Psicothema 2008. Vol. 20, nº 4, pp. 872-882
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Testing for multigroup equivalence of a measuring instrument: A walk through the process

Barbara M. Byrne
University of Ottawa

This article presents an overview and application of the steps taken in testing for the equivalence of a measuring instrument across one or more groups. Following a basic description of, and rationale underlying these steps, the process is illustrated with data comprising response scores to four nonacademic subscales (Physical SC [Ability], Physical SC [Appearance], Social SC [Peers], and Social SC [Parents]) of the Self Description Questionnaire-I for Australian (N= 497) and Nigerian (N= 439) adolescents. All tests for validity and equivalence are based on the analysis of covariance structures within the framework of CFA models using the EQS 6 program. Prospective impediments to equivalence are suggested and additional caveats proposed in the special case where the groups under study represent different cultures.

Comprobando la equivalencia multigrupal de un instrumento de medida: pasos del proceso. Este artículo presenta una visión general y la aplicación de los pasos para comprobar la equivalencia de un instrumento de medida en uno o más grupos. Siguiendo la lógica y una descripción básica de estos pasos, el proceso se ilustra con los datos de las puntuaciones en cuatro subescalas no académicas (Autoconcepto sobre Aptitud Física, Autoconcepto sobre Apariencia Física, Autoconcepto Social en relación con los compañeros y Autoconcepto Social relativo a los padres) pertenecientes al Cuestionario de Auto-Descripción-I para adolescentes australianos (N= 497) y nigerianos (N= 439). Todas las pruebas de validez y equivalencia se basan en el análisis de las estructuras de covarianza dentro del marco de los modelos de análisis factorial confirmatorio, empleando el programa de ecuaciones estructurales EQS 6. Se apuntan posibles impedimentos a la equivalencia y se proponen cautelas adicionales en el caso especial de que los grupos bajo estudio representen culturas diferentes.

In substantive research that focuses on multigroup comparisons, it is typically assumed that the instrument of measurement (e.g., ability tests, assessment/attitudinal scales) is operating in exactly the same way and that the underlying construct has the same theoretical structure and psychological meaning across the groups of interest. As evidenced from reviews of the literature, however, these two critically important assumptions are rarely, if ever tested statistically. The primary approach to addressing this issue of instrument equivalence is to test for the cross-group invariance of its factorial structure using structural equation modeling (SEM) within the framework of a confirmatory factor analytic (CFA) model, a process that can entail several steps depending on the intent of the researcher. The purpose of this pedagogically-oriented article is to outline and illustrate the progression of steps involved in the case where concern focuses solely on the extent to which a measuring instrument is equivalent across two or more independent samples. I demonstrate this procedure based on data comprising item responses to the Physical and Social Self-concept subscales of the Self Description Questionnaire I (Marsh, 1992) for Australian and Nigerian adolescents.

Development of a method capable of testing for multigroup equivalence derives from the seminal work of Jöreskog (1971). Although Jöreskog initially recommended that all tests of equivalence begin with a global test of equivalent covariance structures across groups, this initial test has since been disputed as it often leads to contradictory findings (for details see Byrne, 2006). Testing for equivalence entails a hierarchical set of steps that typically begin with the determination of a well-fitting multigroup baseline model for which sets of parameters are put to the test of equality in a logically ordered and increasingly restrictive fashion. In technical terms, this model is commonly termed the *configural model* and is the first and least restrictive one to be tested (Horn & McArdle, 1992). With this initial model, only the extent to which the same pattern (or configuration) of fixed and freely estimated parameters holds across groups is of interest and thus no equality constraints are imposed. The importance of the configural model is that it serves as the baseline against which all subsequent tests for equivalence are compared. In contrast to the configural model, all remaining tests for equivalence involve the specification of cross-group equality constraints for particular parameters. Specifically, these constrained models test for *measurement equivalence*,

Fecha recepción: 30-12-07 • Fecha aceptación: 26-3-08
Correspondencia: Barbara M. Byrne
School of Psychology
University of Ottawa
K1N6N Ottawa - Ontario (Canadá)
E-mail: bmbch@uottawa.ca

followed by *structural equivalence,* the idea here being that unless it is known that the measurement parameters are operating in the same way across groups, it makes no sense to test for equivalence related to the structural parameters.

Measurement equivalence is concerned with the extent to which parameters comprising the measurement portion of a CFA or full structural equation model are similar across groups. Measurement parameters involve the observed variables (directly measurable) and their links to the unobserved (or latent) variables, which are not directly measurable. These parameters always include the factor loadings and may include the observed variable error variances and any error covariances (commonly termed error correlations). Should a researcher be interested in subsequently testing for latent factor mean differences, then tests for measurement equivalence *must* include a test for the equality of the observed variable intercepts as such equality is assumed in tests for factor mean differences.

In contrast to measurement equivalence, *structural equivalence* focuses on the unobserved variables. As such, structural equivalence is concerned with the equality of relations among the factors (i.e., factor covariances) and may extend to include the factor variances and error residual covariances. Furthermore, should a researcher be interested in testing for cross-group equivalence related to a full structural equation (or path analytic) model, then interest most certainly will extend to testing for the equality of structural regression paths between and among the postulated latent constructs (or factors).

The present article focuses solely on tests for multigroup equivalence related to a CFA model. However, for an extended elaboration of the basic concepts and application techniques associated with CFA, as well as the more complex models, readers are referred to Byrne, 1998; 2001; 2006). I begin this section by first describing how to establish the configural model and then follow with explanation of testing for the validity of this multigroup model. Next, I describe procedures involved in testing first for measurement equivalence, followed by those in testing for structural equivalence. Finally, I close out this section with caveats specific to tests for cross-cultural equivalence.

Establishing the configural model

Establishment of the configural model begins with specification and testing of the hypothesized model (i.e., postulated structure of the measurement instrument under study) for each group separately. These group-specific models are termed *baseline models*. As such, for each group, the baseline model specifies the number of subscales (i.e., factors), the location of the items (i.e., pattern by which the items load onto each factor), and postulated correlations among the subscales (i.e., existence of factor covariances). The validity of these baseline models is tested separately for each group. Ideally, these models will be well-fitting and therefore best fit the data from the perspectives of both parsimony and substantive meaningfulness. If, on the other hand, the model exhibits some evidence of misfit, the researcher may want to re-specify and re-estimate the model, but always with a watchful eye on parsimony as the more complex the model, the more difficult it is to attain group equivalence.

Because measuring instruments are often group-specific in the way they operate, it is possible that baseline models may not be completely identical across groups (see Bentler, 2005; Byrne, Shavelson, & Muthén, 1989). For example, it may be that the best-fitting model for one group includes an error covariance or a cross-loading (i.e., the loading of an item on a factor other than the one for which it was designed), whereas these parameters may not be specified for any of the other groups under study. Presented with such findings, Byrne et al. (1989) showed that by implementing a condition of *partial measurement invariance*, multigroup analyses can still continue given that the recommended conditions are met. As such, some but not all measurement parameters are constrained equal across groups in the testing for structural equivalence. A priori knowledge of such group differences is critical to the application of equivalence-testing procedures, particularly as they relate to cross-cultural samples. Once a well-fitting baseline model has been established for each group separately, these final models are then combined in the same file to form the multigroup model, commonly termed the configural model.

Testing for configural equivalence

The initial step in testing for cross-group equivalence requires only that the same number of factors and their loading pattern be the same across groups; as such, no equality constraints are imposed on the parameters. In other words, the same parameters that were estimated in the baseline model for each group separately are again estimated, albeit within the framework of a multigroup model. Goodness-of-fit related to this multigroup parameterization should be indicative of a well-fitting model. Of import here is that, despite evidence of good fit to the multi-sample data, the only information that we have is that the factor structure is similar, but not necessarily *equivalent* across groups as equivalence of the factors and their related items has not yet been put to the test (i.e., only their overall model fit has been tested).

In essence then, the configural model being tested here is a multigroup representation of the baseline models. This multigroup model serves two important functions. *First*, it allows for equivalence tests to be conducted across the groups *simultaneously*. In other words, parameters are estimated for all groups at the same time. *Second*, in testing for equivalence, the fit of this configural model provides the baseline value against which all subsequently specified invariance models are compared. In contrast to single-group analyses, however, this multigroup analysis yields only one set of fit statistics for overall model fit. When ML estimation is used, the χ^2 statistics are summative and, thus, the overall χ^2 value for the multigroup model should equal the sum of the χ^2 values obtained when the baseline model is tested separately for each group of teachers (with no cross-group constraints imposed). If estimation is based on the robust statistics (to be described later), the corrected χ^2 statistic for each group is not necessarily summative across the groups.

Testing for measurement equivalence

When a researcher is concerned only in the extent to which an instrument is equivalent across independent samples, measurement equivalence generally focuses solely on the invariant operation of the items and, in particular, on the factor loadings. As such, interest centers on the extent to which the content of each item is being perceived and interpreted in exactly the same way across the samples. Testing for invariant factor loadings has been termed tests for «metric equivalence» (Horn & McArdle, 1992) as well as

«measurement unit equivalence» (van de Vijver & Leung, 1997). Although one can also test for the equivalence of measurement error terms, it is now widely accepted that this test is overly restrictive and likely of least interest and importance (Bentler, 2005; Widaman & Reise, 1997) unless the equality of item reliabilities are of interest (see Byrne, 1988 for such an application).

In testing for the equivalence of factor loadings, these parameters are freely estimated for the first group only; for all remaining groups, factor loading estimates are constrained equal to those of Group 1.[1] Provided with evidence of equivalence, these factor loading parameters remain constrained equal while simultaneously testing for the equivalence of additional parameters (e.g., factor covariances). On the other hand, confronted with evidence of nonequivalence related to particular factor loadings, one may proceed with subsequent tests for equivalence if the data meet the recommended conditions of partial measurement equivalence (see Byrne et al., 1989).

Testing for structural equivalence

In contrast to tests for measurement equivalence, which focus on aspects of the observed variables, tests for structural equivalence center on the unobserved (or latent) variables. In testing for multigroup equivalence of a measuring instrument, interest can focus on both the factor variances and their covariances. However, the latter are typically of most interest and therefore constitute the focus in most tests for instrument equivalence. A review of the SEM literature reveals much inconsistency regarding whether or not researchers test for structural equivalence. In particular, these tests are of critical import to construct validity researchers whose interests lie either in testing the extent to which the dimensionality of a construct, as defined by theory, holds across groups (see e.g., Byrne & Shavelson, 1987), or in the extent to which a measuring instrument, developed within the framework of a particular theory, yields the expected dimensional structure of the measured construct in an equivalent manner across groups (see e.g., Byrne & Watkins, 2003).[2] In both instances, the parameters of most interest are the factor covariances.

Testing for cross-cultural equivalence

Of particular import in testing for instrument equivalence is the special case where a researcher wishes to use a measuring instrument that was developed and normed in another country. At first blush, the task would appear simply to be a matter of translating the instrument from one language into another. In sharp contrast, however, this most certainly is not the case! Indeed, the process extends far beyond the issue of translation and involves a comprehensive and rigorous series of procedures that test for the validity of the measure's scores within the new cultural context, as well as for its structural and measurement equivalence with the original instrument and culture. In simple terms, this initial task essentially involves three macro steps: (a) translate the instrument into the desired language, (b) based on the hypothesized factorial structure of the original instrument, test for its validity relative to the newly translated version, and (c) test for measurement and structural equivalence of the translated version with the original one. In reality, however, these three global procedures can be much more complex and involve numerous additional tests before an instrument can be considered sufficiently equivalent to its parent version. For example, in contrast to popular thought, the process of translating an instrument from one language into another involves much more than mere back translation (for details, see Hambleton, Merenda, & Spielberger, 2005). Another example can be item content that is inappropriate and/or meaningless for the culturally new respondents thereby resulting in differential perception of items and hence findings of nonequivalence across the two cultural groups. Given these and many more such examples, the term *test adaptation* is used to describe this more advanced approach to the development and use of translated instruments (see Hambleton et al., 2005).

I have summarized the basic steps in testing for cross-group measurement and structural equivalence when interest focuses only on the extent to which a measuring instrument is equivalent across groups, and have outlined particular cautions in the special case where an instrument is developed in one culture and then adapted for use in another culture. However, as noted earlier, the process of testing for equivalence can involve several additional steps depending on the intent of the study, the particular data under study, and the level of stringency a researcher wishes to apply. For more detailed descriptions of tests for multigroup equivalence, readers are referred to Horn & McArdle (1992), Little (1997), and Widaman and Reise (1997); to Byrne (1998, 2001, 2006) for annotated explanation and illustration of diverse models based on the LISREL, AMOS, and EQS programs, respectively; and to Vandenberg and Lance (2000) for a review of the multigroup equivalence literature. We turn now to an annotated application of testing for the equivalence of a measuring instrument.

Illustrative application

The example data

The samples. The sample data used in this application comprise 497 Australian (266 males; 231 females) and 439 Nigerian (219 males; 220 females) adolescents.[3] An important difference between the two samples, however, reflects on the degree of missing data. Whereas data for the Australians were complete (i.e., no missing values), those for the Nigerians had some missing scores (original N= 465). In addressing this issue of incomplete data, all cases having >8% missing data were deleted from the analyses. For the remaining sample of 439, the randomly missing data were imputed with values derived from a multiple regression in which three item scores from the same congeneric set of indicators (i.e., items measuring the same construct) were used as the predictor variables. Although maximum likelihood (ML) estimation (see Arbuckle, 1996) is now considered the most efficient approach to dealing with missing data, Bentler (2005) notes that when the amount of missing data is very small (the case here), methods such as regression imputation may suffer only marginal loss of efficiency. (For an elaborated discussion of imputation in general, and the preference for regression-based imputation in particular, readers are referred to Byrne, 2001.) Ages of adolescents from both countries ranged from 12 to 14 years (median age= 13 years).

The instrument of measurement. The SDQ-I is a 76-item self-report inventory based on a 5-point Likert-scale format designed for use with children ranging in age from 8 through 12 years. Importantly, the simplistic English wording of the SDQ-I items

makes them suitable for slightly older respondents for whom English is not the first language. The respondent is presented with a series of short statements (e.g., I am good looking), and then asked to select the option which most appropriately reflects his or her level of agreement; choices range from 'false' (1) to 'true' (5). The SDQ-I has been shown to be one of the most psychometrically sound measures of self-concept available (see Byrne, 1996).

Although the SDQ-I comprises 7 subscales that measure both academic and nonacademic self-concepts, only the latter are of interest here. These nonacademic subscales are designed to tap four facets of self-concept: Physical Self-concept relative to one's physical ability (PSC-Ability); Physical Self-concept relative to one's appearance (PSC-Appearance); Social Self-concept relative to one's peers (SSC-Peers); Social Self-concept relative to one's parents (SSC-Parents).

The hypothesized model

The CFA model of SDQ-I factorial structure, as it relates to the four nonacademic subscales, is shown schematically in figure 1.

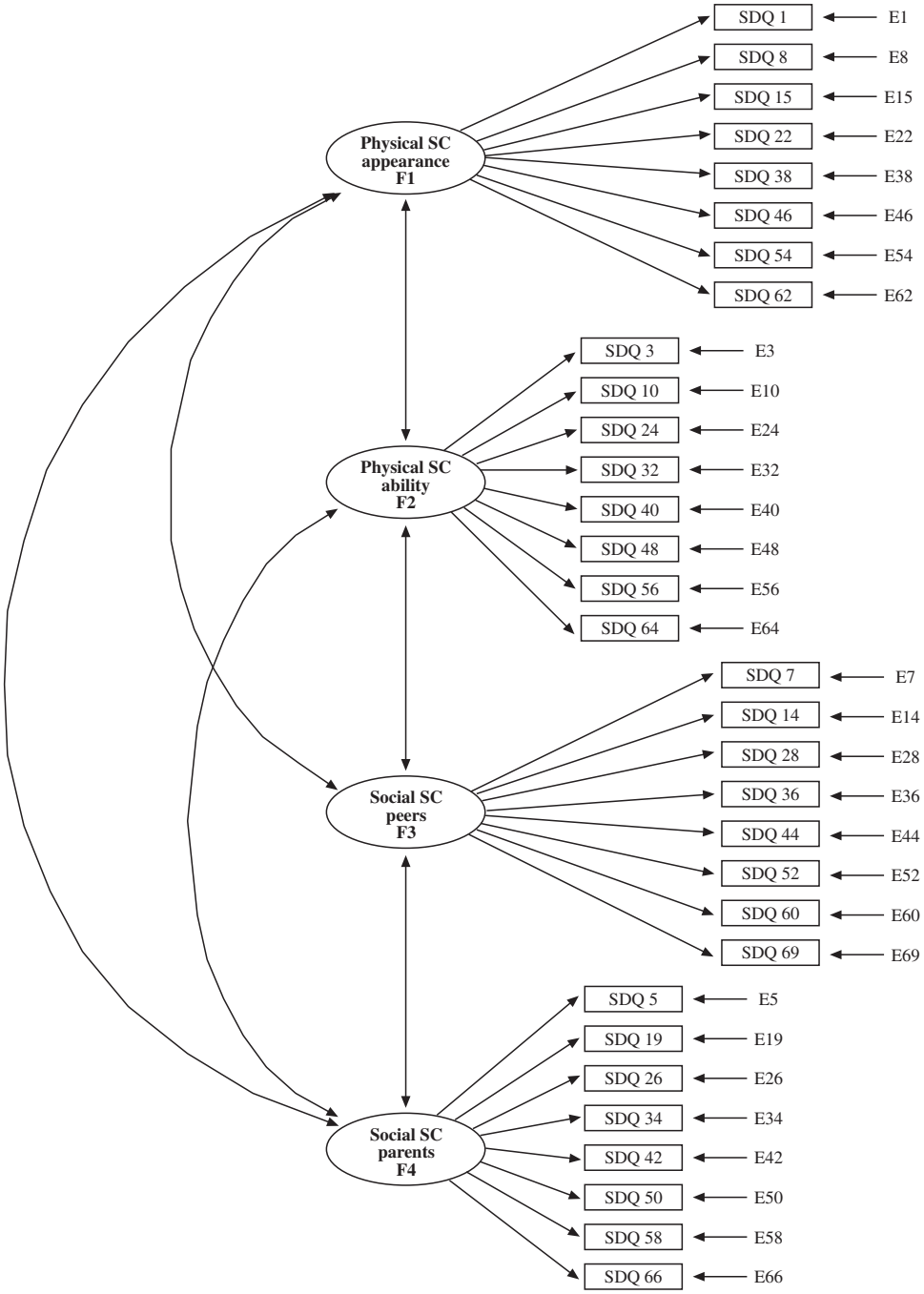

Figure 1. *Hypothesized Model of Factorial Structure for the Self Description Questionnaire-I. Nonacademic Subscales (SDQ-I; Marsh, 1992)*

As such, there are four factors (PSC-Ability; PSC-Appearance; SSC-Peers; SSC-Parents) each of which is measured by 8 items. This model hypothesizes a priori that for both Australian and Nigerian adolescents: (a) each observed variable (i.e., SDQ-I item) has a nonzero loading on the factor it is designed to measure, and zero loadings on all other factors, (b) consistent with theory, the four factors are intercorrelated (as indicated by the double-headed arrows, and (c) measurement error terms associated with the observed variables (the E's) are uncorrelated.

Statistical analyses

All analyses were conducted using the most recent version of the EQS6 program (Bentler, 2005), with all tests of validity and equivalence being based on the analysis of covariance structures within the framework of the CFA model. Given that preanalyses of the data revealed substantial multivariate kurtosis for both groups as evidenced by related Mardia's normalized coefficients of 80.70 (Australians) and 72.98 (Nigerians)[4], analyses were based on the Satorra-Bentler scaled chi-square statistic (S-Bχ^2; Satorra & Bentler, 1988), rather than the usual MLχ^2 statistic as it serves as a correction for χ^2 when distributional assumptions are violated.

Too often, researchers wishing to conduct SEM analyses seem not to be knowledgeable with respect to both the underlying concepts and related issues associated with the critical assumption of multivariate normality. Statistical research has shown that whereas skewness tends to impact tests of means, kurtosis severely affects tests of variances and covariances (DeCarlo, 1997). Given that SEM is based on the analysis of covariance structures, evidence of kurtosis is always of concern. In particular, it is now well known that multivariate kurtosis is exceptionally detrimental to parameter estimation in SEM analyses (see e.g., Curran, West, & Finch, 1996). Thus, it is essential that researchers intent on using this methodology always scrutinize their data for evidence of multivariate normality. Presented with findings of multivariate non-normality, the onus is on the researcher to select the most appropriate analytic approach in addressing this issue.[5]

In testing each of the three models described earlier, evaluation of goodness-of-fit to the sample data was determined on the basis of multiple criteria; the Comparative Fit Index (*CFI; Bentler, 1990), the Root Mean-Square Error of Approximation (*RMSEA; Browne & Cudeck, 1993), and the Standardized Root Mean Square Residual (SRMR). The *CFI represents the robust version of the CFI in that its computation is based on the S-Bχ^2 statistic; it ranges in value from zero to 1.00. Although Hu and Bentler (1999) suggest a value of .95 to be indicative of good fit (see Hu & Bentler, 1999), others argue that it is too restrictive, particularly for multifactor rating scales for which analyses are conducted at the item level (see Marsh, Hau, & Wen, 2004). Thus, *CFI values in the range of .92 through .94 may also be considered as reasonable indicators of good model fit. The *RMSEA is a robust version of the usual RMSEA and takes into account the error of approximation in the population. It asks the question «How well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?» (Browne & Cudeck, 1993, pp. 137-138). This discrepancy, as measured by the *RMSEA, is expressed per degree of freedom, thus making it sensitive to model complexity; values less than .05 indicate good fit, and values as high as .08 represent reasonable errors of approximation in the population. For completeness, I also include the 90% confidence interval provided for *RMSEA (see Steiger, 1990). Finally, the SRMR is the average standardized residual value derived from fitting the hypothesized variance covariance matrix to that of the sample data. Its value ranges from zero to 1.00, with a value less than .08 being indicative of a well-fitting model (Hu & Bentler, 1999).

Results

Establishing the configural model. As noted earlier, a prerequisite to testing for instrument equivalence is to establish a well-fitting baseline model for each group separately.

Initial testing of the hypothesized model for Australian adolescents yielded only a marginally good fit to the data (S-B$\chi^2_{(458)}$= 1059.24; SRMR= .07; *CFI= .90; *RMSEA= .05, 90% C.I.= .047, .055). A review of the LMTest statistics (i.e., indices of model misfit)[6] revealed one cross-loading (F3«SDQ38) and two error covariances (SDQ40/SDQ24; SDQ26/SDQ19) to be markedly misspecified. The related SDQ items are as follows:

- Item 38: Other kids think I am good looking
- Item 24: I enjoy sports and games
- Item 40: I am good at sports
- Item 19: I like my parents
- Item 26: My parents like me

Given that the cross-loading of Item 38 on Factor 3 (Social [Peers]) seemed reasonable, this parameter was added to the model first and the model re-estimated. This respecification revealed a slight improvement in model fit (S-B$\chi^2_{(457)}$= 1003.66; SRMR= .06; *CFI= .91; *RMSEA= .05, 90% C.I.= .045, .053). Given the obvious overlap of content between Items 19 and 26, and possible content overlap between Items 24 and 40, albeit the first statement regarding perceived ability in sports is descriptive, whereas the second is evaluative, the model was subsequently respecified and reestimated with these two error covariances included. This reparameterization resulted in a further slight improvement in model fit (S-B$\chi^2_{(455)}$= 903.88; SRMR= .06; *CFI= .92; *RMSEA= .05, 90% C.I.= .040, .049). Although a review of the LMTest statistics suggested the addition of a second cross-loading to the model (F1→SDQ32), this parameter was not incorporated for two reasons: (a) considerations of parsimony, and (b) the questionable meaningfulness of this cross-loading across males and females. The item content reads as «I have good muscles» and thus, given its possible gender-specificity, argues against the specification and estimation of this parameter. The resulting baseline model which I considered to be the most appropriate for the Australian sample, despite its modestly adequate fit, comprised one cross-loading and two error covariances as detailed above.

Turning next to establishment of a baseline model for the Nigerian sample, we find that in contrast to the Australians, results revealed a fairly well-fitting model (S-B$\chi^2_{(458)}$= 732.93; SRMR= .06; *CFI= .92; *RMSEA= .04, 90% C.I.= .032, .042). A review of the LMTest statistics revealed only one parameter that could be regarded as misspecified and this was an error covariance between Items 26 and 19, which replicated the same finding for Australian adolescents. Thus, the model was subsequently respecified and reestimated with this parameter freely estimated. This respecification yielded some improvement in goodness-of-fit,

thereby resulting in a fairly well-fitting model (S-B$\chi^2_{(457)}$= 702.49; SRMR= .06; *CFI= .93; *RMSEA= .04, 90% C.I.= .030, .040). Given no further clear evidence of poorly specified parameters, this model was deemed the most appropriate baseline model for Nigerian adolescents.

Having determined baseline models for both groups under study, we now combine them into one file for purposes of testing cross-group equivalence. Consistent with the baseline testing strategy, this multigroup model comprises two differentially specified baseline models that are schematically presented in figure 2.

Review of this schema shows the same error covariance between Items 19 and 26 (indicated by the double-headed arrow) for both the Nigerians and the Australians. The additional error covariance between Items 24 and 40, together with the cross-loading of Item 38 on Factor 3 (indicated by the one-headed arrow leading from F3 to the original placement of Item 38 on F1) is specified only for the Australians.

Testing for configural equivalence. As noted earlier, the multigroup model under test in this first step of testing for instrument equivalence is one in which no equality constraints are imposed. This configural model simply incorporates the baseline models for both groups and allows for their simultaneous analyses. Given that the model fits reasonably well, we can conclude that both the number of factors and the pattern of their item loadings are similar across Australians and Nigerians.

To assist readers in making the link between the graphical portrayal of this model in figure 2 and its related textual specification, the EQS input file is presented in the Appendix. For those who may not be familiar with EQS notation a brief explanation is provided here: V's= observed variables (in this case, single items); F's= factors; E's= measurement errors; *'s= estimated parameters. Baseline model specification is presented first for Australian adolescents, followed by that for Nigerian adolescents. For the Australians, note that V38 (Item 38) is estimated to load on Factor 1 (the original loading), as well as on Factor 3. Within the Covariances paragraph, the two error covariances (E26,E19; E40,E24) are shown to be estimated. Although the covariance between error terms E26 and E19 are also specified for Nigerian adolescents, it is important to note the different labelling, which necessarily occurs as a consequence of the location of the items within this group's data base. In the line immediately following the Label paragraph (**/LABELS**), note that the first three entries represent ID, SEX, and AGE. Because (a) these variables are not included in the Australian data base and, (b) all variables are automatically given a V label as they are entered into the EQS data base, these additional three variables for the Nigerians cause the program to label Item 1 as V4, rather than V1 as is the case for the Australians (for a nonmathematical introduction to SEM together with various annotated applications based on the EQS program, readers are referred to Byrne, 2006).

As expected, goodness-of-fit statistics related to the testing of this configural model yielded a modestly well-fitting model (S-B$\chi^2_{(912)}$= 1610.76; SRMR= .06; *CFI= .92; *RMSEA= .04, 90% C.I.= .037, .044) thereby suggesting that the configural model represents the data fairly well. Thus, we conclude that the factorial structure of the four SDQ-I nonacademic subscales is optimally represented as a four-factor model, with the pattern of factor loadings specified in accordance with the postulated configural model.

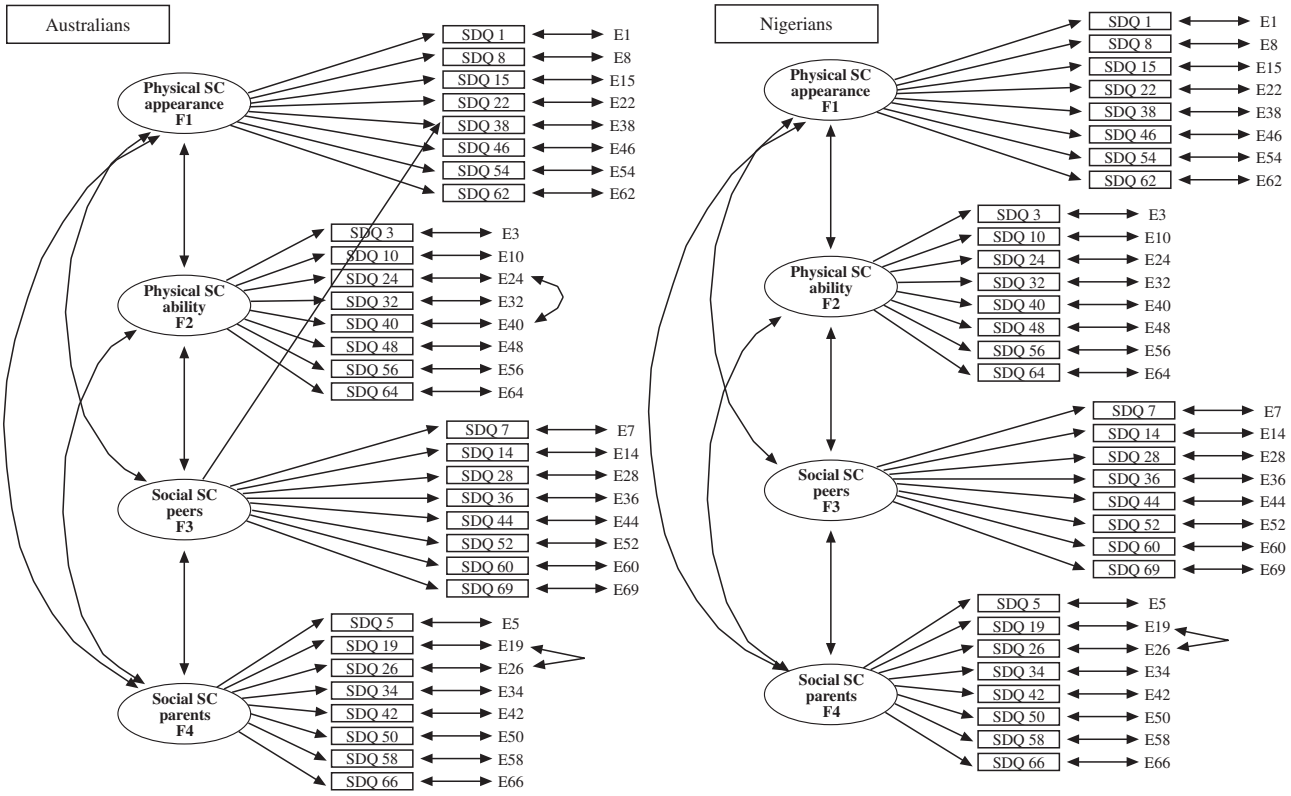

Figure 2. *Baseline Models of SDQ-I Structure for Australian and Nigerian Adolescents*

Testing for measurement equivalence. As noted earlier, in testing for measurement equivalence, factor loadings are estimated only for the first group (in this case, the Australians) and then constrained equal for the second group. That is to say, factor loading values for the Nigerians were constrained equal to the values estimated for the Australians. However, it is important to note that, because Item 38 was specified differently for the two groups, no equality constraint was imposed on this parameter. Furthermore, given that one of the error covariances (E26,E19) was specified for both groups of adolescents, this parameter was also tested for its equivalence. In EQS, equality constraints are specified in the **/CONSTRAINTS** paragraph of the input file. These equality constraints, as they relate to the factor loadings and the one error covariance, are shown in table 1.

Goodness-of-fit statistics related to this constrained model revealed very negligible decrement in overall fit compared with results for the configural model (S-B$\chi^2_{(940)}$= 1721.48; SRMR= .07; *CFI= .92; *RMSEA= .04, 90% C.I.= .039, .045). In testing for equivalence, the models of interest are necessarily nested and thus can be compared in pairs by computing the difference in their overall χ^2 values and the related degrees of freedom; the test is known as the Likelihood Ratio Test. This χ^2 difference value ($\Delta\chi^2$) is distributed as χ^2, with degrees of freedom equal to the difference in degrees of freedom (Δdf). Analogously, the same comparisons can be made based on the ΔS-B$\chi^2_{(\Delta df)}$, albeit a correction to the value is needed as this difference is not distributed as χ^2 (Bentler, 2005; for specification and application of this formula, see Byrne, 2006). If this value is statistically significant, in the comparison of two nested models, it suggests that the constraints specified in the more restrictive model do not hold (i.e., the two models are not equivalent across groups). If, on the other hand, the $\Delta\chi^2$ value is statistically nonsignificant, this finding suggests that all specified equality constraints are tenable.

Table 1
EQS input for equality constraints in test of measurement equivalence

```
/CONSTRAINTS
(1,V8,F1) = (2,V11,F1);
(1,V15,F1) = (2,V18,F1);
(1,V22,F1) = (2,V25,F1);
(1,V46,F1) = (2,V49,F1);
(1,V54,F1) = (2,V57,F1);
(1,V62,F1) = (2,V65,F1);
(1,V10,F2) = (2,V13,F2);
(1,V24,F2) = (2,V27,F2);
(1,V32,F2) = (2,V35,F2);
(1,V40,F2) = (2,V43,F2);
(1,V48,F2) = (2,V51,F2);
(1,V56,F2) = (2,V59,F2);
(1,V64,F2) = (2,V67,F2);
(1,V14,F3) = (2,V17,F3);
(1,V28,F3) = (2,V31,F3);
(1,V36,F3) = (2,V39,F3);
(1,V44,F3) = (2,V47,F3);
(1,V52,F3) = (2,V55,F3);
(1,V60,F3) = (2,V63,F3);
(1,V69,F3) = (2,V72,F3);
(1,V19,F4) = (2,V22,F4);
(1,V26,F4) = (2,V29,F4);
(1,V34,F4) = (2,V37,F4);
(1,V42,F4) = (2,V45,F4);
(1,V50,F4) = (2,V53,F4);
(1,V58,F4) = (2,V61,F4);
(1,V66,F4) = (2,V69,F4);
(1,E26,E19) = (2,E29,E22);
```

Decisions of equivalence based on this difference test originated with the LISREL program (Jöreskog, 1971; Jöreskog & Sorböm, 1993) as it represented the only way to identify evidence of nonequivalence. More recently, however, researchers (e.g., Cheung & Rensvold, 2002; Little, 1997) have argued that this $\Delta\chi^2$ value is an impractical and unrealistic criterion upon which to base evidence of equivalence. Thus, there has been a trend towards basing comparative models on the difference between the CFI values (ΔCFI or *CFI) as a more practical approach to determining the extent to which models are equivalent. Until the recent simulation research of Cheung and Rensvold (2002), however, use of the ΔCFI difference value has been of a purely heuristic nature. Following an extensive study of the properties of 20 goodness-of-fit indices within the context of invariance testing, Cheung and Rensvold (2002) arbitrarily suggested a ΔCFI (or *CFI) difference value not exceeding 0.01.

The two equivalence testing approaches just described operate basically at a macro level, the most recent strategy being the more practical of the two. However, a more precise approach to identifying parameters that are not group-equivalent is provided in EQS by means of the Lagrange Multiplier Test (LMTest), a multivariate test of equality. As such, one examines the probability value associated with the LMTest χ^2 statistic assigned to each constrained parameter; values less than 0.05 indicate nonequivalence. Due to space limitations, readers are referred to Byrne (2006) for a more complete explanation of the LMTest, together with example applications and interpretation of related findings. Based on the present study, however, results revealed four items to be operating differentially across Australian and Nigerian adolescents:

Item 24 – I enjoy sports
Item 40 – I am good at sports
Item 50 – My parents are easy to talk to
Item 66 – My parents and I have a lot of fun together

Whereas Items 24 and 40 relate to the Physical (Ability) SC Scale, Items 50 and 66 belong to the Social (Parents) SC Scale. Of import from a psychometric perspective is the total equivalence of the common error covariance across Australians and Nigerians suggesting that for both groups of adolescents, there is strong overlap of item content..

Testing for structural equivalence. In testing for structural equivalence, interest focuses on the factor covariances. Although some researchers may also wish to test for the equality of the factor variances, these parameters are typically of little interest. I draw your attention to three important aspects of this test. *First*, equality constraints related to items found not to be equivalent across the groups (Items 24, 40, 50, 66) are no longer specified. Rather the factor loadings for these four items are freely estimated for each group. *Second*, equality constraints are now specified for the four factor covariances (see last six lines of the **/CONSTRAINTS** paragraph shown in table 2). *Finally*, the equality of these structural parameters is tested while concomitantly maintaining the equality of all remaining factor loadings (i.e., those found to be equivalent across groups). Thus, it is easy to understand why the equivalence-testing criteria become increasingly stringent as one progresses from tests of the measurement model, to tests of the structural model.

Goodness-of-fit statistics related to this model were somewhat weaker than for the two previous two models (S-B$\chi^2_{(942)}$= 1729.91; SRMR= .09; *CFI= .91; *RMSEA= .04, 90% C.I.= .039, .045), which is not unexpected. Findings related to these analyses revealed all factor covariances involving Factor 4 (Social SC [Parents]) to be nonequivalent across Australian and Nigerian adolescents. In addition, two further items were found not to be operating in the same way across the groups (Items 56 and 52).

Table 2
EQS input for equality constraints in test of structural equivalence

```
/CONSTRAINTS
(1,V8,F1)= (2,V11,F1);
(1,V15,F1)= (2,V18,F1);
(1,V22,F1) = (2,V25,F1);
(1,V46,F1) = (2,V49,F1);
(1,V54,F1) = (2,V57,F1);
(1,V62,F1) = (2,V65,F1);
(1,V10,F2) = (2,V13,F2);
(1,V32,F2) = (2,V35,F2);
(1,V48,F2) = (2,V51,F2);
(1,V56,F2) = (2,V59,F2);
(1,V64,F2) = (2,V67,F2);
(1,V14,F3) = (2,V17,F3);
(1,V28,F3) = (2,V31,F3);
(1,V36,F3) = (2,V39,F3);
(1,V44,F3) = (2,V47,F3);
(1,V52,F3) = (2,V55,F3);
(1,V60,F3) = (2,V63,F3);
(1,V69,F3) = (2,V72,F3);
(1,V19,F4) = (2,V22,F4);
(1,V26,F4) = (2,V29,F4);
(1,V34,F4) = (2,V37,F4);
(1,V42,F4) = (2,V45,F4);
(1,V58,F4) = (2,V61,F4);
(1,E26,E19) = (2,E29,E22);
(1,F1,F2)= (2,F1,F2);
(1,F1,F3) = (2,F1,F3);
(1,F1,F4) = (2,F1,F4);
(1,F2,F3) = (2,F2,F3);
(1,F2,F4) = (2,F2,F4);
(1,F3,F4) = (2,F3,F4);
```

Table 3
Test for invariance of the SDQ-I nonacademic subscales: goodness of fit statistics

Model	S-Bχ^2	df	*CFI	*RMSEA	*RMSEA 90% C.I.
1. No constraints imposed	1610.76	912	.92	.04	.037, .044
2. Factor loadings and common error covariance[a] constrained equal across groups	1721.48	940	.92	.04	.039, .045
3. Equivalent factor loadings, error covariance, factor covariances constrained equal across groups	1729.91	942	.91	.04	.039, .045

[a] Error covariance between Items 19 and 26

*CFI= robust version of the Comparative Fit Index; *RMSEA= robust version of the Root Mean Square
Error of Approximation; C.I. = confidence interval

As a synopsis of these three tests for equivalence, all model fit statistics and list of nonequivalent parameters are shown in tables 3 and 4, respectively.

Summary

The purpose of this article was to present an overview of the steps taken in testing for the equivalence of a measuring instrument across one or more groups. Following a basic description of, and rationale underlying these steps, I illustrated this process based on response scores to four nonacademic subscales of the SDQ-I for Australian and Nigerian adolescents; these subscales represented measures of Physical SC (Ability), Physical SC (Appearance), Social SC (Peers), and Social SC (Parents). Analyses were conducted using the most recent version of the EQS SEM program.

Based on the LMTest, a multivariate test capable of pinpointing misspecified parameters in the model, findings from this example application revealed evidence of both measurement and structural nonequivalence. At the measurement level, 3 items measuring physical SC (Ability), 2 items measuring social SC (Parents) and 1 item measuring Social SC (Peers) were found not to be operating equivalently across Australian and Nigerian adolescents. At the structural level, there was strong evidence of nonequivalence with respect to relations between the social SC (Parents) subscale and each of the other subscales. These findings seem clear in pointing to major differences between Australian and Nigerian adolescents with respect to self-perceived physical ability and self-perceived relations with parents. The next investigative task, then, is to determine whether these differences represent true cultural discrepancies, or rather, are a function of other extraneous factors. Although this follow-up focus is critical to all research that tests for instrument equivalence, particularly that which involves different language versions, this work again exceeds the scope of the present article. Nonetheless, given the particularly challenging task of testing for, and establishing instrument equivalence across culture, I now highlight some of the many ways by which various circumstances can obstruct this process.

It is evident that differences in social norms and values, and resulting socialization practices can play a major role in creating cultural differences in the meaning and/or structure of a measured construct and the perception of its related item content. Thus,

Table 4
Nonequivalent parameters across australian and nigerian adolescents

Parameter	Item content	Related factor(s)
Factor loadings		
Item 24	I enjoy sports sports and games	PSC (Ability)
Item 40	I am good at sports	PSC (Ability)
Item 50	My parents are easy to talk to	SSC (Parents)
Item 66	My parents and I have a lot of fun together	SSC (Parents)
Item 56	I am a good athlete	PSC (Ability)
Item 52	I have more friends than most other kids	SSC (Peers)
Covariances		
Factors 1 & 4 PSC (Appearance)/SSC (Parents)		
Factors 2 & 4 PSC (Ability)/SSC (Parents)		
Factors 3 & 4 SSC (Peers)/SSC (Parents)		

PSC= physical self-concept; SSC= social self-concept

social phenomena alone can contribute importantly to the presence of measurement and/or structural *nonequivalence*. The issue of *measurement nonequivalence* can be examined through various sources, including differential meaning of the activities or behaviors measured (i.e., functional nonequivalence), differential psychometric properties of the scale or test (i.e., metric nonequivalence), and various types of bias. Common bias examples include those that occur during data collection (i.e. method bias) and those due to inadequate item development or translation (i.e. item bias). Method bias can arise from particular characteristics of the instrument (e.g., response styles such as acquiescence or extremity ratings) or its administration (e.g., communication problems between interviewer and interviewee). Item bias can occur as a consequence of differences in the appropriateness of item content (e.g., use of a term or colloquialism that is not understood in at least one of the cultural groups), inadequate item formulation (e.g., unclear wording), or unsuitable item translation (see van de Vijver & Leung, 1997 for an elaboration of these sources of bias).

The issue of *structural nonequivalence* focuses on the extent to which the construct(s) and, in particular its dimensionality, is dissimilar across samples. Here, we would want to investigate various aspects of possible construct bias. For example, pertinent to each group under study: (a) Do the items adequately target all domains of the construct they are designed to measure?; (b) Are all dimensions of the construct relevant?; (c) Are the behaviors being tapped by the items germane? Answers to these and other related questions are essential in probing why the dimensionality of a construct might differ across samples.

Scientific inquiry that seeks to ascertain the extent to which a measuring instrument is equivalent across independent samples, particularly across different cultural samples, clearly represents a challenging, albeit intriguing mission. Nonetheless, it is a task that is critical to the appropriate and responsible use of tests and assessment scales with diverse groups. Undoubtedly, for researchers charged with the task of testing for instrument equivalence, thorough knowledge of the procedure and familiarity with the methodological literature are essential. I am hopeful that readers will find this article helpful in providing such information, thereby making the task less onerous.

Footnotes

1 Determination of which group should serve as Group 1 is purely arbitrary.

2 Wells and Marwell (1976) noted over thirty years ago that measurement and theory are inseparately wed. Thus, one tests either for the validity of a theory (assuming accurate measurements) or tests for the validity of the measuring instrument (assuming an accurate theory), but cannot validate both simultaneously.

3 Given that SEM is grounded in large sample theory, sample sizes of at least 200 are strongly recommended (Boomsma & Hooglund, 2001).

4 When samples are very large and multivariately normal, Mardia's normalized estimate is distributed as a unit normal variate such that large values reflect significant positive kurtosis and large negative values reflect significant negative kurtosis. Bentler (2005) has suggested that, in practice, values >5.00 are indicative of data that are non-normally distributed.

5 An anonymous reviewer requested a discussion of power. However, within the framework of SEM, the assessment of power is very complex and clearly extends beyond the scope of this paper. Unlike simple procedures such as ANOVA for which alternative hypotheses pertain to only a few parameters, in SEM there are considerably more parameters. In addition to sample size, power is affected by the size and location of misspecified model parameters, the number of degrees of freedom, and the size of the noncentrality parameter. For an elaboration of this topic, interested readers are referred to Kaplan (1995).

6 In contrast to the LISREL and AMOS programs, which identify misspecified parameters univariately, EQS determines misspecification multivariately based on the Lagrange Multiplier Test (LMTest).

Appendix

```
/TITLE
Testing for Invariance of Nonacademic SCs (Aus/Niger) «MultInv1»
Configural Model (No Constraints)
Group 1 = Australians
/SPECIFICATIONS
DATA='C:\EQS61\files\papers\anmacs\AUSREG1.ESS'; Groups=2;
VARIABLES= 77; CASES= 497;
METHODS=ML,ROBUST;
MATRIX=RAW;
/LABELS
V1=SDQ1; V2=SDQ2; V3=SDQ3; V4=SDQ4; V5=SDQ5; V6=SDQ6; V7=SDQ7;
V8=SDQ8; V9=SDQ9; V10=SDQ10; V11=SDQ11; V12=SDQ12; V13=SDQ13;
V14=SDQ14; V15=SDQ15; V16=SDQ16; V17=SDQ17; V18=SDQ18; V19=SDQ19;
V20=SDQ20; V21=SDQ21; V22=SDQ22; V23=SDQ23; V24=SDQ24; V25=SDQ25;
V26=SDQ26; V27=SDQ27; V28=SDQ28; V29=SDQ29; V30=SDQ30; V31=SDQ31;
V32=SDQ32; V33=SDQ33; V34=SDQ34; V35=SDQ35; V36=SDQ36; V37=SDQ37;
V38=SDQ38; V39=SDQ39; V40=SDQ40; V41=SDQ41; V42=SDQ42; V43=SDQ43;
V44=SDQ44; V45=SDQ45; V46=SDQ46; V47=SDQ47; V48=SDQ48; V49=SDQ49;
V50=SDQ50; V51=SDQ51; V52=SDQ52; V53=SDQ53; V54=SDQ54; V55=SDQ55;
V56=SDQ56; V57=SDQ57; V58=SDQ58; V59=SDQ59; V60=SDQ60; V61=SDQ61;
V62=SDQ62; V63=SDQ63; V64=SDQ64; V65=SDQ65; V66=SDQ66; V67=SDQ67;
V68=SDQ68; V69=SDQ69; V70=SDQ70; V71=SDQ71; V72=SDQ72; V73=SDQ73;
V74=SDQ74; V75=SDQ75; V76=SDQ76; V77=GEN;
/EQUATIONS
V1 = F1+E1;
V8 = *F1+E8;
V15 = *F1+E15;
V22 = *F1+E22;
V38 = *F1+ *F3 + E38;
V46 = *F1+E46;
V54 = *F1+E54;
V62 = *F1+E62;
V3 = F2+E3;
V10 = *F2+E10;
V24 = *F2+E24;
V32 = *F2+E32;
V40 = *F2+E40;
V48 = *F2+E48;
```

Appendix (continuated)

```
V56 = *F2+E56;
V64 = *F2+E64;
V7 = F3+E7;
V14 = *F3+E14;
V28 = *F3+E28;
V36 = *F3+E36;
V44 = *F3+E44;
V52 = *F3+E52;
V60 = *F3+E60;
V69 = *F3+E69;
V5 = F4+E5;
V19 = *F4+E19;
V26 = *F4+E26;
V34 = *F4+E34;
V42 = *F4+E42;
V50 = *F4+E50;
V58 = *F4+E58;
V66 = *F4+E66;
/VAR
E1 =*; E3 =*; E5 =*; E7 =*; E8 =*; E10 =*; E14 =*; E15 =*; E19 =*; E22 =*;
E24 =*; E26 =*; E28 =*; E32 =*; E34 =*; E36 =*; E38 =*; E40 =*; E42 =*; E44 =*;
E46 =*; E48 =*; E50 =*; E52 =*; E54 =*; E56 =*; E58 =*; E60 =*; E62 =*; E64 =*;
E66 =*; E69 =*;
F1 to F4 = *;
/COVARIANCES
F1 to F4 = *;
E26,E19 = *; E40,E24 = *;
/END
/TITLE
Testing for Invariance of Nonacademic Self-concepts (CFA of PSAp, PSAb, SSPe, SSPa)
Group 2 = Nigerians
/SPECIFICATIONS
DATA='C:\EQS61\files\papers\anmacs\nigerreg.ESS';
VARIABLES= 79; CASES= 439;
METHODS=ML,ROBUST;
MATRIX=RAW;
/LABELS
V1=ID; V2=SEX; V3=AGE; V4=SDQ1; V5=SDQ2; V6=SDQ3; V7=SDQ4; V8=SDQ5;
V9=SDQ6; V10=SDQ7; V11=SDQ8; V12=SDQ9; V13=SDQ10; V14=SDQ11;
V15=SDQ12; V16=SDQ13; V17=SDQ14; V18=SDQ15; V19=SDQ16; V20=SDQ17;
V21=SDQ18; V22=SDQ19; V23=SDQ20; V24=SDQ21; V25=SDQ22; V26=SDQ23;
V27=SDQ24; V28=SDQ25; V29=SDQ26; V30=SDQ27; V31=SDQ28; V32=SDQ29;
V33=SDQ30; V34=SDQ31; V35=SDQ32; V36=SDQ33; V37=SDQ34; V38=SDQ35;
V39=SDQ36; V40=SDQ37; V41=SDQ38; V42=SDQ39; V43=SDQ40; V44=SDQ41;
V45=SDQ42; V46=SDQ43; V47=SDQ44; V48=SDQ45; V49=SDQ46; V50=SDQ47;
V51=SDQ48; V52=SDQ49; V53=SDQ50; V54=SDQ51; V55=SDQ52; V56=SDQ53;
V57=SDQ54; V58=SDQ55; V59=SDQ56; V60=SDQ57; V61=SDQ58; V62=SDQ59;
V63=SDQ60; V64=SDQ61; V65=SDQ62; V66=SDQ63; V67=SDQ64; V68=SDQ65;
V69=SDQ66; V70=SDQ67; V71=SDQ68; V72=SDQ69; V73=SDQ70; V74=SDQ71;
V75=SDQ72; V76=SDQ73; V77=SDQ74; V78=SDQ75; V79=SDQ76;
/EQUATIONS
V4 = F1+E4;
V11 = *F1+E11;
V18 = *F1+E18;
V25 = *F1+E25;
V41 = *F1+E41;
V49 = *F1+E49;
V57 = *F1+E57;
V65 = *F1+E65;
V6 = F2+E6;
V13 = *F2+E13;
V27 = *F2+E27;
V35 = *F2+E35;
V43 = *F2+E43;
V51 = *F2+E51;
V59 = *F2+E59;
V67 = *F2+E67;
V10 = F3+E10;
V17 = *F3+E17;
V31 = *F3+E31;
V39 = *F3+E39;
V47 = *F3+E47;
V55 = *F3+E55;
V63 = *F3+E63;
V72 = *F3+E72;
V8 = F4+E8;
V22 = *F4+E22;
V29 = *F4+E29;
V37 = *F4+E37;
V45 = *F4+E45;
V53 = *F4+E53;
V61 = *F4+E61;
V69 = *F4+E69;
/VAR
E4 =*; E6 =*; E8 =*; E10 =*; E11 =*; E13 =*; E17 =*; E18 =*; E22 =*; E25 =*;
E27 =*; E29 =*; E31 =*; E35 =*; E37 =*; E39 =*; E41 =*; E43 =*; E45 =*; E47 =*;
E49 =*; E51 =*; E53 =*; E55 =*; E57 =*; E59 =*; E61 =*; E63 =*; E65 =*; E67 =*;
E69 =*; E72 =*;
F1 to F4 = *;
/COVARIANCES
F1 to F4 = *;
E29,E22 = *;
/Print
 Fit=all;
/END
```

References

Arbuckle, J.L. (1996). Full information estimation in the presence of incomplete data. In G.A. Marcoulides & R.E. Schumacker (Eds.): *Advanced structural equation modeling: Issues and techniques* (pp. 243-277). Mahwah NJ: Erlbaum.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P.M. (2005). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software (www.mvsoft.com).

Boomsma, A., & Hoogland, J.J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.): *Structural equation modeling: A festschrift in honor of Karl Jöreskög*. Lincolnwood, IL: Scientific Software.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.): *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B.M. (1988). The Self Description Questionnaire III: Testing for equivalent factorial validity across ability. *Educational and Psychological Measurement*, *48*, 397-406.

Byrne, B.M. (1996). *Measuring self-concept across the lifespan: Issues and instrumentation*. Washington, DC: American Psychological Association.

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications and programming.* Mahwah, NJ: Erlbaum.

Byrne, B.M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming.* Mahwah, NJ: Erlbaum.

Byrne, B.M. (2006). *Structural equation modeling with EQS: Basic concepts, applications and programming (2nd edition)* Mahwah, NJ: Erlbaum.

Byrne, B.M., & Shavelson, R.J. (1987). Adolescent self concept: Testing the assumption of equivalent structure across gender. *American Educational Research Journal*, *24*, 365-385.

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

Byrne, B.M., & Stewart, S.M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, *13*, 287-321.

Byrne, B.M., & Watkins, D. (2003). The issue of measurement equivalence revisited. *Journal of Cross-cultural Psychology*, *34*, 155-175.

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

Curran, P.J., West. S.G., & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16-29.

DeCarlo, L.T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*, 292-307.

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

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

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.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Kaplan, D. (1995). Statistical power in structural equation modeling. In Hoyle, R.H. (Ed.): *Structural equation modeling: Concepts, issues and applications* (pp. 100-117), Thousand Oaks, CA: Sage.

Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53-76.

Marsh, H.W. (1992). *Self Description Questionnaire (SDQ) I: A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and research monograph*. Macarthur, New South Wales, Australia: Faculty of Education, University of Western Sydney.

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

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.): *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180.

van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks: Sage.

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

Wells, L.E., & Marwell, G. (1976). *Self-esteem: Its conceptualization and measurement*. Beverly Hills: Sage.

Widaman, K.F., & Reise, S.P. (1997). Exploring the measurement equivalence of psychological instruments: Applications in the substance use domain. In K.J. Bryant, M. Windle, & S.G. West (Eds.): *The science of prevention* (pp. 281-324). Washington, D.C.: American Psychological Association.