Assessing Model Fit: Caveats and Recommendations for Confirmatory Factor Analysis and Exploratory Structural Equation Modeling

John L. Perry

Leeds Trinity University

Adam R. Nicholls

University of Hull

Peter J. Clough

University of Hull

Lee Crust

University of Lincoln

Author Note. John L. Perry is with the Department of Sport, Health, and Nutrition, Leeds Trinity University, Leeds, LS18 5HD. Adam R. Nicholls and Peter J. Clough are with the Department of Psychology, University of Hull, Hull, HU6 7RX, UK. Lee Crust is with the School of Sport and Exercise Sciences, University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, UK. Correspondence concerning this article should be addressed to John Perry, e-mail: j.perry@leedstrinity.ac.uk

Abstract

Confirmatory factor analysis (CFA) is commonly used to assess measurement models in sport and exercise psychology. Frequently used as a yardstick for their adequacy, are specific cutoff values proposed by Hu and Bentler (1999). The purpose of this study was to investigate the appropriateness of using the CFA approach with these cutoff values for typical multidimensional measures. Further, we sought to examine how a model could be respecified to achieve acceptable fit, and explored whether exploratory structural equation modeling (ESEM) provides a more appropriate assessment of model fit. Eight measures commonly used in sport and exercise psychology research were examined using CFA and ESEM. Despite demonstrating good validity in previous research, all eight failed to meet the cutoff values proposed by Hu and Bentler. ESEM improved model fit in all multidimensional measures. In conclusion, we suggest that model misfit in this study demonstrates the inadequacy of using CFA cutoff values. Further, we recommend ESEM as a preferred approach to examining model fit in multidimensional measures.

Keywords: confirmatory factor analysis, exploratory structural equation modeling, modification indices

Assessing Model Fit: Caveats and Recommendations for Confirmatory Factor Analysis and Exploratory Structural Equation Modeling

Jöreskog (1969) developed confirmatory factor analysis (CFA) to examine psychometric models and the use of CFA has risen exponentially in recent time. Searches on SPORTdiscus revealed that 180 papers employing CFA techniques were published from 1990-1999, compared to 549 papers from 2000-2009. In part, this is due to the expansion of structural equation modeling methods that firstly require the researcher to obtain a satisfactory measurement of model fit before proceeding to the main analysis. This use has added to the more traditional approach of using CFA purely to examine the factorial validity of a measure.

Theoretically, CFA represents an objective test of a theoretical model. In practice, conducting all factor analytic procedures requires a series of judgments. By far the most important judgment made in CFA is whether a model is deemed to be acceptable or not. Logically, the process of accepting or rejecting models is fairly simple, in that the aim is to avoid concluding that a good model is bad, and that a bad model is good (MaCallum, Browne, & Sugawara, 1996). This is typically achieved by examining the absence or presence of misspecifications, which are errors between the prescribed model and the estimated parameters. In structural equation modeling, of which CFA is one form, the goodness of a model is typically determined by the absence (good) or presence (bad) of misspecifications (Saris, Satorra, & van der Veld, 2009). The clearest of all the parameters for making judgments on the acceptability of model fit is the chi-square (χ^2). However, as initially observed by Bentler and Bonett (1980) and many thereafter (e.g., Saris et al.), because this statistic is sensitive to sample size, if will reject models that have only a trivial misspecification, thus leading to increased type II error. The solution appears to be to use a selection of fit indices that calculate exact model fit based on chi-square (e.g., standardized root mean square residual or goodness of fit index), relative fit indices that compare the hypothesized model to an independent baseline model (e.g., Tucker-Lewis index or incremental fit index), and noncentrality-based indices that test the alternative hypothesis rather than the null (e.g., Bentler's comparative fit index or the root mean square error of approximation).

Hu and Bentler (1999) proposed cutoff criteria for all commonly cited fit indices by examining rejection rates on hypothetical models. These proposed criteria are referred to as a matter of routine in studies using any kind of structural equation methods. While reference to Hu and Bentler's suggested cutoffs is not necessarily an issue itself, the extent to which many researchers view these recommendations as golden rules potentially creates an substantial amount of type one errors. Marsh, Hau, and Wen (2004) keenly and accurately point out that Hu and Bentler offered caution about using such cutoff values and concisely explain the dangers of overgeneralizing the findings from Hu and Bentler in search of golden rules. Indeed, Marsh et al. refer to a traditional cutoff values amounting to "little more than rules of thumb based largely on intuition and have little theoretical justification" (p. 321). It is surprising therefore that such cutoff values are blindly accepted so regularly in sport and exercise psychology without even acknowledgement of their limitations.

The use of CFA techniques for examining factorial validity and identifying acceptable levels of fit is certainly not straightforward. Hopwood and Donnellan (2010) illustrated the difficulty very effectively by examining eight common personality measurements. By conducting CFAs, the authors found that none of the scales used came close to Hu and Bentler's recommended cutoff values. Interestingly, even the best performing measure achieved a model fit well below the commonly accepted criteria, despite commonly being accepted as an appropriate assessment of personality. The length and complexity of personality measures means that employing the same requirements of such models compared to short, simple models is simply not appropriate. A CFA model typically constrains items to loading on only one factor, resulting in misspecification for each cross-loading. Long, complex measures therefore, have much less chance of achieving an acceptable fit. In providing their own caveat for using CFA, Hopwood and Donnellan describe what they call *The Henny Penny Problem* after the character from the children's tale who lamented that the sky was falling after an acorn fell on his head. The authors point out that claims that a measure is invalid because of a weak CFA fit is exaggerated and ignores other types of validity such as content and criterion-related validity. Such personality assessments could perhaps perform better in a CFA by reducing their size and/or complexity, but if this is at cost of predictive or other forms of validity, it is simply not a virtuous academic pursuit.

When encountering misspecifications in a CFA model, the researcher has several options. They can either (a) determine that the misspecification is irrelevant and proceed, (b) concede that the misspecification is significantly relevant and therefore reject the model, or (c) modify the model to achieve an acceptable fit. Such modification can be achieved using the modification indices provided in CFA output. The modification indices (MI) provide an estimate increase in the chi-square for each fixed parameter if it were to be freed. In independent cluster models (ICM; Marsh et al., 2009), covariances between items from questionnaires are typically fixed to zero. By identifying significant modification indices and allowing them to be estimated, chi-square will be increased, thus yielded a better statistical model fit. The use of MI to respecify poorly fitting models was effectively demonstrated by MacCullum (1986) and further recommended by Saris, den Ronden, and Satorra (1987) and Saris et al. (2009). It should be noted however, that all of these authors also urge caution because this data driven approach does not necessarily hold any theoretical relevance. Indeed, MacCullum found that in half of the models tested in a simulation study, MI did not find a true model. Several authors (e.g., Brown, 2006; Kaplan, 2009; Kline, 2005) have referred to such respecification as atheoretical, claiming that it is merely capitalizing on chance within a sample. The process of using MI is seldom reported and therefore presumably, seldom conducted in sport and exercise psychology.

ESEM provides an alternative to CFA, which is effectively an integration of exploratory factor analysis (EFA) and CFA methods. CFA assesses an a priori model that typically allows observed variables to load only onto their intended factor. Typically, all loadings, regardless of their significance, onto other latent variables are constrained to zero (Figure 1). In Figure 1, *y* represents the latent variables, which are typically subscales in self-report psychology measures, while *x* represents each observed variable, typically an item within a questionnaire, and *e* represents the residual error. This is a typical CFA model, often referred to as an ICM (Marsh et al., 2009). This means that all non-significant cross-loadings will contribute to model misspecification (Ashton & Lee, 2007). This misspecification is defined by Hu and Bentler (1998, p. 427) as when “one or more parameters are fixed to zero were population values are non-zeros (i.e., an underparameterized misspecified model)". Clearly in many psychometric measures, particularly long, multidimensional scales, this can become a substantial issue. Moreover, questionnaires that are aggregated to enable an overall score to be derived as well as individual subscale scores to include appropriate internal consistency must have moderate to high inter-correlations and therefore, many non-zero cross-loadings. Church and Burke (1994) explained that ICMs are too restrictive for research where secondary or cross-loadings are likely, such as personality research. It is this reason why Hopwood and Donnellan (2010), and others before them, found such difficulty in obtaining a satisfactory CFA fit on personality scales. ESEM provides standard errors for all rotated parameters. As such, it allows all observed variables to load on all latent variables (Figure 2). This overcomes the issue of secondary, often non-significant cross-loadings causing irrelevant model misspecification, and therefore, the potential rejection of a good model. This was expertly demonstrated by Marsh et al. (2010), who assessed the 60-item NEO Five-Factor Inventory using CFA and ESEM methods. The authors found that ESEM noticeably outperformed CFA in goodness of fit and construct validity.

Given the exponential rise in the use of CFA, it is crucial to examine the potential limitations of the technique. The purpose of this study was to firstly assess the likelihood that common quantitative measures in sport and exercise psychology can meet the cutoff values proposed by Hu and Bentler (1999) with independent samples. Secondly, we tested the extent to which manipulation of the model according to modification indices was a valid approach to achieving model fit. Thirdly, we conducted ESEM on all multidimensional scales to examine if this is likely to be a preferred alternative to CFA. We hypothesized that the majority of measurement scales used in the study would fall below the cutoff values proposed by Hu and Bentler (1999) and all chi-square values would suggest model misfit (i.e., < .001). We also hypothesized that while modification indices would significantly improve model fit, it would not be clear whether approach is merely sample-specific data manipulation. Finally, we hypothesized that ESEM would provide a better model fit on all measurement scales, proportional to the amount of factors and whether the factors provide an aggregated score.

Methods

We collated data from using eight commonly used psychometric scales in sport and exercise psychology. The measures were selected to represent a range of complexities in terms of the number of items (10-48) and factors (1-10). The measures also represent a variety of interrelationships between subscales, where some have highly correlated subscales and others have relatively independent subscales. Participant information for each scale used is displayed in Table 1. All samples were gathered using athletes from a range of individual and team sports following ethical approval from a UK-based higher education institution.

Measures

Coping Inventory for Competitive Sport (CICS; Gaudreau & Blondin, 2002). The CICS examines 10 coping subscales using 39 items requiring a response on a five-point Likert-type scale anchored from 1 = *Does not correspond at all to what I did or thought* to 5 = *Corresponds very strongly to what I did or what I thought*. For the purposes of this study, the CICS was only considered as a 10-factor model and hierarchical models were not assessed. Gaudreau and Blondin presented an acceptable CFA fit when the CICS was published, also demonstrating sufficient concurrent and divergent validity. Fletcher (2008) examined the psychometric properties of the CICS over a 10-week period, concluding that the measure is strong, obtaining meaningful and interpretable data.

Stress Appraisal Measure (SAM; Peacock & Wong, 1990). The SAM is contains seven subscales with 28-items items in total requiring a response on a five-point Likert-type scale anchored from 0 = *Not at all* to 5 = *Extremely*. At the time of publication, Peacock and Wong presented support for the internal consistency and construct validity of the SAM.

Mental Toughness Questionnaire-48 (MTQ48; Clough, Earle, & Sewell, 2002). The MTQ48 contains six subscales on 48-items items requiring a response on a five-point Likert-type scale from 1 = *Strongly disagree* to 5 = *Strongly agree*. Perry, Clough, Earle, Crust, and Nicholls (2013) found support for the factorial validity and reliability of the scale, with a sample of 8207 participants, adding to previous support for the criterion validity, which has associated higher mental toughness with pain tolerance (Crust & Clough, 2005), attendance at injury rehabilitation clinics (Levy, Polman, Clough, Marchant, & Earle, 2006), coping and optimism (Nicholls, Polman, Levy, & Backhouse, 2008), the use of psychological strategies (Crust & Azadi, 2010), and different managerial positions (Marchant, Polman, Clough, Jackson, Levy, & Nicholls, 2009).

Sport Motivation Scale-6 (SMS-6; Mallett, Kawabata, Newcombe, Otero-Forero, & Jackson, 2007). The SMS-6 assesses a six-factor model of sport motivation on 24 items requiring a response on a seven-point Likert-type scale from 1 = *Does not correspond at all* to 5 = *Corresponds exactly*. Mallett et al. claimed Improved model fit compared to its earlier incarnation (The sport motivation scale, Pelletier et al., 1995), the SMS-6 also demonstrated concurrent validity. More recently, Kawabata and Mallett (2013) provided further support for the discriminant validity of the SMS-6.

Sport Emotion Questionnaire (SEQ; Jones, Lane, Bray, Uphill, & Catlin, 2005). The SEQ examines five emotions using 22 items requiring a response on a five-point Likert-type scale from 0 = *Not at all* to 5 = *Extremely*. Participants are asked to indicate the extent to which they experience each emotion at the time of completing the SEQ. At the time of publication, Jones et al. demonstrated good model fit, concurrent and construct validity, and internal consistency.

Coping Self-Efficacy Scale (CSES; Chesney, Neilands, Chambers, Talyor, & Folkman, 2006). The CSES consists of 26 items and three subscales requiring a response on an 11-point Likert-type scale from 0 = *Cannot do at all* to 10 = *Certain can do*. In publishing the CSES, Chesney et al. present satisfactory model fit, concurrent validity, and internal consistency.

Connor-Davidson Resilience Scale 10 (CD-RISC 10; Campbell-Sills & Stein, 2007). The CD-RISC is a 10-item unidimensional scale, with its items being rated on a 5-point Likert-type scale. The questions are anchored at 0 = *not true at all* and 4 = *true nearly all of the time.* The CD-RISC has regularly demonstrated very good psychometric properties including during its translation into Spanish (Notario-Pacheco, Solera-Martínez, Serrano-Parra, Bartolomé-Gutiérrez, García-Campayo, & Martínez-Vizcaíno, 2011) and Chinese (Wang, Shi, Zhang, & Zhang, 2010).

General Self-efficacy Scale **(**GSE; Schwarzer & Jerusalem, 1995) The GSE is a 10-item unidimensional scale that assesses self-efficacy. Items of the GSE rated on a 4-point

Likert-type scale anchored at 1 = *Not at all True* and 4 = *Exactly True.* The internal consistency and construct validity of the GSE has been support in a host of countries (Luszczynska, Gutiérrez-Doña, & Schwarzer, 2005; Scholz, Gutiérrez-Doña, Sud, & Schwarzer, 2002) across the world.

Procedure

All data was collected using pen and paper method in the presence of researchers to ensure authenticity rather than using online data collection.

Data Analysis

Preliminary analysis checked for missing data and outliers before univariate skewness and kurtosis and multivariate kurtosis were examined. CFA was conducted on all measurement scales using Mplus 7.0 (Muthén & Muthén, 2012). Model fit was assessed using chi square (χ^2), the comparative fit index (CFI), the Tucker-Lewis index (TLI), standardized root mean residual (SRMR), and root mean squared error of approximation (RMSEA). Chi-square and SRMR represented absolute fit indices and CFI and TLI provided incremental indices, and RMSEA presented a parsimony-adjusted measure. All analyses used the robust maximum likelihood method (MLR) with epsilon value .05, and geomin rotation which is the default in Mplus.

To examine how easily fixed a model could be, we used modification indices to correlate observed variables until a better model fit was found, using an iterative process, as recommended by Oort (1998). In each analysis, all MI with a value > 10 in the “WITH”” statements were sequentially selected one at a time to enable observed variables to correlate. Oort demonstrated that the process should be iterative, whereby only one modification is made at once, as others may contain biases based on the existing structure. This enabled us to firstly assess if this generated an acceptable model fit. Secondly, if it did, we identified the amount of modifications required to achieve the fit. However, this begins to deviate from the intended theoretical design of the original model. To assess if this had deviated, we cross-validated our respecified model by testing model fit on two random halves of the original sample. If there was a clear difference (ΔCFI > .1) between the model fits, the modified model was deemed to have failed cross-validation.

For all multidimensional scales, ESEM was conducted, employing the same fit indices as CFA. As ESEM provides a more subjective overview and therefore, the model fit alone cannot be relied on without then examining the individual loadings. To assess this, we computed the proportion of items that loaded on intended factors, the number of significant cross-loadings, and the number of significant cross-loadings that were greater than the loading onto the intended factor.

Results

Confirmatory Factor Analyses

A summary of fit indices from the confirmatory factor analyses are displayed in Table 2. It is worth noting that of the eight measurement scales assessed; all chi-square statistics results were significant. Moreover, none of the measures achieved cutoff values for CFI and TLI of > .95, as recommended by Hu and Bentler (1999). Indeed, the SEQ was the only questionnaire to reach the sometimes applied more relaxed cutoff value of > .90 for CFI and TLI. While all met the recommended SRMR cutoff of < .08, only two of the eight achieved an RMSEA of < .05. With the exception of the CSES, all measures demonstrated a high proportion of items loading correctly onto their intended factor.

To examine whether these models could be ‘fixed’ using the modification indices, values of > .10 from the “WITH” statements were correlated as part of the model. The results of these modifications are displayed in Table 3. With modifications, all model fits improved significantly and achieved CFI and TLI > .90, SRMR < .06, and RMSEA < .06. All chi-square values remained significant. However, such modifications of course change the existing model and such a data-driven approach may yield sample-specific model fit rather than anything substantive. To partially examine this, all samples were randomly split in half and tested using the modified model. The results of this cross-validation are displayed in Table 4. For some measures, such as the CICS and SEQ, the modified model was successfully cross-validated, because no significant change in model fit was observed. For most of the measures, it appears that the use of the MI may deviate from the original model, though the extent to which this is theoretically substantial requires further investigation.

Exploratory Structural Equation Modeling

All multidimensional measurement scales presented significantly improved model fit using ESEM (see Table 5). On average, CFI increased by .082, TLI increased by .070, SRMR reduced by .032, and RMSEA reduced by .018. All chi-square significance values remain significant ($p < .001$).

As ESEM allows all observed variables to load onto all latent variables, it is important to examine the loadings of each item to assess whether they have loaded onto their intended factor. Further, cross-loadings should be checked, as significant cross-loadings or cross-loadings greater than the loading onto the intended factor represent a misspecification in the model. Approximately 90% of items loading onto their intended factor appears to be the norm, allowing for some cross-loadings. As expected, the only aggregated measure, the MTQ48, included a greater number of significant cross-loadings. Consequently, the increase in model fit for this measure between CFA and ESEM was greater.

Discussion

The purpose of this study was to (a) assess the likelihood that common quantitative measures in sport psychology can meet proposed cutoff values, (b) examine the extent to which a model can be reasonably respecified using the MI, and (c) evaluate the ability of ESEM to provide a more appropriate estimate of model fit than CFA.

The results suggest that Hu and Bentler's (1999) proposed, and commonly implemented, cutoff values for a host of fit indices are unrealistic for most measures to achieve on a sample independent from that which they were developed with. Consequently, we urge caution for researchers when employing the CFA technique. As a minimum, they should acknowledge the limitations of the approach and rigid cutoff values to prevent the "Henny Penny" problem described by Hopwood and Donnellan (2010). Those referring to Hu and Bentler's suggested cutoff values as golden rules when conducting CFA on complex, multidimensional models would be well advised to review the hypothetical models used in the original paper to establish such cutoffs. Hu and Bentler presented a simple model that contained 15 observed variables and three factors. Each factor had five loadings of .70 - .80 and all cross-loadings were fixed to zero. Further, they examined a 'complex' model that enabled just three cross-loadings across the same matrix. This is a long way from the complexity of many of the measures commonly used in sport and exercise psychology and another example of the dangers in overgeneralizing Hu and Bentler's (1999) findings, a topic discussed in much greater depth by Marsh et al. (2004).

The extent to which a misspecified model can be fixed remains contentious. In this study we have demonstrated, that from purely a statistical point of view, it is feasible to respecify the model using the MI. However, we urge caution when conducting this method, as all respecifications must be theoretical acceptable. This could be an acceptable approach as long as restrictions are placed on permissible modifications (MacCullum, 1986). Said differently, researchers should determine whether it is theoretically plausible for model respecification. An example might be freeing parameters between items within the same subscale, or perhaps creating a higher-order model that allows covariances between some subscale items that are theoretically related.

ESEM is an emerging technique that is used either supplementary with CFA or instead of CFA. There are several studies that utilize ESEM very effectively for the development and/or validation of a multidimensional measure outside of the sport domain (e.g., Marsh, Nagengast, Morin, Parada, Craven, & Hamilton, 2011; Marsh et al., 2010). In this study we have demonstrated that this technique is a desirable alternative to CFA using scales frequently used in a sport context. Other than rare exceptions (e.g., Morin & Maïno, 2011), the use of ESEM in the sport psychology literature is limited at present. We propose that in researchers could make a theoretical judgment on the appropriateness of the technique. For true ICMs where subscales within are measure are theoretically unrelated or even opposed, CFA should provide an accurate representation of the model fit. If encountering misspecifications, researchers may consider the use of MI to improve model fit but do so with caution, and be able to theoretically justify their respecifications. The vast majority of multidimensional scales in sport and exercise psychology however, are not true ICMs, because we can logically expect to find secondary loadings, particularly within highly correlated subscales or aggregated subscales. Under these circumstances, ESEM provides a more appropriate assessment of model fit than CFA and should be used from the outset.

The variety of measures examined in this paper, with the relatively large sample sizes is certainly strength. There are however, some limitations to acknowledge. Firstly, something about the sampling? Secondly, we did not calculate the statistical power of each modification index, as recommended by Saris et al. (2009). This is because our use of MI was for demonstration purposes only. Further, the extent to which MI substantially change each model requires further investigation, as we provided cross-validation only by splitting the original sample. A true measure of this would be to improve a model fit using the MI on one large sample and then use a completely independent sample to cross validate the new model.

In summary, we have demonstrated here that the proposed cutoff values by Hu and Bentler (1999) are unrealistic for most commonly used scales in sport and exercise psychology. The fact that none of the measures used achieved the suggested cutoff values leads us to one of two conclusions; either all of the measures we assessed are inadequate, or the cutoff values are not appropriate. Because all of the measures used have previously provided evidence of their suitability, to accept the former is likely to lead to the rejection of many highly useful self-report measures. We feel the latter conclusion is a more true, progressive, and helpful conclusion. Further, we recommend that researchers using genuine ICMs seek to examine the MI to improve model fit after performing a CFA. Finally, researchers examining more complex, multidimensional or aggregated models should conduct ESEM in place of CFA.

References

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*, 150-166. doi: 10.1177/1088868306294907

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: The Guilford Press.

Campbell-Sills, L., & Stein, M, B. (2007) Psychometric Analysis and Refinement of the Connor–Davidson Resilience Scale (CD-RISC): Validation of a 10-Item Measure of Resilience. *Journal of traumatic stress, 20,* 1019-1028. doi: 10.1002/jts.20271

Chesney, M. A., Neilands, T. B., Chambers, D. B., Taylor, J. M., & Folkman, S. (2006). A Validity and reliability study of the Coping Self-Efficacy Scale. *British Journal of Health Psychology, 11*, 421-437. doi: 10.1348/135910705X53155

Clough, P., Earle, K., & Sewell, D. (2002) Mental toughness: the concept and its measurement. In I. Cockerill (Ed.), *Solutions in sport psychology*, (pp. 32-43). London: Thomson.

Crust, L. & Azadi, K. (2010).Mental toughness and athletes' use of psychological strategies. *European Journal of Sport Science*, *10*, 43-51. doi:10.1080/17461390903049972

Crust, L. & Clough, P. (2005).Relationship between mental toughness and physical endurance. *Perceptual & Motor Skills*, *100*, 192-194.doi:10.2466/pms.100.1.192-194

Fletcher, R. (2008). Longitudinal factorial invariance, differential, and latent mean stability of the Coping Inventory for Competitive Sport. In M. P. Simmons & L. A. Foster (Eds.) *Sport and exercise psychology research advances*, (pp. 293-306). New York, NY: Nova

Science Publishers Inc.

Gaudreau, P. & Blondin, J-P. (2002). Development of a questionnaire for the assessment of coping strategies employed by athletes in competitive sport settings. *Psychology of Sport and Exercise, 3*, 1-34. doi: 10.1016/S1469-0292(01)00017-6

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

Hu, L.T., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424 – 453.

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

Jones, M. V., Lane, A. M., Bray, S. R., Uphill, M., & Catlin, J. (2005). Development and Validation of the Sport Emotion Questionnaire. *Journal of Sport and Exercise Psychology, 27,* 407-431. *Retrieved from*: http://journals.humankinetics.com/jsep

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202. doi: 10.1007/BF02289343

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousands Oaks, CA: Sage.

Kawabata, M., & Mallett, C. J. (2013). Examining discriminant validity issues of the Sport Motivation Scale-6. *Journal of Sports Sciences, 31*, 173-181. doi: 10.1080/02640414.2012.723132

Kline, R. B. (2005). *Principles and practice of structural equation modeling (2nd ed.).* New York: The Guildford Press.

Levy, A., Polman, R., Clough, P., Marchant, D., & Earle, K. (2006).Mental toughness as a determinant of beliefs, pain, and adherence in sport injury rehabilitation. *Journal of Sports Rehabilitation*, *15*, 246-254. Retrieved from http://journals.com/jsr

Luszczynska, A., Gutiérrez-Doña, B., & Schwarzer, R. (2004). General self-efficacy in various domains of human functioning: Evidence from five countries. *International Journal of Psychology, 40*, 80-89. doi: 10.1080/00207590444000041

MacCullum, R. (1986). Specification Searches in Covariance Structure Modeling. *Psychological Bulletin, 100*, 107-120. doi: 10.1037//0033-2909.100.1.107

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

Mallett, C. J., Kawabata, M., Newcombe, P., Otero-Forero, A., & Jackson, S. (2007). Sport Motivation Scale-6 (SMS-6): A revised six-factor sport motivation scale. *Psychology of Sport and Exercise, 8*, 600–614. doi: 10.1016/j.psychsport.2006.12.005

Marchant, D. C., Polman, R. C. J., Clough, P. J., Jackson, J. G., Levy, A. R., & Nicholls, A. R. (2009). Mental toughness in the work place: Managerial and age differences. *Journal of Managerial Psychology, 27,* 428-437. doi: 10.1108/02683940910959753

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 overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341. doi: 10.1207/s15328007sem1103_2

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22,* 471– 491. doi: 10.1037/a0019227

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

Marsh, H.W., Nagengast, B., Morin, A.J.S., Parada, R.H., Craven, R.G. & Hamilton, L.R. (2011). Construct validity of the multidimensional structure of bullying and victimization: An application of exploratory structural equation modelling. *Journal of Educational Psychology, 103*, 701-732. doi: 10.1037/a0024122

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

Muthén, L.K. and Muthén, B.O. (1998-2012). *Mplus user’s guide. Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Nicholls, A. R., Polman, R. C., Levy, A. R., & Backhouse, S. (2008). Mental toughness, optimism, and coping among athletes. *Personality & Individual Differences*, *44*, 1182-1192. doi: 10.1016/j.paid.2007.11.011

Notario-Pacheco, B., Solera-Martínez, M., Serrano-Parra, M. D., Bartolomé-Gutiérrez, R., García-Campayo, J., & Martínez-Vizcaíno, V. (2011). Reliability and validity of the Spanish version of the 10-item Connor-Davidson Resilience Scale (10-item CD-RISC) in young adults. *Health and Quality of Life Outcomes, 9*, 63-68. doi: 10.1186/1477-7525-9-63

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 5*, 107-124. doi: 10.1080/10705519809540095

Peacock, E., & Wong, P. (1990). The stress appraisal measure (SAM): A multidimensional approach to cognitive appraisal. *Stress Medicine, 6*, 227–236. doi: 10.1002/smi.2460060308

Pelletier, L. G., Fortier, M. S., Vallerand, R. J., Tuson, K. M., Brie` re, N. M., & Blais, M. R. (1995). Toward a new measure of intrinsic motivation, extrinsic motivation, and amotivation in sports: The Sport Motivation Scale (SMS). *Journal of Sport and Exercise Psychology, 17*, 35–54. *Retrieved from*: http://journals.humankinetics.com/jsep

Perry, J. L., Clough, P. J., Earle, K., Crust, L., & Nicholls. A. R. (2013). Factorial validity of the Mental Toughness Questionnaire 48. *Personality and Individual Differences, 4,* 587-592. doi: 10.1016/j.paid.2012.11.020

Saris, W. E., den Ronden, J., & Satorra, A. (1987). Testing structural equation models. In P. Cuttance & R. Ecob (Eds.), *Structural modeling by example* (pp. 202–220). New York: Cambridge University Press.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 561-582. doi: 10.1080/10705510903203433

Scholz, U., Gutiérrez-Doña, B., Sud, S., & Schwarzer, R. (2002). Is general self-efficacy a universal construct? Psychometric findings from 25 countries. *European Journal of Psychological Assessment, 18*, 242-251. doi: 10.1027//1015-5759.18.3.242

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston, *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). Windsor: NFER-NELSON.

Wang, L., Shi, Z, Zhang, & Zhang, Z. (2010). Psychometric properties of the 10-item Connor-Davidson Resilience Scale in Chinese earthquake victims. *Psychiatry and Clinical Neurosciences, 64*, 499-504. doi: 10.1111/j.1440-1819.2010.02130.x

Table 1

Demographic details for each measurement scale

Instrument	Number of items	Factors	Participants		
			Male	Female	Age
CICS	39	10	1798	750	21.75 (5.10)
SAM	28	7	934	327	22.21 (5.50)
MTQ48	48	6	407	218	26.30 (11.85)
SMS-6	24	6	364	158	24.10 (8.46)
SEQ	22	5	1257	431	21.92 (5.16)
CSES	26	3	674	311	20.73 (4.63)
CD-RISC	10	1	408	250	26.91 (11.40)
GSE	10	1	364	158	24.10 (8.46)

Note. CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale; CD-RISC = Connor-Davidson Resilience Scale; GSE = General Self-efficacy Scale**.**

Table 2

Summary of fit indices for measures using CFA

Measure	χ^2	*df*	sig.	CFI	TLI	SRMR	RMSEA	% Loadings > .5	% Loadings > .4
CICS	4063.6	657	<.001	.873	.856	.050	.045	89.74	97.44
MTQ48	2683.1	1065	<.001	.804	.793	.053	.049	68.75	85.42
CSES	1375.8	296	<.001	.797	.777	.068	.061	61.54	73.07
SAM	1959.5	329	<.001	.851	.829	.069	.063	78.57	89.29
SEQ	1390.2	199	<.001	.914	.901	.057	.058	100.00	100.00
SMS-6	766.3	237	<.001	.877	.857	.060	.065	95.83	100.00
CD-RISC	133.7	35	<.001	.850	.807	.037	.065	70.00	90.00
GSE	213.0	35	<.001	.838	.792	.064	.099	60.00	80.00

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; RMSEA = root mean square error of approximation. CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale; CD-RISC = Connor-Davidson Resilience Scale; GSE = General Self-efficacy Scale**.**

Table 3

Model fits using modification indices

Measure	Modifications	χ^2	*df*	sig.	CFI	TLI	SRMR	RMSEA
CICS	84	1908.8	573	<.001	.950	.935	.041	.030
MTQ48	62	1532.0	1003	<.001	.936	.928	.039	.029
CSES	51	562.5	245	<.001	.940	.921	.043	.036
SAM	88	666.9	241	<.001	.961	.939	.050	.037
SEQ	47	574.2	152	<.001	.970	.954	.049	.040
SMS-6	27	365.3	210	<.001	.964	.953	.048	.038
CD-RISC	3	67.9	32	<.001	.945	.923	.028	.041
GSE	12	64.1	23	<.001	.963	.927	.036	.059

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; RMSEA = root mean square error of approximation. CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48;

SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale; CD-RISC = Connor-Davidson Resilience Scale; GSE = General Self-efficacy Scale.

Table 4

Model fits using modification indices for cross-validation

Measure	χ^2		*df*	CFI		TLI		SRMR		RMSEA	
	Sample a	Sample b		Sample a	Sample b	Sample a	Sample b	Sample a	Sample b	Sample a	Sample b
CICS	1276.4	1309.0	573	.948	.946	.933	.930	.045	.043	.031	.032
MTQ48	1286.5	1307.1	1003	.926	.935	.917	.926	.049	.046	.030	.031
CSES	465.0	396.0	245	.924	.944	.899	.926	.051	.046	.043	.035
SAM	409.8	477.7	241	.970	.956	.953	.931	.051	.057	.033	.039
SEQ	365.5	371.3	152	.968	.971	.951	.956	.051	.052	.040	.040
SMS-6	307.2	301.5	210	.951	.963	.936	.951	.060	.054	.043	.040
CD-RISC	67.5	44.8	32	.940	.983	.916	.976	.039	.032	.058	.035
GSE	48.0	40.6	23	.949	.972	.900	.946	.045	.037	.066	.053

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; RMSEA = root mean square error of approximation. CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale; CD-RISC = Connor-Davidson Resilience Scale; GSE = General Self-efficacy Scale.

Table 5

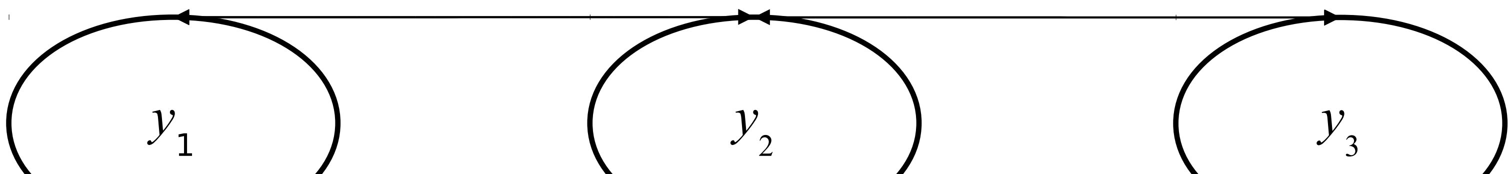
Summary of fit indices for measures using ESEM

Measure	χ^2	*df*	sig.	CFI	TLI	SRMR	RMSEA	% items loading onto intended factor ($p < .01$)	Proportion of significant cross-loadings	% items loading greater onto non-intended factor
CICS	1759.7	396	<.001	.949	.905	.017	.037	89.74	8.46	17.95
MTQ48	1621.5	855	<.001	.907	.878	.031	.038	93.75	20.83	35.42
CSES	968.5	250	<.001	.865	.824	.044	.054	88.46	10.26	15.38
SAM	648.1	203	<.001	.959	.924	.019	.042	92.86	7.14	23.08
SEQ	669.8	131	<.001	.961	.932	.018	.048	90.91	11.82	13.64
SMS-6	339.8	147	<.001	.955	.916	.024	.050	91.67	4.17	4.17

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; RMSEA = root mean square error of approximation. CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale.

Figure 1

An illustration of model structure with estimated parameters in confirmatory factor analysis



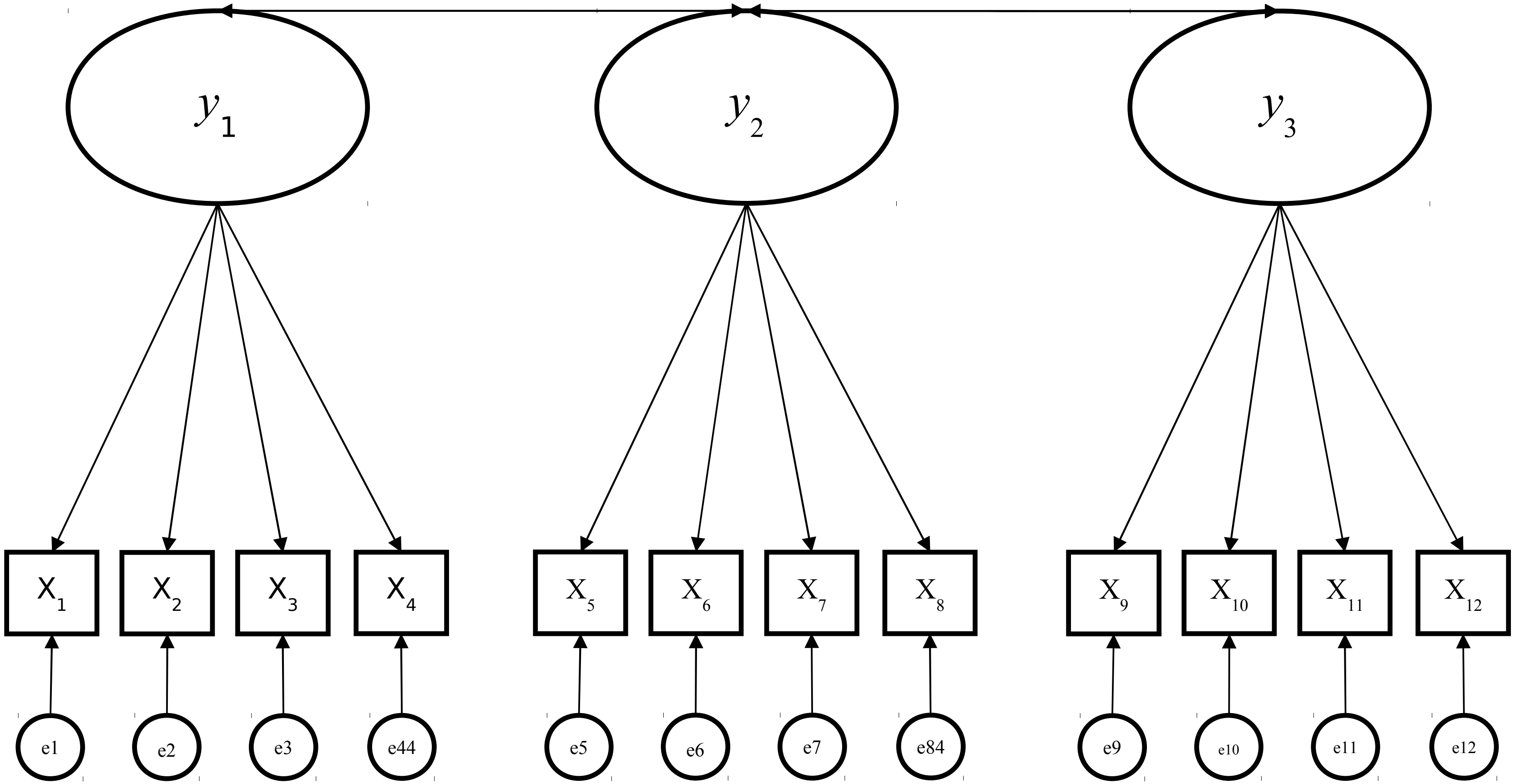
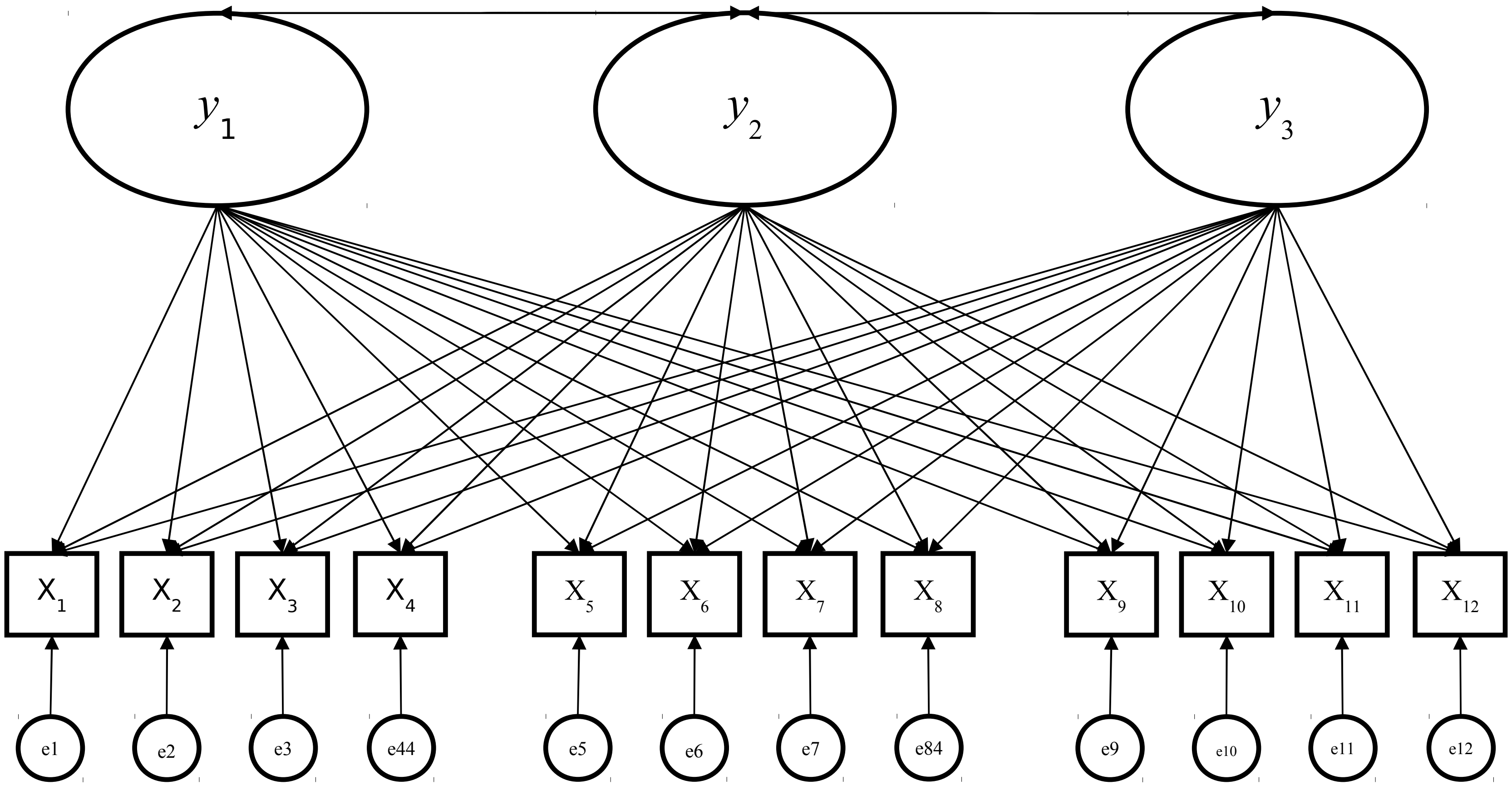


Figure 2

An illustration of model structure with estimated parameters in exploratory structural equation modeling

1

1



y_1
y_2
y_3
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_{10}
X_{11}
X_{12}
e1
e2
e3
e44
e5
e6
e7
e84
e9
e10
e11
e12