finch, 2020, who found that the use of fit indices outperformed parallel analysis when factor loadings were small).

Given such issues, our recommendation for now – which is consistent with the conclusion of Montoya and Edwards (2021) and may change in the future – is to continue using parallel analysis to determine the number of factors to be extracted in EFA. Alternatively, it may be useful to use a combination of methods, such as the use of fit cut-offs (when using statistical software that provide fit indices, such as some packages in *R*) alongside parallel analysis. For the interested reader, other traditional and newer factor retention methods are reviewed and assessed in Auerswald and Moshagen (2019), who likewise advocate for the use of combined methods for determining the number of factors to

be extracted in EFA. In either case, it may still be beneficial to report fit statistics for a final selected model, as this provides a general index of factor structure adequacy (Finch, 2020). As with CFA, we recommend the reporting of a combination of fit indices (see Swami & Barron, 2019, for details).

2.8. Isn't confirmatory factor analysis better than exploratory factor analysis?

The idea that CFA is in some way "better" than EFA is a common misconception that we continue to see some body image researchers subscribe to, often implicitly though occasionally also explicitly. The idea manifests in a number of ways, such as when researchers believe CFA is a "gold standard" for establishing measurement equivalence, prioritise CFA over EFA (e.g., making decisions about score dimensionality based on CFA alone or discounting the evidence provided by EFA in favour of that provided by CFA), and when they discount EFA entirely as an analytic strategy for test adaptation (i.e., not conducting and reporting an EFA prior to conducting a CFA). Such beliefs likely stem from a perception that EFA is purely "exploratory" and hence less hypothesis-driven than CFA (Marsh et al., 2014); that is, in comparison with CFA, EFA is sometimes be viewed as inherently flawed because results are based on subjective interpretations of dimensionality rather than, say, more "objective" fit indices.

However, as we hinted at previously (Swami & Barron, 2019), these assumptions are inherently problematic because EFA and CFA have different purposes. In broad outline, EFA is a useful tool for understanding the latent dimensionality of scores in a dataset (i.e., a data-driven approach), whereas CFA is useful for understand whether data fit *a priori* hypothesised measurement models (i.e., a hypothesis-driven approach). Because EFA is not restricted by modelling limitations, it helps researchers determine the best factorial solution for their dataset. On the other hand, because CFA *is* restricted by modelling limitations, it helps researchers determine whether hypothesised (or earlier data-driven) models fit the data

in their dataset. As such, there is little value in viewing EFA and CFA as oppositional strategies; instead, EFA and CFA should be viewed as complementary tools in the arsenal of researchers conducting test adaptation (Swami & Barron, 2019).

Consider the following example. There is a hypothetical instrument – we will call it the Questionnaire of Body Image (QBI) – that consists of 10 items. In the parent study, the developers of the QBI showed, through both EFA and CFA, that scores on the QBI are unidimensional (i.e., all 10 items load onto a single factor). Now, a different scholar decides they would like to use the QBI in a new national or linguistic group; they've appropriately translated the instrument and asked a large number of respondents to complete it. When analysing the data, this researcher decides that CFA is superior to EFA and, therefore, only applies the former analytic method. In practice, this means they would only test the fit of the unidimensional model of QBI scores. However, this leaves open – and untested – the possibility that QBI scores in this dataset (or in this linguistic and national group) are in fact multidimensional (i.e., scores reduce to more than one dimension). Had the researcher adopted an EFA-to-CFA strategy, they would have identified the multidimensional model through EFA and been able to compare the fit of both the multidimensional and parent unidimensional model of QBI scores through CFA.

It may be tempting to view these issues as trivial. After all, if the unidimensional model of QBI scores fits the data in CFA, does it really matter that an EFA shows scores to be multidimensional? We would counter that these issues are non-trivial for a number of reasons. First, from a theoretical point-of-view, considering alternative representations that may account for variation in the data, as well as preferred factor solutions, is important as it helps to prevent confirmation bias among researchers (e.g., when researchers only seek support for a particularly factorial representation; Kline, 2016). To return to our earlier example, identifying the fact that QBI scores are unidimensional in some contexts but

multidimensional in others may allow researchers to ask new research questions about the construct being measured, identify hitherto neglected factors of interest, or develop new understandings of the meaning and experience of the construct. From a practical point-of-view, researchers and practitioners need to be certain that the way instruments are scored in particular contexts accurately reflects latent dimensionality, especially when instruments are used for diagnostic purposes. Perhaps most importantly, given that most body image instruments are initially developed in Anglophone contexts and often with White respondents, there is a need to ensure that voices of diverse communities – as can sometimes be expressed through data-driven analytic approaches – are not marginalised and rendered invisible.

As such, we continue to advocate for the use of an EFA-to-CFA strategy (see also Worthington & Whittaker, 2006) for test adaptation, where this is feasible (e.g., where sample and subsample sizes are large enough; EFA and CFA should not be conducted on the same subsamples). There may, of course, be occasions when it is not be practical to run both EFA and CFA, such as when a target population is difficult to sample in sufficiently large numbers. In these cases, we suggest that it is in fact EFA, and not CFA, that is the superior first-pass analytic approach in the context of test adaptation. Not only does EFA offer a more thorough understanding of item behaviour in localised contexts, but there may also be reasons to doubt the utility of CFA as a standalone analytic strategy in these cases, particularly when researchers are dealing with potentially multidimensional constructs. Specifically, in CFA, items are only allowed to load on to their respective hypothesised latent factors, whereas cross-loadings are forced to be zero (Marsh et al., 2009, 2010; Morin et al., 2016). In other words, the items associated with each factor are assumed to be “pure” indicators of that factor and there will be no associations between items and non-target conceptually-related constructs (Marsh et al., 2014).

When dealing with multidimensional constructs, such an assumption would seem to be highly improbable (Marsh et al., 2014; Morin et al., 2016); that is, CFA alone may be too restrictive. Of course, in many cases, parent studies use EFA to eliminate items that cross-load onto more than one factor. Even here, however, items may still be fallible indicators of target constructs and therefore may still have residual associations with non-target constructs (Asparouhov & Muthén, 2009; Marsh et al., 2013, 2014). Alternatively, it is quite possible that items demonstrate hitherto un-theorised or unexpected cross-loadings in new national or linguistic groups, possibly because the meaning of individual items or of the constructs themselves differ across groups. In these cases, expecting cross-loadings to be zero in CFA typically results in inflated estimates of factor correlations and hence model misspecification (Asparouhov & Muthén, 2009; Asparouhov et al., 2015; Marsh et al., 2011, 2014). Moreover, current methods for dealing with these issues in CFA (e.g., *post hoc* modifications) have been problematised (Marsh et al., 2010). In short, CFA may simply not be a useful strategy for test adaptation when used in isolation.

2.9. Should we be using exploratory structural equation modelling instead?

Given the limitations associated with CFA discussed in Section 2.8, some researchers have suggested relying on exploratory structural equation modelling instead (ESEM; Marsh et al., 2013, 2014; Morin et al., 2013, 2020), which can be conducted using programmes such as MPlus and *R*. In broad outline, ESEM is an analytic strategy that relaxes the requirement of zero cross-loadings while providing access to information usually restricted to CFA, such as goodness-of-fit statistics, residual correlations, and standard error estimates; that is, ESEM combined aspects of both EFA and CFA (Morin et al., 2013) and has been shown to result in improved fit and less strongly correlated factors than CFA solutions (e.g., Chiorri et al., 2016; Guay et al., 2015). This, in turn, both improves the discriminant validity of the factors and provides a more accurate representation of the data (Morin et al., 2013). In many cases,

ESEM is viewed as primarily confirmatory in its approach (Marsh et al., 2014); that is, through the use of target rotation, researchers using ESEM are able to make *a priori* predictions about an expected factor structure (i.e., researchers are able to specify which items are pure measures of a factor while allowing for cross-loadings on other items). However, ESEM can also be used as an exploratory tool (with an oblique geomin rotation), where all items are specified to load on all factors (Asparouhov & Muthén, 2009).

ESEM may be particularly useful in test adaptation studies when dealing with (potentially) multidimensional constructs. Let us briefly return to our example of the fictional QBI. Assume for instance that the factor structure of the QBI has been found to be multidimensional in a number of national contexts, but that there are various multidimensional models available (e.g., a 2- and 3-factor model). In this scenario, ESEM may be a good alternative to the EFA-to-CFA method, as it would allow for testing of the competing models in a less restrictive manner compared to CFA. However, there may be some scenarios where ESEM can be problematic: ESEM is less useful when dealing with large, complex models and when sample sizes are small-to-moderate, because in these situations ESEM-derived models may lack parsimony (Marsh et al., 2014; but see Marsh et al., 2020, who have developed set-ESEM as a way of dealing with these concerns).

ESEM may also be best used in cases where there is at least some minimal theory that provides a basis for hypothesis-driven testing. In terms of our fictional QBI, it is the availability of various hypothesised models in the earlier literature that makes ESEM a particularly useful method of analysis. Nevertheless, given that researchers typically discard items that cross-load following EFA (i.e., the data-driven, hypothesised model that researchers end up with following EFA produces is typically one without cross-loading items), we see both the ESEM and the EFA-to-CFA approaches as viable and alternative analytic strategies for test adaptation. The decision to use ESEM versus EFA-to-CFA is

ultimately one that needs to be made based on the richness of previous work and available theorising. Conversely, an EFA-to-CFA strategy should be favoured when both extant theorising and current hypothesising point to a unidimensional factor structure.

2.10. How should global multidimensionality be modelled in test adaptation studies?

When dealing with multidimensional constructs, there are a number of ways in which global multidimensionality (i.e., how lower-order constructs co-exist with a global construct) in particular can be modelled. In body image research, the most common method is to model global multidimensionality using a higher-order factor model. In these models, it is assumed that items load onto their respective lower-order factors, which in turn load onto a general factor (see Figure 1). Examples of this include the suggestion that the two lower-order facets of drive for muscularity (i.e., attitudinal and behavioural facets) as measured using the Drive for Muscularity Scale (McCreary & Sasse, 2000) load onto a higher drive for muscularity factor (McCreary et al., 2004) and that the lower-order facets of acceptance of cosmetic surgery (i.e., Intrapersonal, Social, and Consider) as measured using the Acceptance of Cosmetic Surgery Scale load onto a higher-order attitudinal dimension (Henderson-King & Henderson-King, 2005).

In these higher-order models, it is assumed that associations between indicators and the higher-order factors are indirect; that is, that the associations are mediated by the lower-order factors. It is also assumed that associations between the indicators and the unique part of the first-order factor are mediated by the lower-order factor (for discussions, see Brunet et al., 2016; Gignac, 2016). However, there is growing interest in bifactor models, which may provide more realistic representations of multidimensional associations compared to higher-order models (Morin et al., 2016, 2020; Reise, 2012). In bifactor models, items are allowed to define a global G-factor and one specific S-factor, with all S-factors specified as orthogonal to one another and in relation to the G-factor (Gignac, 2016; Morin et al., 2016; Reise, 2012).

This method allows for the total item covariance matrix to be separated into a global component that explains the variance shared among responses to all items, and specific factors that explain the covariance associated with items subsets not already explained by the global component (Morin et al., 2016, 2020; Rodriguez et al., 2016; see Figure 2).

Bifactor analyses have become increasingly popular, as have bifactor-ESEM analyses (Marsh et al., 2013, 2014; Morin et al., 2013), but their application to conceptualisations of body image research remains limited (for exceptions, see Maïano et al., 2021; Meadows & Higgs, 2020). Historically, it may be suggested that scholars have relied on higher-order models of body image constructs "by default"; that is, body image researchers may not have considered bifactor models of body image constructs simply because of the relative invisibility of bifactor analyses in this area of research. However, the bifactor model offers scholars wider applicability and less ambiguous conceptual accounts of constructs compared with higher-order factor models (for reviews, see Chen et al., 2007; Morin et al., 2016, 2020; Reise, 2012). This, in turn, raises difficult questions for scholars working on test adaptation. For example, when dealing with multidimensional constructs of body image that have been previously modelled as higher-order factors, should scholars working on test adaptation assess bifactor models instead (or in addition to higher-order factor models)?

When there are good theoretical reasons to expect a bifactor model, and given some of the benefits of higher-order factor models (Bornovalova et al., 2020; Chen et al., 2007; Morin et al., 2016, 2020; Reise, 2012), we suggest that there may be value in assessing the fit of bifactor models – perhaps alongside higher-order factor models. However, researchers who are intent on applying bifactor models in the context of test adaptation need to be aware of a number of issues. First, applying bifactor models in test adaptation studies will raise problems comparing test adaptation results with parent studies, though doing so may also provide scholars with the impetus to re-assess earlier models of multidimensional body image

constructs. Second, because bifactor models absorb as much item variance as possible into global or specific factors, there is a risk that such models will be over-fitted (Bonifay & Cai, 2017; Bonifay et al., 2017). In a similar vein, there may also be some situations when S-factors are difficult to interpret: unlike lower-order factors (which retain their substantive meaning and, together, define the meaning of their higher-order factors), S-factors are essentially unrelated to the G-factor and all other S-factors, which may not make theoretical sense when dealing with body image constructs. Finally, there is a likelihood that the pattern of factor loadings that define global factors may be unstable across different samples (for a discussion, see Giordano & Waller, 2020). The latter can be particularly problematic in test adaptation work because unstable global factors can make it difficult to understand the meaning of such factors across national or linguistic groups.

However, given that bifactor analyses have rarely been applied in body image research, we suggest that some of these issues need to be examined empirically before firm recommendations can be drawn. Certainly, we recommend that body image scholars – whether working on test adaptation or not – should develop greater familiarity with bifactor models and develop theoretical understandings of multidimensional body image constructs accordingly. In that sense, we are minded to repeat Box and Draper's (1987) adage that "all models are wrong, but some are useful." All models are wrong because they merely represent simplifications of constructs. However, there is now growing evidence that bifactor models offer a useful way of conceptualising and understanding multidimensional constructs. At a minimum, an awareness of bifactor modelling is important when considering multidimensional body image constructs, even if scholars continue to show a preference for higher-order factor models. More broadly, test adaptation studies are an important arena in which an understanding of bifactor models can begin to shape body image research more

generally, though we also feel it is important to mention Decker's (2021, p. 39) useful advice here: "Don't use a bifactor model unless you believe the true structure is bifactor."

2.11. When is it necessary to determine measurement invariance and what method should be used to do so?

A prerequisite of any meaningful comparison of latent scores across groups is ensuring that scores on the instrument are functioning in a similar way (i.e., capturing the same construct) across those groups (Vandenberg & Lance, 2000). If the measurement properties of an instrument fundamentally differ across two groups, then measurement biases could occur, leading to artefactual results (Chen, 2008). To that end, assessment of measurement invariance is a necessary step in any instance in which researchers wish to compare latent scores across groups. Examples of such groups include those stratified by nationality, gender, developmental stages (e.g., early adulthood versus late adulthood), race, and sexual orientation. A less common example is to examine the same individuals across different time-points (e.g., the experience of having a procedure that alters physical appearance, such as a cosmetic surgery or a mastectomy, could fundamentally change an individual's understanding and experience of body image).

Measurement invariance can be determined at different levels, including configural, metric, scalar, and strict levels of invariance (for definitions and cut-off indices, see Swami & Barron, 2019), with scalar invariance typically considered a minimum threshold for comparison of latent means (Chen, 2007; Putnick & Bornstein, 2016). If scalar invariance is not achieved, response bias at the item level (i.e., *differential item functioning* or DIF) may be suspected (Zumbo, 2003). In this case, a partially invariant model may be identified by fixing the intercept for one or more items. In a partially invariant model, individuals with the same score on the latent variable achieve different scores on a differential item. Evidence of partial

scalar invariance may be sufficient for comparison of latent means (Vandenberg & Lance, 2000).

Assessment of measurement invariance is typically computed using either multi-group CFA (Chen, 2007, 2008) or multiple-indicators multiple-causes (MIMIC) modelling (Muthén, 1989). Each method has associated strengths and weaknesses (for a discussion, see Bauer, 2017). For categorical group variables (i.e., where clearly demarcated groups can be discerned, such as with comparisons across national groups), we recommend assessing measurement invariance using multi-group CFA. In multi-group CFA, the data from each categorical group are concurrently fitted to parallel models. As such, model parameters can be constrained to test increasingly restrictive models (e.g., to examine configural, metric, scalar, strict, and partial invariance). However, multigroup CFA does not accommodate continuous variables, such as age or body mass index. To examine continuous variables using multigroup CFA, it is first necessary to transform the data into discrete categories, but this may cause biased results or information loss (MacCallum et al., 2002).

For researchers wishing to examine the effects of continuous variables, we recommend the use of MIMIC models (Muthén, 1989). In addition to accommodating continuous variables, MIMIC models are also advantageous in that latent factors can be regressed onto multiple predictors (which may include continuous covariates). To that end, MIMIC modelling can be used to evaluate DIF for multiple traits concurrently and can facilitate the identification of main effects and interactions (e.g., age, gender identity, age x gender identity; Woods, 2009). A further advantage of MIMIC modelling is that it may have greater power than multi-group CFA to detect DIF with smaller sample sizes (Muthén, 1989; Woods, 2009).

However, there are also several disadvantages associated with MIMIC modelling, which lead us to recommend preferential use of the multi-group CFA approach for

researchers investigating categorical variables. Briefly, because MIMIC models are applied to the entire population (rather than specifying a model for each group, as in multi-group CFA), MIMIC models presume configural invariance is met, which may not always be the case. In addition, while multi-group CFA allows for different factor loadings across two groups, factor loadings in MIMIC models are presumed to be constant, which means that non-uniform DIF cannot be identified using MIMIC modelling (Bauer, 2017). Given the relative strengths and weaknesses of the multi-group CFA and MIMIC approaches, some researchers have used a combination of the two methods (Tóth-Király et al., 2017) or alternative approaches such as Moderated Nonlinear Factor Analysis (Bauer, 2017), which interested readers may wish to explore.

2.12. Are test adaptation issues only of relevance to studies with different national or linguistic groups?

In short, no. Test adaptation methods can be used anytime scholars wish to examine the equivalence of an instrument in any social identity group, such as racial and sexual minority groups within a nation. Where researchers have determined that a translation of an instrument is needed for a new social identity group, we recommend following the guidelines reported in Swami and Barron (2019) and supplemented above. Where a translation is not needed, researchers may proceed to an examination of measurement equivalence, while bearing in mind different analytic options that are available. For instance, in cases where there are good theoretical reasons to be believe that the factorial validity of scores on instrument are likely to be divergent in a new social identity group, scholars may wish to begin by using an EFA-to-CFA or ESEM approach. On the other hand, where theory points to factor structure equivalence, researchers may find it more useful to proceed to an examination of measurement invariance across social identity groups. In both cases, it is important that all decision-making is theoretically-driven and justified.

2.13. Are test adaptation issues only of relevance to body image instruments?

We are being slightly facetious with this question: clearly, the issues we have discussed above and those covered in more detail in Swami and Barron (2019) are relevant when translating and validating a wide range of instruments assessing various constructs in the social sciences. While we have focused on issues directly relevant to body image given the focus of this journal, the same guidelines and recommendations apply equally to all manner of instruments, including those related to body image constructs (e.g., eating behaviours, disordered eating, weight stigma, self-objectification) and those that are not as closely aligned with body image research. Indeed, we encourage scholars from across disciplines in the social sciences not only to pay closer attention to issues of test adaptation, but also to persuade journal editors in various disciplines to provide adequate space for the reporting of test adaptation studies. Test adaptation studies remain an important component of research in the social sciences precisely because they are an important step in ensuring that historically marginalised populations are represented in ongoing research. When editors actively discourage test adaptation studies or refuse to consider such studies for publication, they do a disservice to the entire field (Ziegler, 2021).

3. Conclusion

Test adaptation studies play a crucial role in body image, particularly in terms of understanding the way hypothesised constructs are understood and experienced in different social identity, linguistic, national, and cultural groups. More broadly, test adaptation studies can ensure that understandings of body image are not constricted to particular identities and are instead inclusive and respectful of differences within and across groups. Body image researchers have a proud track record of giving voice to often marginalised groups but, to ensure that we keep doing so, researchers need to immerse themselves in developments in

test adaptation theory. Indeed, as theory and practice continue to evolve, it is important that body image scholars develop a fuller awareness of, and keep up-to-date with, test adaptation methods. Having an understanding of these issues will help ensure that *Body Image* continues to enjoy success in its mission of considering and understanding issues dealing with "body image in diverse cultural contexts" (Cash, 2004, p. 3).

References

Andersen, N., & Swami, V. (2021). *A bibliometric review of publications in Body Image, 2004 to 2020: Science mapping research on body image.* Manuscript submitted for publication.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*(3), 397-438. https://doi.org/10.1080/1075510903008204

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management*, *41*(6), 1561-1577. https://doi.org/10.1177/0149206315591075

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468-491. https://doi.org/10.1037/met0000200

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507-526. https://doi.org/10.1037/met0000077

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143-156. https://doi.org/10.1509/jmkr.38.2.143.18840

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). Response biases in marketing research. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research: Uses, misuses, and future advances* (pp. 95-109). Sage.

Baumgartner, H., & Weijters, B. (2015). Response biases in cross-cultural management. In S. Ng & A. Y. Lee (Eds.), *Handbook of culture and consumer behavior* (pp. 150-180). Oxford University Press.

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.

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, *20*(6), 573-584. https://doi.org/10.1080/13645579.2016.1252188

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465-484. https://doi.org/10.1080/00273171.2017.1309262

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184-186. https://doi.org/10.1177/2167702616657069

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18-27. https://doi.org/10.1016/j.biopsych.2020.01.013

Box, G., & Draper, N. R. 1987). *Empirical model-building and response surfaces*. Wiley.

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

Brunet, J., Gunnel, K. E., Teixeira, P., Sabiston, C. M., & Bélanger, M. (2016). Should we be looking at the forest or the trees? Overall psychological need satisfaction and individual needs as predictors of physical activity. *Journal of Sport and Exercise Psychology*, *38*(4), 317-330. https://doi.org/10.1123/jsep.2016-0256

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105. https://doi.org/10.1037/h0046016

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

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

Cash, T. F., Melnyk, S. E., & Hrabosky, J. I. (2004). The assessment of body image investment: An extensive revision of the Appearance Schemas Inventory. *International Journal of Eating Disorders*, *35*(3), 305-316. https://doi.org/10.1002/eat.10264

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. https://doi.org/10.1111/j.1365-2648-2007.04242.x

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

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*(5), 1005-1018. https://doi.org/10.1037/a0013193

Chiorri, C., Marsh, H. W., Ubbiali, A., & Donati, D. (2016). Testing the factor structure and measurement invariance across gender of the Big Five Inventory through exploratory structural equation modeling. *Journal of Personality Assessment*, *98*(1), 88-99. https://doi.org/10.1080/00223891.2015.1035381

Clark, D. A. & Bowles, R. P. (2018) Model fit and item factor analysis: Overfactoring, underfactoring, and a program to guide interpretation. *Multivariate Behavioral Research*, *53*(4), 544-558. https://doi.org/10.1080/00273171.2018.1461058

Colina, S., Marrone, N., Ingram, M., & Sánchez, D. (2017). Translation quality assessment in health research: A functionalist alternative to back-translation. *Evaluation and the Health Professions*, *40*(3), 267-293. https://doi.org/10.1177/0163278716648191

Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*(4), 407-422. https://doi.org/10.1177/002224378001700401

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302. https://doi.org/10.1037/h0040957

Cummins, R. A., & Gullone, E. (2000). Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. In *Proceedings, Second International Conference on Quality of Life in Cities* (pp.74-93). National University of Singapore.

Decker, S. L. (2021). Don't use a bifactor model unless you believe the true structure is bifactor. *Journal of Psychoeducational Assessment*, *39*(1), 39-49. https://doi.org/10.1177/0734282920977718

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research*, *8*(2), 181-192. https://doi.org/10.1093/her/8.2.181

Dignard, N. A. L., & Jarry, J. L. (2019). The Body Appreciation Scale-2: Item interpretation and sensitivity to priming. *Body Image*, *28*, 16-24. https://doi.org/10.1016/j.bodyim.2018.10.005

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*(3), 362-388. https://doi.org/10.1080/00273170902938969

Douglas, S. P., & Craig, C. S. (2007). Collaborative and iterative translation: An alternative approach to back translation. *Journal of International Marketing*, *15*(1), 30-43. https://doi.org/10.1509/jimk.15.1.030

Finch, W. H. (2020). Using fit statistic differences to determine the optimal number of factors to retain in exploratory factor analysis. *Educational and Psychological Measurement*, *80*(2), 217-241. https://doi.org/10.1177/0013164419865769

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*(1), 93-111. https://doi.org/10.1037/met000064

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, *55*, 57-68. https://doi.org/10.1016/j.intell.2016.01.006

Giordano, C., & Waller, N. G. (2020). Recovering bifactor models: A comparison of seven methods. *Psychological Methods*, *25*(2), 143-156. https://doi.org/10.1037/met0000227

Guay, F., Morin, A. J. S., Litalien, D., Valois, P., & Vallerand, R. J. (2015). Application of exploratory structural equation modeling to evaluate the Academic Motivation Scale. *The Journal of Experimental Education*, *83*(1), 51–82. https://doi.org/10.1080/00220973.2013.876231

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences*, *44*(4), 932-942. https://doi.org/10.1016/j.paid.2007.10.034

Henderson-King, D., & Henderson-King, E. (2005). Acceptance of cosmetic surgery: Scale development and validation. *Body Image*, *2*(2), 137-149. https://doi.org/10.1016/j.bodyim.2005.03.003

Hernández, A., Hidalgo, M. D., Hambleton, R. K., & Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema*, *32*(3), 390-398. https://doi.org/10.7334/psicothema2019.306

International Test Commission. (2017). *ITC guidelines for translating and adapting tests* (2nd ed.). International Test Commission. https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory: Monograph supplement 9. *Psychological Reports*, *3*, 635-694. https://doi.org/10.2466/pr0.3.7.635-694

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19-40. https://doi.org/10.1037/1082-989X.7.1.19

Maïano, C., Morin, A. J. S., Aimé, A., Lepage, G., & Bouchard, S. (2021). Psychometric properties of the Body Checking Questionnaire (BCQ) and the Body Checking Cognitions Scale (BCCS): A bifactor exploratory structural equation modeling approach. *Assessment*, *28*(2), 632-646. https://doi.org/10.1177/1073191119858411

Maneesriwongul, W., & Dixon, J. K. (2004). Instrument translation process: A methods review. *Journal of Advanced Nursing*, *48*(2), 175-186. https://doi.org/10.1111/j.1365-2648.2004.03185.x

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM) and Set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research*, *55*(1), 102-119. https://doi.org/10.1080/00273171.2019.1602503

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(3), 439-476. https://doi.org/10.1080/10705510903008220.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, *29*(4), 322-346. https://doi.org/10.1177/0734282911406657

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. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, *49*(6), 1194-1218. https://doi.org/10.1037/a0026913

McCabe, M., Tatangelo, G., Watson, B., Fuller-Tyszkiewicz, M., Rodgers, R. F., Aimé, A., & Ricciardelli, L. (2019). Development and testing of a model for risk and protective factors for eating disorders and higher weight among emerging adults: A study protocol. *Body Image*, *31*, 139-149. https://doi.org/10.1016/j.bodyim.2019.10.001

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*(6), 297-304. https://doi.org/10.1080/07448480009596271

McCreary, D. R., Sasse, D. K., Saucier, D. M., & Dorsch, K. D. (2004). Measuring the drive for muscularity: Factorial validity of the Drive for Muscularity Scale in men and women. *Psychology of Men and Masculinity*, *5*(1), 49-58. https://doi.org/10.1037/1524-9220.5.1.49

Meadows, A., & Higgs, S. (2020). A bifactor analysis of the Weight Bias Internalization Scale: What are we really measuring? *Body Image*, *33*, 137-151. https://doi.org/10.1016/j.bodyim.2020.02.013

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23. https://doi.org/10.3102/0013189X023002013

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/0013164420942899

Morin, A. J. S., Arens, A., & Marsh, H. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*(1), 116-139. https://doi.org/10.1080/10705511.2014.961800

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: a second course* (pp. 395-436). Information Age Publishing, Inc.

Morin, A. J. S., Myers, N. D., & Lee, S. M. (2020). Modern factor analytic techniques: Bifactors models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (4th ed., pp. 1044-1073). Wiley. https://doi.org/10.1002/9781119568124.ch51

Muthén, B. O. (1989). Latent variable modelling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585. https://doi.org/10.1007/BF02296397

Nunnally, J. (1978). *Psychometric theory*. McGraw-Hill.

Ozolins, U., Hale, S., Cheng, X., Hyatt, A., & Schofield, P. (2020). Translation and back-translation methodology in health research: A critique. *Expert Review of Pharmacoeconomics and Outcomes Research*, *20*(1), 69-77. https://doi.org/10.1080/14737167.2020.1734453

Perneger, T. V., Leplège, A., & Etter, J.-F. (1999). Cross-cultural adaptation of a psychometric instrument: Two methods compared. *Journal of Clinical Epidemiology*, *52*(11), 1037-1046. https://doi.org/10.1016/S0895-4356(99)00088-8

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*(1), 13-43.

Preacher, K. J., Zhang, G., Kim, C. & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28-56. https://doi.org/10.1080/00273171.2012.710386

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1-15. https://doi.org/10.1016/S0001-6918(99)00050-5

Putnick, D. K., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696. https://doi.org/10.1080/00273171.2012.715555

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223-237. https://doi.org/10.1080/00223891.2015.1089249

Russell, C., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology*, *77*(3), 336-342. https://doi.org/10.1037/0021-9010.77.3.336

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, *37*(4), 189-199. https://doi.org/10.3102/0013189X08319569

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2009). Theory of Test Translation Error. *International Journal of Testing*, *9*(2), 78-91. https://doi.org/10.1080/15305050902880835

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the Theory of Test Translation Error. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on the PISA Research Conference 2009* (pp. 71-85). Springer Verlag.

Swami, V., & Barron, D. (2019). Translation and validation of body image instruments: Challenges, good practice guidelines, and reporting recommendations for test adaptation. *Body Image*, *31*, 204-220. https://doi.org/10.1016/j.bodyim.2018.08.014

Swami, V., Özgen, L., Gökçen, E., & Petrides, K. V. (2015). Body image among female university students in Turkey: Concurrent translation and validation of three body image measures. *International Journal of Culture and Mental Health*, *8*(2), 176-191. https://doi.org/10.1080/17542863.2014.917117

Tóth-Király, I., Bõthe, B., Rigó, A., & Orosz, G. (2017). An illustration of the Exploratory Structural Equation Modeling (ESEM) framework on the Passion Scale. *Frontiers in Psychology*, *8*, 1968. https://doi.org/10.3389/fpsyg.2017.01968

Tylka, T. L., Alleva, J. M., Calogero, R. M., Fuller-Tyszkiewicz, M., Jackson, T., Murnen, S., Murray, S. B., Rodgers, R. F., Swami, V., & Webb, J. B. (2020). Editor's response to Clarivate Analytic's decision to suppress *Body Image* from receiving a 2019 impact factor. *Body Image*, *34*, iii-v. https://doi.org/10.1016/j.bodyim.2020.08.001

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*(2), 135-147. https://doi.org/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*(1), 4-70. https://doi.org/10.1177/109442810031002

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The effects of familiarity with the response category labels on item response to Likert scales. *Journal of Consumer Research*, *40*(2), 368-381. https://doi.org/10.1086/670394

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, *15*(1), 96-110. https://doi.org/10.1037/a0018721

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 Outcomes (PRO) measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*(2), 94-104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Wild, D., Eremenco, S., Mear, I., Martin, M., Houchin, Gawlicki, M., Harrendran, A., Wiklund, I., Chong, L. Y., von Maltzahn, R., Cohen, L., & Molsen, E. (2008). Multinational trials-recommendations on the translation required, approaches to using the same language in different countries, and the approaches to support pooling the data: The ISPOR Patient Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. *Value in Health*, *12*(4), 430-440. https://doi.org/10.1111/j.1524-4733.2008.00471.x

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1-27. https://doi.org/10.1080/00273170802620121

Worthington, R., & Whittaker, T. (2006). Scale development research: A content analysis and recommendations for best practice. *Counseling Psychologist*, *34*(6), 806-838. https://doi.org/10.1177/0011000006288127

Zhao, X., & Solano-Flores, G. (2021). Testing across languages in international comparisons: Cultural adaptation of consensus-based test translation review procedures. *Journal of Multilingual and Multicultural Development*. Advanced online publication. https://doi.org/10.1080/01434632.2020.1852242

Zhao, X., Solano-Flores, G., & Qian, M. (2018). International test comparisons: Reviewing translation error in difference source language-target language combinations. *International Multilingual Research Journal*, *12*(1), 17-27. https://10.1080/19313152.2017.1349527

Ziegler, M. (2021). Psychological test adaptation and development – A new beginning. *Psychological Test Adaptation and Development*. Advance online publication. https://doi.org/10.1027/2698-1866/a000001

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*(2), 136-147. https://doi.org/10.1191/0265532203lt248oa

Figures

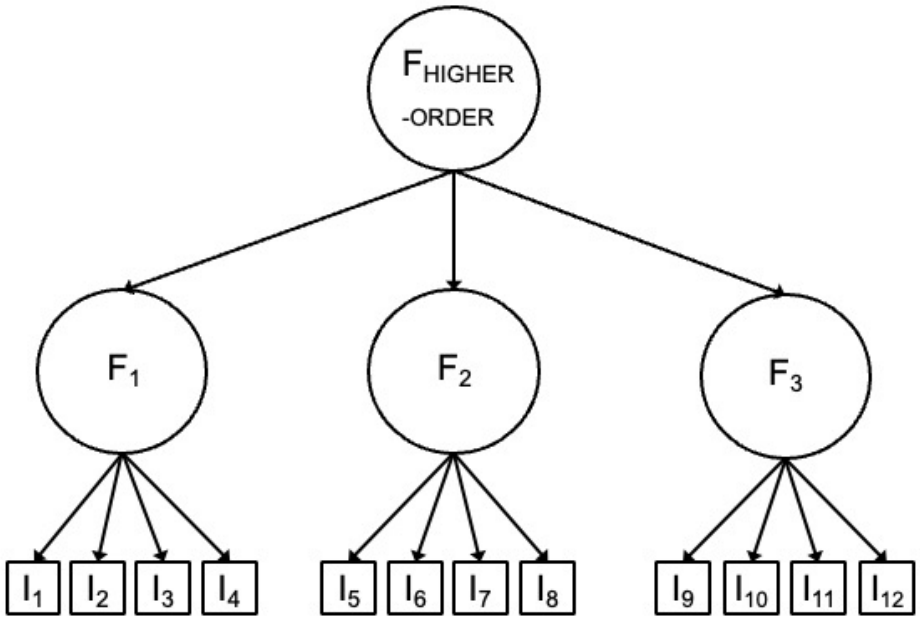

Figure 1. Example of a higher-order factor model in which items (I) load onto three factors (F), which in turn load onto a higher-order factor. Residual variances omitted for clarity.

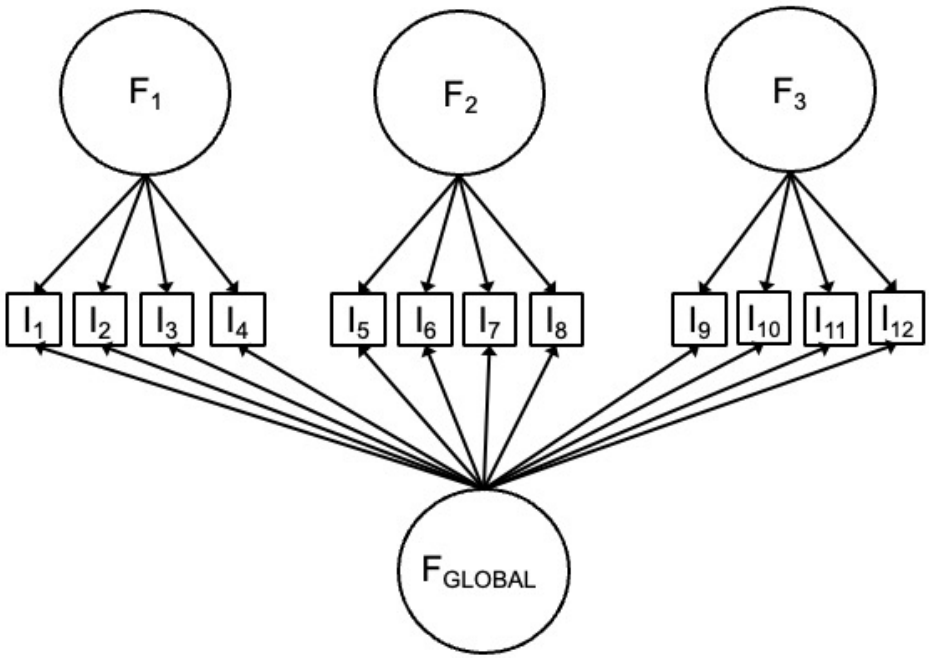

Figure 2. Example of an orthogonal bifactor model where items (I) load onto specific factors (F) and a global factor. Residual variances omitted for clarity.