PhUSE 2013

Paper HE06

Around the world in three statistical models: determining the level of measurement invariance across countries of a PRO instrument

Dirk Heerwegh, Business & Decision Life Sciences, Brussels, Belgium

ABSTRACT

Because many clinical trials are multinational, Patient Reported Outcome (PRO) measures are often used in different cultural settings (usually in a translated version). In such cases, evidence should show that the measurement properties are adequately similar between the different versions of the PRO instrument.

This paper focuses on the use of confirmatory factor analysis (CFA) applied across groups (multiple group CFA) to determine the level of measurement invariance. Typically, a series of three statistical models is tested, each applying more stringent invariance requirements: configural invariance, weak invariance, and strong invariance. This paper explains these concepts and provides a Sas® macro that automatically tests this series of models and outputs summary statistics that allow the user to easily determine the level of invariance supported by the data.

KEY WORDS

Confirmatory factor analysis, measurement invariance, measurement equivalence, differential item functioning.

1. BACKGROUND

Patient Reported Outcomes (PROs) are measurements of (aspects of) a patient's health status which are directly reported by the patient without any interpretation by a physician or anyone else (FDA, 2006). Because PROs often relate to abstract, multi-faceted concepts (e.g. "quality of life"), the measurement instrument usually contains multiple questions ("items") to ensure that all aspects of the concept are covered. For instance, the Medicare Health Outcomes Survey (HOS, cf. http://www.hosonline.org/Content/Default.aspx) contains 12 items to measure physical and mental health status.

Users of validated PRO instruments can follow the guidelines provided by the developer of the instrument to combine the scores on the individual items into summary scores (sometimes referred to as scale scores). For instance, for the Veterans SF-12 survey, a three step procedure is used to calculate the physical and mental component summary values based on 12 items (Spiro et al., 2004).

When a measurement instrument is modified and/or applied outside of the population for which it was originally designed, as is the case when an instrument is translated, it may be required to provide evidence that the translated instrument is still behaving equivalently (FDA, 2006). Many PRO instruments have already been translated into different languages and have been appropriately validated (at the time of writing, the PROQOLID website described 582 original instruments and 1349 translations; http://www.proqolid.org/).

For many users, the key point is to select and implement the appropriate measurement instrument in the desired language(s). After data collection, the analysis can proceed according to the guidelines made available by the issuer of the instrument. This paper however focuses on situations where there is no appropriate instrument readily available and it is necessary to modify an existing measurement instrument. In such cases, one has to establish whether a modified survey instrument has equivalent measurement properties as the original instrument. In general terms, one needs to assess the level of measurement equivalence or measurement invariance across groups. The groups can be thought of as "language groups" (when the instrument is translated), but also as "survey mode groups" (e.g. when adapting a paper questionnaire to an online questionnaire), or any other groups formed by relevant categorical variables (e.g. adolescents vs. adults, males vs. females, etc.).

A general approach to assessing measurement equivalence/invariance across groups is to analyze the data with a specific statistical method, termed Multiple Group Confirmatory Factor Analysis (MGCFA). This is an extension of Confirmatory Factor Analysis (CFA), which is comparable to Exploratory Factor Analysis (EFA, e.g. done with Sas® PROC FACTOR). Just like EFA, CFA attempts to reduce the number of observed variables into latent factors based on the commonalities in data[1]. For instance, based on the 12 items in the Veterans VR-12 survey, an EFA could be used to extract 2 factors (one factor relating to physical health and one relating to mental health). This effectively reduces the number of variables to work with in further analysis from 12 to 2.

[1] We will use the terms "factor", "latent factor" and "latent variable" interchangeably throughout the paper. These terms all refer to the entities extracted from the set of what we call "observed variables" or "manifest variables" (i.e. the variables that were actually measured with a survey item).

CFA differs from EFA in that it imposes an *a priori* model on the data, and tests the degree to which it is plausible that the data were generated by the proposed model. Again taking the Veterans VR-12 survey as an example, one can not only ask CFA to extract 2 factors from the data, but also to do so in accordance with a theoretical model which specifies which items can load on which factors. **Figure 1** shows the theoretical model imposed on the VR-12 survey. Question 2a ("moderate activities") for instance, should load on the Physical Component Summary (PCS), but not on the Mental Component Summary (MCS). Technically, this means that the factor loading of Q2a on MCS should be restricted to 0 (zero). Such restrictions give rise to an imposed factor structure (graphically represented by the absence and presence of arrows in the model). Obviously, each restriction can bring with it a certain degree of model misfit, and the purpose of the CFA analysis is to assess whether the degree of misfit is within the range of what is statistically acceptable.

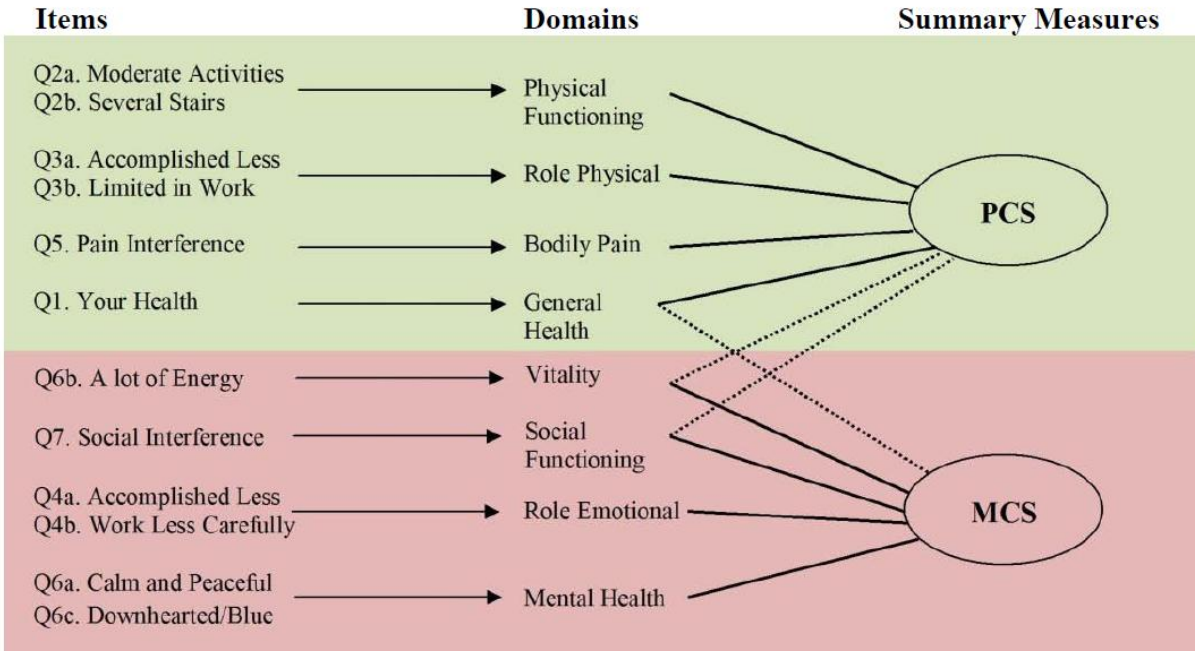

Figure 1. Theoretical model mapping the 12 items from the Veterans Rand survey to 2 summary measures (PCS-Physical Component Summary, and MCS-Mental Component Summary). Source: Centers for Medicare (2012), p. 6.

CFA can be conducted with Sas® PROC CALIS, and other software packages are available that were specifically designed to perform CFA (and related analyses that are more generally termed Structural Equation Models, abbreviated as SEM), e.g. Mplus (www.statmodel.com), LISREL (www.ssicentral.com/lisrel/index.html), AMOS (http://www-03.ibm.com/software/products/us/en/spss-amos/), EQS (http://www.mvsoft.com/), and several packages within R (http://www.r-project.org/), such as *lavaan* (http://cran.r-project.org/web/packages/lavaan/index.html), and *sem* (http://cran.r-project.org/web/packages/sem/index.html). In this paper, we will use Sas® and R.

This paper is organized as follows. In section 2, a CFA model will be fitted in a single group. This section is meant for readers who are unfamiliar with CFA and can be skipped by others. The section introduces a simple 3-factor model and the syntax in Sas® (PROC CALIS) and R (lavaan) to fit the models. It explains key elements of the output and explains how model fit can be improved by investigating modification indexes and implementing model modifications.

Section 3 introduces the concept of Measurement Equivalence / Invariance (ME/I) and explains which statistical models can be used to establish ME/I.

2. FITTING A CFA MODEL

To illustrate the syntax and results of a CFA model, we will use data from Holzinger and Swineford (1939) which comes as a pre-installed dataset in the lavaan package in R. Although this is not a dataset from the life sciences field, it is a good dataset to illustrate the basic concepts of the statistical model.

2.1 A CFA Model

The Holzinger and Swineford dataset originates from a social study in which twenty-six tests were used that intended to measure a general factor and four specific factors. Nineteen of these tests intend to measure four specific ability domains as shown in **Figure 2**.

SPATIAL TESTS
- — Visual perception test
- — Cubes
- — Paper form board
- — Lozenges

SPEED TESTS
- — Add
- — Code
- — Counting groups of dots
- — Straight and curved capitals

VERBAL TESTS
- — General information
- — Paragraph comprehension
- — Sentence completion
- — Word classification
- — Word meaning

MEMORY
- — Word recognition
- — Number recognition
- — Figure recognition
- — Object-number
- — Number-figure
- — Figure-word

Figure 2: Holzinger and Swineford variables. Source: Holzinger and Swineford (1939).

The tests were administered to seventh and eighth grade students in two schools, the Grant-White School (n = 145) and the Pasteur School (n = 156). The dataset for this analysis includes 9 variables thought to represent or measure three domains: spatial ability, verbal ability, and speed, as shown in **Figure 3**.

Spatial tests (visual ability factor)
- x1 – visual perception
- x2 – cubes
- x3 – lozenges

Verbal tests (verbal ability factor)
- x4 – paragraph comprehension
- x5 – sentence completion
- x6 – word meaning

Speed tests (speed factor)
- x7 – speeded addition
- x8 – speeded counting of dots
- x9 – speeded discrimination straight and curved capitals

Figure 3: CFA theoretical model.

It is assumed that variables x1-x3 measure a first latent factor (visual ability), variables x4-x6 measure a second latent factor (verbal ability), and variables x7-x9 measure a third latent factor (speed) (**Figure 3**). Graphically, this theoretical model can be represented as shown in **Figure 4**. This model will be fitted in Sas® (PROC CALIS) and R (lavaan) in the next 2 subsections.

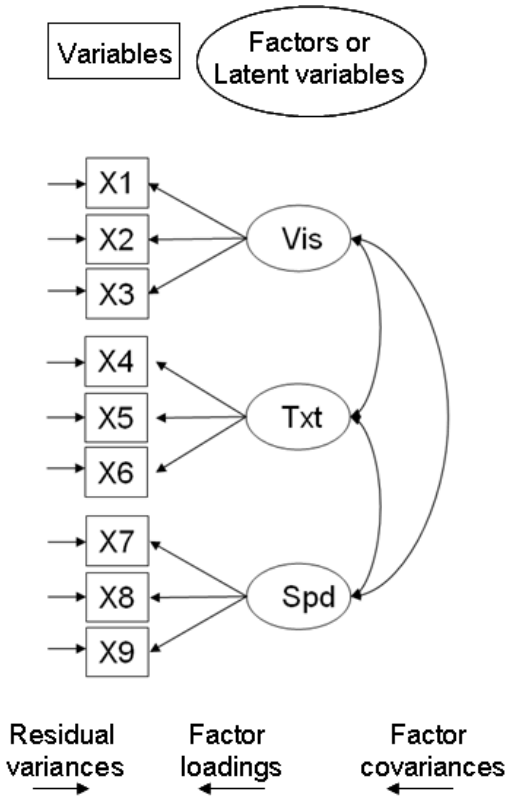

Figure 4. Graphical representation of the theoretical model. The straight arrows pointing from the factors (in ovals) to the x-variables (in squares or rectangles) represent the factor loadings; the short straight arrows pointing from the left to the x-variables represent the residual variances of the x-variables, and the curved arrows connecting the factors represent the factor covariances (the factor variances are also estimated, although usually not explicitly indicated in the model).

The equivalent representation in terms of a set of linear equations, is as follows:

$x_1 = \lambda_1 \text{ VIS} + \varepsilon_1$

$x_2 = \lambda_2 \text{ VIS} + \varepsilon_2$

$x_3 = \lambda_3 \text{ VIS} + \varepsilon_3$

$x_4 = \lambda_4 \text{ TXT} + \varepsilon_4$

$x_5 = \lambda_5 \text{ TXT} + \varepsilon_5$

$x_6 = \lambda_6 \text{ TXT} + \varepsilon_6$

$x_7 = \lambda_7 \text{ SPD} + \varepsilon_7$

$x_8 = \lambda_8 \text{ SPD} + \varepsilon_8$

$x_9 = \lambda_9 \text{ SPD} + \varepsilon_9$

These equations and the graphical representation show that a number of parameters will be estimated in this model:

- Factor loadings (estimated in the Lambda matrix)[2]
- Residual variances of the observed variables (estimated in the Epsilon matrix)
- Variances and covariances of the latent factors (estimated in the Phi matrix)

[2] Different software packages may use different matrices in which these elements are estimated. LISREL for instance, estimates residual variances in the Theta-Delta or Theta-Epsilon matrix depending on whether the observed variables are x or y variables. Note that intercepts are not necessarily estimated in single group models. In multiple group models, they are estimated in the nu matrix or in the tau matrix.

For reasons of statistical identification and to fix the scale of the latent variables, one of the lambdas (of each factor) is set to 1 or the variance of the latent factor is set to 1. The default of many software packages is to set (fix) the factor loading (lambda) of the first indicator of each latent factor to 1.

For the current model, this means that 6 lambdas, 9 residual variances, 3 factor variances, and 3 factor covariances have to be estimated for a total of 21 parameters. From 9 observed variables, there are 9(9+1) / 2 = 45 moments, implying that the model will have 45 – 21 = 24 degrees of freedom. Having a positive number of degrees of freedoms is one requirement for having a statistically defined model.

2.2. FITTING A CFA MODEL IN SAS® (PROC CALIS)

The syntax presented below is only one of the available alternatives in PROC CALIS. In this case, the LINEQS (for linear equations) was chosen, because it corresponds closely to the linear equations representation offered above. Alternatives representations available in PROC CALIS are e.g. PATH (allowing specifying models by using the causal paths of the variables – more closely resembling the graphical representation of the model) and LISMOD (which mimics the LISREL method of model specification).

```
PROC CALIS DATA=hs METHOD=ml;
 VAR x1-x9;
 LINEQS
   x1 =  1   * F1 + E1,
   x2 =  l21 * F1 + E2,
   x3 =  l31 * F1 + E3,
   x4 =  1   * F2 + E4,
   x5 =  l52 * F2 + E5,
   x6 =  l62 * F2 + E6,
   x7 =  1   * F3 + E7,
   x8 =  l83 * F3 + E8,
   x9 =  l93 * F3 + E9;
  VARIANCE
   E1-E9 = ve1-ve9,
   F1 = phi11,
   F2 = phi22,
   F3 = phi33;
  COV
   F2 F1 = phi21,
   F3 F1 = phi31,
   F3 F2 = phi32;
RUN;
```

In LINEQS, each observed variable (x1-x9) is described in terms of a factor loading (either set to 1 for the fixed factor loadings, or named l_{ij} with *l* being short for lambda, *i* = the item index [1 through 9], and *j* = the latent factor index [1 through 3]), a latent factor (F1, F2, and F3), and a residual variance (E1 through E9). The x-variables are also called the endogenous variables (variables that are influenced by other variables in the model) and the latent factors (F1-F3) are called exogenous variables (they are not influenced by any variable in the model).

In VARIANCE, the residual variances E1-E9 are repeated and assigned to variable names (ve1-ve9). The variance of the latent factors (F1-F3) are estimated and stored in the variables phi11, phi22, and phi33 (these variable names are chosen to clarify that they are located on the diagonal of the Phi matrix).

In COV, the covariances are specified. Since we are allowing covariances between the three latent factors, three additional elements in the Phi matrix are specified (the variable names are chosen to clarify that these covariances are the below-the-diagonal elements of the Phi matrix).

The results from this model are shown in Appendix 1. From the results, it seems that the global model fit is less than optimal. The chi-squared value is 85.02, which is statistically significant with 24 degrees of freedom ($p<0.0001$). Unlike “normal” chi square testing, we hope to find a non-significant chi-square test statistic, because the chi-squared value indicates the degree of discrepancy between the theoretical model and the data. While the chi-squared test is sensitive to sample size (Cheung & Rensvold, 2002), other fit indexes also suggest that the model doesn’t fit very well. The Root Mean Squared Error of Approximation (RMSEA) is above 0.06 (it is 0.0921)

and the Tucker-Lewis Index (TLI)/Non-normed Index is less than 0.95 (it is 0.896) (for an overview of recommended cut-off values, refer to Vandenberg & Lance, 2000).

At this stage, it should be noted that the model fit can potentially be improved by relaxing one or several model constraints. For instance, one could allow an observed variable to load on two latent factors, and/or one could allow a covariance between residual errors. Software packages routinely allow the user to request "modification indexes". These inform the user which constraints can be released to improve model fit by a certain degree. A modification index is essentially the expected drop in the chi-squared value if the parameter were estimated. As a lower chi-squared value is desirable, larger values of the modification index are of most interest[3]. Some packages also output Estimated Parameter Changes (EPCs). These tell us by how much a parameter value would change if the model modification were allowed. The ultimate guide in modifying the model should always be founded in theory (i.e. modifications should make sense from a theoretical point of view). Also, model modifications imply a (gradual) shift from a true confirmatory context to one of data exploration. Model modifications may therefore be useful in exploratory research (e.g. when developing or refining measurement instruments), but not in true confirmatory research.

To illustrate the use of modification indexes, we will rerun the CFA model in PROC CALIS while adding the keyword MOD to the PROC CALIS statement:

```
PROC CALIS DATA=hs METHOD=ml MOD;
```

The results are shown in Appendix 2. Sas® conveniently outputs the rank order of the 10 largest modification indexes by category (the modification indexes are called Lagrange Multipliers, which explains the abbreviation "LM Stat" in the Sas® output). We are interested in potential cross-loadings[4] (section "Stats for Paths from Exogenous Variables") and error covariances (sections "Error Variances and Covariances")[5].

Inspection of the modification indexes reveals that model fit could be improved by allowing a path from F1 on x9. Remember that x9 is a test item in the "Speed" test. More specifically, it is the "speeded discrimination straight and curved capitals" test. Clearly, there is a visual component in this test item, so it makes sense that this item would load on both the "Speed" and the "Visual" factors.

Note that there is also a suggestion to allow an error covariance between items x7 (speeded addition) and x8 (speeded counting of dots). The presence of an error covariance suggests presence of a "minor factor". This means that the latent factor F3 is not sufficiently able to explain the (strong) covariance between x7 and x8 alone. Perhaps these two items share something very specific in common which causes them to be more strongly correlated than what is explained by the model. A general rule however is to apply modifications one at the time, so we will first refit the model with an additional loading of x9 on F1, after which we can evaluate whether more modifications are still needed. A slight modification to the program is therefore made (see highlighted part):

[3] The critical value of the chi-square distribution is 3.84 for 1 degree of freedom (at alpha=0.05). Therefore, all MIs above 3.84 would in principle refer to model modifications that would significantly improve model fit. However, from a practical point of view, usually we are looking for MIs of at least 10 or even 20 to avoid over fitting the model.

[4] A cross-loading is an alternative term to mean that one observed variable loads onto more than one factor. If all observed variables only load on a single latent factor, there are no cross-loadings and the factor structure is said to be "simple" (and is sometimes preferred because it allows for an easier interpretation of the substantive meaning of factors).

[5] We are not interested in allowing paths "from Endogenous Variables" (to endogenous variables) because we want all variables to be influenced only by the latent factors, not by any other x-variables. We are also not interested in allowing "paths with New Endogenous Variables" because the theoretical model assumes that the variance and covariance between the observed variables are caused by the latent factors, and not the other way around.

```
PROC CALIS DATA=hs METHOD=ml;
 VAR x1-x9;
 LINEQS
   x1 =  1   * F1 + E1,
   x2 =  l21 * F1 + E2,
   x3 =  l31 * F1 + E3,
   x4 =  1   * F2 + E4,
   x5 =  l52 * F2 + E5,
   x6 =  l62 * F2 + E6,
   x7 =  1   * F3 + E7,
   x8 =  l83 * F3 + E8,
   x9 =  l93 * F3 + l91 * F1 + E9;
  VARIANCE
   E1-E9 = ve1-ve9,
   F1 = phi11,
   F2 = phi22,
   F3 = phi33;
  COV
   F2 F1 = phi21,
   F3 F1 = phi31,
   F3 F2 = phi32;
RUN;
```

The global fit indexes suggest that the model fit has considerably improved. The chi-square value has dropped to 52.12 with 12 degrees of freedom. Because the two models are nested, we can perform a chi-squared difference test. This suggests that the fit has improved significantly (chi² difference = 85.02 – 52.12 = 32.90 for 24 – 23 = 1 degree of freedom; $p<0.0001$)[6].

The chi-squared test statistic of the modified model is still statistically significant (chi-square = 52.12, df=23, p=0.0005), the RMSEA is 0.0651 (<0.08 but not < 0.05), and the TLI/Non-normed fit index is 0.9480 (> 0.90 but not > 0.95). This suggests that the global model fit is adequate (but not perfect). The modification indexes do not suggest any further modifications that would both improve model fit greatly (all LM Stats that we consider relevant are < 10) and would be theoretically easily defensible. Therefore, no further model modifications are applied.

2.3. FITTING A CFA MODEL IN R (LAVAAN)

The lavaan syntax is more compact than that used in PROC CALIS. As can be seen below, only three lines are needed to specify the model (lavaan will automatically fix the factor loading to 1 for x1, x4, and x7 on vis, txt, and spd, respectively). Also note that we did not use F1, F2, and F3 as factor names, but have chosen for slightly clearer labels.

```
library (lavaan)

model <- 'vis =~ x1 + x2 + x3
          txt =~ x4 + x5 + x6
          spd =~ x7 + x8 + x9'

fit<-cfa(model,data=HolzingerSwineford1939)

summary(fit, stand=T, fit.measures=T, mod=T)
```

[6] Note that the modification index was 36.29, which closely (but not exactly) corresponds to the achieved drop in the realized chi-squared value of 32.90.

The results are shown in Appendix 3. Comparison to the results from PROC CALIS indicates that the results are equal (with the only exception being the chi-squared value which is slightly different: 85.0221 in PROC CALIS and 85.306 in lavaan). The modification indexes (requested by mod=T) suggest the same modifications as those obtained from PROC CALIS. We leave it as an exercise to adapt the lavaan model to include a factor loading of x9 on vis.

3. MEASUREMENT EQUIVALENCE / INVARIANCE (ME/I)

3.1. BASIC CONCEPTS

Establishing measurement equivalence / invariance across groups is a logical prerequisite to conducting substantive cross-group comparisons (e.g. testing group mean differences or testing the equality of regression parameters across groups; Vandenberg & Lance, 2000). For instance, to evaluate the effect of a treatment on quality of life, one should test whether the treatment and control groups are responding to the measurement instrument in the same manner to exclude artificial findings. This does not mean that the test scores should be equal – we would indeed hope to find important differences in the "people" parameters (e.g. a higher mean on the latent factors in the treated patients). But we hope to find no significant differences in the "measurement" parameters – i.e. the measurement instrument should be "neutral" with respect to the different groups. Examples of bias can e.g. be found in IQ or ability tests, where certain items may give males (or females) an advantage because they are more of interest to males (or females). For instance, association tasks with gender-biased words (e.g. hunting) could give one group a (dis-)advantage.

Even if no substantive cross-group comparisons are planned, it is still necessary to establish ME/I when a PRO instrument is translated or is used in another cultural (sub-) group. It is perhaps unlikely that we are interested in comparing the mean of the latent factors across language groups, but even then we would like to know whether the instrument has equivalent measurement properties in the different groups. In this case, establishing ME/I is an aim in itself.

3.2. STATISTICAL METHODS TO ESTABLISH ME/I

A variety of statistical techniques is available to demonstrate ME/I (Wild *et al.*, 2009). Some techniques (e.g. Item Response Theory) focus on detection of Differential Item Functioning (DIF). DIF occurs when people from different groups with the same value on the latent factor give a different response to a test item. Obviously, presence of DIF implies that the measurement instrument is not fully equivalent or invariant across the groups. Another statistical method to investigate measurement equivalence is multiple group CFA (MGCFA), which is a specific type of Structural Equation Modelling (SEM; Vandenberg & Lance, 2000)[7]. In this discussion, we are focusing on MGCFA models, aimed at evaluating measurement equivalence of multi-item composite measures (i.e. PRO instruments such as VR-12 and SF-36).

To demonstrate ME/I across groups, one should test a series of hypotheses regarding measurement equivalence in a specific sequence (Vandenberg & Lance, 2000):

1. A test of "configural invariance". This baseline model evaluates whether the same pattern of fixed and free factor loadings can be specified for each group. This boils down to *simultaneously* fitting the same model in all groups while placing no equality constraints across the groups. Configural equivalence must be established in order for subsequent tests to be meaningful.
2. A test of "metric invariance". This test evaluates whether the scales are in the same metric across the groups. Technically, it implies that the factor loadings are constrained to be equal across the groups. At least partial metric invariance must be established before moving on to the next level of ME/I.
3. A test of "scalar" equivalence. This test evaluates whether the scales have the same origin across the groups. Technically, this means adding a cross-group equality constraint of the intercepts.

Note that it is possible to interpret the results from the metric and scalar equivalence tests in terms of uniform and non-uniform DIF. In case scalar equivalence is not attained (i.e. the intercepts not equal across the groups) but metric equivalence is satisfactorily demonstrated (i.e. the factor loadings or slopes are equal across groups), then one is in a situation of uniform DIF (parallel lines with different intercepts). In case metric equivalence is not obtained (the factor loadings or slopes are not equal across the groups), then non-uniform DIF is present (non-parallel lines).

Figure 5 shows the case where an observed item has a lower intercept in Group 2 as compared to Group 1. This implies that two individuals with the same score on the latent trait (e.g. with the same true ability level or with the same true life satisfaction score) are not expected to have equal scores on the observed item under investigation: despite the same true value on the latent trait, the individual from Group 1 is expected to score higher on the observed item than the individual from Group 2. This suggests that the item is biased (and the amount of bias is captured in the intercept parameter – tau). Because the lines run parallel to each other, the amount of bias is equal across the whole range of the possible values of the factor. Therefore, this corresponds to uniform DIF.

[7] A CFA model typically includes only a "measurement model", which refers to a model that specifies how latent factors are measured by observed variables. A SEM would add a structural part to that by including covariates (e.g. treatment, age, gender, etc.) that influence the latent factors.

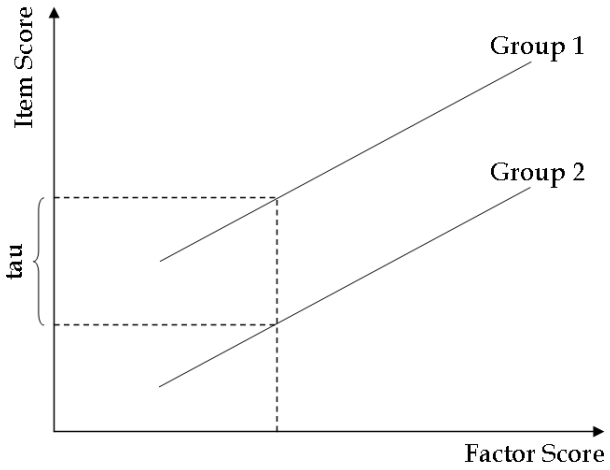

Figure 5. The case of unequal item intercepts in a multiple group CFA framework corresponds to uniform DIF.

Interestingly, this series of tests can be extended with subsequent models (Lance & Vandenberg, 2000). Most interestingly to us, one can also test whether (in addition to equal factor loadings and intercepts of the observed variables) the residual variances of the items are invariant across the groups. This would indicate equal across-groups reliability of the individual observed variables (note that an overall scale reliability can also be calculated with CFA, cf. Raykov, n.d.; Gu *et al.*, 2009).

3.3. STATISTICAL TESTS

The metric and scalar equivalence models mentioned in 3.2 are tested against the configural equivalence model and the metric equivalence model, respectively, with a chi-squared difference test[8]. Non-significant chi-squared differences indicate that the added constraints imposed on the model do not lead to a significant worsening of the model fit, and the level of ME/I is therefore supported by the data. It has been noted that, similarly to the chi-squared test statistic, the chi-squared difference test statistic is sensitive to sample size, which has led to some (relatively underdeveloped) research into appropriate cut-off values of differences in other fit statistics such as the CFI (Cheung & Rensvold, 2002).

Since the configural equivalence model is the baseline model, its fit is judged only by the overall groups chi-squared test statistic (not by comparison to another model). Since the chi-squared values of each group separately are additive, the overall groups model fit can be assessed by adding up the chi-squared values of the different groups (this is done automatically by most if not all software packages).

3.4. PARTIAL INVARIANCE

When the aim is to make substantive cross-group comparisons, it can be useful to perform tests of partial invariance. These tests are conducted with the ideas that measurement may be invariant across some but not all groups, or that some but not all of the measure's components are invariant across groups, and that implementing controls for partial measurement invariance renders across-group comparisons permissible (Vandenberg & Lance, 2000).

Even when the aim is not to make cross-group comparisons, it can be of interest to be able to pin-point the source(s) of the model misfit. This information can be used to guide further development of the measurement instrument.

Modification indexes can be used to determine which cross-group equality constraints need to be relaxed to obtain a better fitting model. Based on the modification indexes, one could for instance allow a factor loading or an intercept to be different across the groups.

3.5. A WORKED EXAMPLE WITH THE VR-12 INSTRUMENT

In this section, measurement invariance of the VR-12 instrument across two language versions will be tested in a random sample from the Medicare HOS data collected in 2009-2011. The research question at hand is whether the measurement properties are equivalent between the English and the Spanish version of the VR-12. Baseline survey results (as opposed to the follow-up survey results) are used and the analysis is restricted to a random sample of the full dataset (performing the analysis on a sample also allows performing model modifications if necessary, followed by a cross-validation of the final model on the remainder of the full sample).

The sample was drawn from the full dataset with PROC SURVEYSELECT. Only records were selected that pertain to respondents who completed the baseline survey for at least 80% by mail [c12srvdisp = 'M10'], who

[8] This is appropriate when the models are nested within each other (i.e. when one model can be derived from another model by setting model parameters equal to zero).

completed the survey themselves [c12cmpwho = 1], and who did so in the English or Spanish language [c12srvlang IN (1, 2)]]). Also, only respondents from region 2 (New York) were selected because this was the only region in which a substantial number of respondents completed the survey in Spanish. Two hundred respondents from each language group were selected at random. The programs are shown in Appendix 4; the dataset is available online at the PhUSE Wiki.

Measurement invariance is tested using the three models as specified in section 3.2. Statistical methods to establish ME/I (configural equivalence, metric equivalence, and scalar equivalence). These models are tested with the Sas® macro %measurementInvariance. This macro automatically tests these models and returns the results. Technically, it relies on the R packages lavaan and semTools to run the actual analyses. Therefore, the model specification has to be done using lavaan's syntax. The model to be fitted to the data is graphically displayed in **Figure 4**. The corresponding lavaan syntax is shown below.

```
pcs =~ q2a + q2b + q3a + q3b + q5 + q1 + q6b + q7

mcs =~ q1 + q6b + q7 + q4a + q4b + q6a + q6c
```

Close investigation of the theoretical model reveals that the items are mapped onto the summary measures PCS and MCS via domains. Technically, one could say that this is a second order CFA model in which items load on domains, and domains load onto the summary measures. The implication for the current example is that we need to take into account that some domains are measured by a single item (Bodily Pain, General Health, Vitality, and Social Functioning), while others are measured by two items (Physical Functioning, Role Physical, Role Emotional, and Mental Health). Items that belong to the same domain can be expected to correlate more strongly with each other than with the items that belong to another domain. For example, question q2a is expected to correlate more strongly with question q2b than with q3a or any of the other items that measure PCS. Expected "minor factors" (see section 2.2. Fitting a CFA model in Sas® (PROC CALIS)) can be included by adding error covariances to the model. In lavaan syntax, this is accomplished by connecting the two variable names with a double tilde. To specify an error covariance between q2a and q2b, write q2a ~~q2b.

The %measurementInvariance macro expects the measurement model to be specified in a Sas® data set. Note the addition of error covariances between the following pairs of variables in the model: q2a with q2b, q3a with q3b, q4a with q4b, and q6a with q6b.

```
DATA model;
  INFILE CARDS;
  LENGTH ln $ 500;
  INPUT ln &;
CARDS;
pcs =~ q2a + q2b + q3a + q3b + q5 + q1 + q6b + q7
mcs =~ q1 + q6b + q7 + q4a + q4b + q6a + q6c
q2a ~~ q2b
q3a ~~ q3b
q4a ~~ q4b
q6a ~~ q6b
;
RUN;
```

The macro itself is called with the following statement:

```
%measurementInvariance(model=model,data=lib.c12a_puf,group=C12SRVLANG,
mi=TRUE,export_data=TRUE);
```

The macro parameters are explained in Table 1.

Table 1. Macro parameters of the Sas® macro %measurementInvariance

Macro Parameter	Function	Possible values	Notes
model	Names the dataset that contains the model specification	Any valid Sas® dataset name	Dataset is expected in the WORK library. One record per model specification statement.
data	Names the dataset that contains the data to which the model will be fit	Any valid Sas® dataset name	A dataset at the level of the individual is required (i.e. summary data such as a covariance or correlation matrix is not currently supported).
group	Names the grouping variable	Variable has to be present in the Sas® dataset specified in the data parameter	Note that R is case sensitive, so this variable name has to be written exactly as it exists in the dataset.
mi	Whether or not modification indexes are requested	TRUE FALSE	The macro will return the modification indexes for the first non-baseline model that is not supported by the data. Default value = FALSE
mi_cutoff	Cut-off used for printing the modification indexes	Non-negative number	Only useful is mi is set to TRUE. Default value = 10
export_data	Whether or not the analysis dataset needs to be exported for R (in CSV format)	TRUE FALSE	This parameter exists to avoid exporting the same data set multiple times in case one wishes to fit multiple models on the same data. Default value = FALSE

The results of the Sas® macro %measurementInvariance are saved in a PDF document. The table that gives the results of the measurement equivalence testing (**Figure 6**) indicates that the baseline model (M1. Configural Equivalence) fits the data well. Although the chi-squared value is statistically significant (p=0.009), the CFI value is above 0.95 and the RMSEA is below 0.05. Looking at model M2, it is clear that metric equivalence could be defended based on the fit indexes of model M2 alone. However, it is also clear that the model fit has significantly worsened in comparison to model M1 (M2 vs. M1: p=0.009)[9]. A more detailed investigation of model M2 should be undertaken to assess the potentially problematic item(s), to which we will turn shortly. With an RMSEA > 0.08, model M3 does not fit the data well.

[9] The equality of the p-values for both quoted Chi-square tests is due to rounding. A chi^2 value of 127.246 for 92 degrees of freedom gives a p-value of 0.008819 (rounded to 0.009 in the output). The chi^2 difference test (155.351-127.246=28.105 for 105-92=13 degrees of freedom) gives a p-value of 0.008749 (also rounded to 0.009 in the output).

Measurement Equivalence Testing: Results of three statistical models

Model	Chisq. (df), p	CFI	RMSEA	Incremental fit
M1. Configural Equivalence	127.246 (92), p=0.009	0.989	0.047	N.A.
M2. Metric equivalence	155.351 (105), p=0.001	0.985	0.053	M2 vs M1: p=0.009
M3. Scalar equivalence	244.097 (115), p=0.000	0.960	0.081	M3 vs M2: p=0.000
M4. Equal latent means	268.096 (117), p=0.000	0.954	0.087	M4 vs M3: p=0.000

Figure 6. Results regarding the measurement equivalence testing.

The next part of the output repeats much of the information from the first table (**Figure 7**), but adds comparisons of each model with the baseline model (M3 vs. M1, M4 vs. M1). The earlier interpretations remain valid.

Measurement Equivalence Testing: Detailed results of three statistical models

Model	Chisq. (df), p	CFI	RMSEA	Incremental fit
M1. Configural Equivalence	127.246 (92), p=0.009	0.989	0.047	N.A.
M2. Metric equivalence	155.351 (105), p=0.001	0.985	0.053	M2 vs M1: p=0.009, delta CFI = 0.00
M3. Scalar equivalence	244.097 (115), p=0.000	0.960	0.081	M3 vs M2: p=0.000, delta CFI = 0.02
				M3 vs M1: p=0.000, delta CFI = 0.02
M4. Equal latent means	268.096 (117), p=0.000	0.954	0.087	M4 vs M3: p=0.000, delta CFI = 0.00
				M4 vs M1: p=0.000, delta CFI = 0.03

Figure 7. Detailed results regarding the measurement equivalence testing.

The final part of the output shows the suggested model modifications (see **Figure 8**). Because the metric equivalence model (M2) had a significantly worse fit than the baseline model, this part of the output shows the modification indexes that could be use to improve the metric equivalence model. This output shows that the factor loading of q5 could potentially be allowed to be different in the two groups. It would suggest that "pain interference" (q5) is more strongly associated with PCS in one group than in the other group. To know in which group the association is larger, the factor loadings can be investigated. This information is currently not provided by the Sas® macro. An alternative is to use the expected parameter changes, as explained below.

Suggested model modifications (Modification index value > 10) for the Metric Equivalence model

Model modification	Group number	Modification Index Value	Expected Parameter Change
pcs =~ q5	1	10.033	0.244
pcs =~ q5	2	10.033	-0.244

Figure 8. List of modification indexes for the measurement equivalence model.

The expected parameter change is 0.244 for group 1, and -0.244 for group 2 (**Figure 8**). This means that, if the model were modified and the factor loading for q5 were freed across the groups, the factor loading in group 1 is expected to be the current (equal) factor loading in groups 1 and 2 + 0.244. (Equivalently, one could say that the modified model is expected to show a factor loading for q5 in group 2 which equals the current (equal) factor loading in groups 1 and 2 - 0.244).

Because the factor PCS is "anchored" on item q2a (the factor loading of the first item that loads on PCS is set to 1 by default), PCS is scaled in the same direction as q2a (i.e. higher scores represent better health). As a consequence, higher scores on PCS are associated with lower scores on q5 (since this is a negatively worded

item), and therefore, q5 must have a negative factor loading. Hence, the factor loading should be less negative in group 1 than in group 2. So, the association between PCS and q5 is stronger in group 2 than in group 1.

Using R independently (output not reproduced), the validity of this reasoning is supported: in the model with an equal factor loading of q5, the factor loading is -1.885 (a negative factor loading as predicted). In the model in which the factor loading for q5 was freed, it was -1.612 in group 1, and -2.118 in group 2. Using the expected parameter change values, we would have estimated that the factor loading of q5 in group 1 would become -1.885 + 0.244 = -1.641, and in group 2, -1.885 – 0.244 = -2.129. The estimated values are not exactly equal to those obtained when actually fitting the model, but they are close approximations.

Whether model modifications are pursued or not depends on the purpose of the analysis (confirmatory vs. exploratory). Just to show the flexibility of the Sas® macro, the next model will allow a different factor loading for q5, by using different labels for the factor loading of q5:

```
DATA model;
  INFILE CARDS;
  LENGTH ln $ 500;
  INPUT ln &;
CARDS;
pcs =~ q2a + q2b + q3a + q3b + c(l1,l2)*q5 + q1 + q6b + q7
mcs =~ q1 + q6b + q7 + q4a + q4b + q6a + q6c
q2a ~~ q2b
q3a ~~ q3b
q4a ~~ q4b
q6a ~~ q6b
;
RUN;

%measurementInvariance(model=model,data=lib.c12a_puf,group=C12SRVLANG,
mi=TRUE);
```

Measurement Equivalence Testing: Results of three statistical models

Model	Chisq. (df), p	CFI	RMSEA	Incremental fit
M1. Configural Equivalence	127.246 (92), p=0.009	0.989	0.047	N.A.
M2. Metric equivalence	144.854 (104), p=0.005	0.987	0.048	M2 vs M1: p=0.128
M3. Scalar equivalence	232.792 (114), p=0.000	0.964	0.078	M3 vs M2: p=0.000
M4. Equal latent means	257.551 (116), p=0.000	0.957	0.084	M4 vs M3: p=0.000

Suggested model modifications (Modification index value > 10) for the Scalar Equivalence model

Model modification	Group number	Modification Index Value	Expected Parameter Change
q6b ~1	1	22.436	0.21
q6b ~1	2	22.436	-0.21
q6a ~1	1	11.246	0.153
q6a ~1	2	11.246	-0.153

Figure 9. Results regarding the measurement equivalence testing of the modified model in which item q5 is allowed to have a different factor loading across the two groups.

The results (**Figure 9**) indicate that partial metric invariance is supported by the data, since the chi-square difference test between model M1 and model M2 is not statistically significant. Therefore, all factor loadings

except that of q5 can be constrained to be equal across the two groups and we can conclude partial metric equivalence with the exception of item q5. The output shows that the Scalar equivalence model does not fit the data well (RMSEA = 0.078), and the chi-squared difference test is also statistically significant (p=0.000). The model modification indexes suggest that the intercepts of the observed items q6a and q6b are probably different in both groups. The intercept of q6b ("a lot of energy") and q6a are expected to be higher in group 1 than in group 2 ("calm and peaceful"). These model modifications can be allowed with the following syntax in which different labels for the intercepts signify that these should be estimated separately in both groups. The lavaan syntax for intercepts is *item ~ label * 1*.

```
DATA model;
  INFILE CARDS;
  LENGTH ln $ 500;
  INPUT ln &;
CARDS;
pcs =~ q2a + q2b + q3a + q3b + c(l1,l2)*q5 + q1 + q6b + q7
mcs =~ q1 + q6b + q7 + q4a + q4b + q6a + q6c
q2a ~~ q2b
q3a ~~ q3b
q4a ~~ q4b
q6a ~~ q6b
q6a~c(t1,t2)*1
q6b~c(t3,t4)*1
;
RUN;

%measurementInvariance(model=model,data=lib.c12a_puf,group=C12SRVLANG,
mi=TRUE);
```

Measurement Equivalence Testing: Results of three statistical models

Model	Chisq. (df), p	CFI	RMSEA	Incremental fit
M1. Configural Equivalence	127.246 (92), p=0.009	0.989	0.047	N.A.
M2. Metric equivalence	144.854 (104), p=0.005	0.987	0.048	M2 vs M1: p=0.128
M3. Scalar equivalence	167.597 (112), p=0.001	0.983	0.054	M3 vs M2: p=0.004
M4. Equal latent means	202.783 (114), p=0.000	0.973	0.067	M4 vs M3: p=0.000

Figure 10. Results regarding the measurement equivalence testing of the second modified model in which item q5 is allowed to have a different factor loading across the two groups and items q6a and q6b are allowed to have a different intercept across the two groups.

The results (**Figure 10**) indicate that the scalar equivalence model is acceptable (RMSEA is close to 0.05) but model fit is still significantly less good than that of the partial metric equivalence model. There are no more model modification indexes > 10, so further model modifications are not pursued in this example. It is concluded that at least partial metric equivalence is obtained for the English and Spanish language versions of this measurement instrument. Full scalar equivalence is not attained; at least two items are found to have different intercepts in both groups, and even when these across-group differences are allowed for, the global model fit indexes fail to suggest proper model fit.

Next steps in a real analysis could be to cross-validate these findings on a new sample from the original dataset, or could consist of treating some of the variables as ordinal in an attempt to arrive at a better fitting model. Neither one of these analyses will be conducted in this paper, and it should also be noted that the Sas® macro currently does not support analysis for ordinal data (see the section of planned developments).

5. CONCLUSIONS

One of the statistical methods to investigate measurement properties of an instrument across groups is to conduct multiple group CFA. This paper explained the basic concepts of CFA, and showed how CFA models can be fit in Sas® and R (lavaan) with a worked example. The paper also explained the basic concepts of measurement equivalence in the framework of multiple group CFA and provided which statistical models answer which aspects

of measurement invariance. It then presented a Sas® macro which can be used to automatically test the appropriate statistical models to assess the level of measurement equivalence. A worked example was provided with syntax and a provided dataset. It was also shown how partial measurement non-invariance can be tested and allowed for with the example, and how the model results are interpreted.

6. TECHNICAL IMPLEMENTATION USING SAS® AND R

Although Sas® does provide a procedure to fit (MG)CFA models (PROC CALIS), the current technical implementation relies on R to fit the (MG)CFA models. The main reason is because the R package semTools contains a function which performs the statistical tests that are required.

The basic workflow is that Sas® is used to pre-process the data (if needed) and to specify the measurement model, while R is used to fit the model, after which the results are returned to Sas®. The results are saved in a PDF file. The model is specified in Sas in lavaan syntax, and should be stored in a Sas® dataset.

The following sections describe the main steps in the Sas® macro %measurementInvariance.

6.1. EXPORTING THE SAS® DATASET

The Sas® dataset has to be exported for R to be able to analyze it. This is accomplished with the Sas® macro %exportlib as featured in Holland (2005), with a small variation of adding a cleaning up step at the end of the macro by deleting the temporary Sas® file created to export the datasets. The macro exports all datasets within a specified library to a CSV file format in a specified destination folder.

6.2. CREATION OF R SYNTAX

Based on the model specification supplied by the user in the Sas® program, a complete R syntax file is built. Basically, there are three parts in the R syntax file. The first part makes sure that the required packages (lavaan and semTools) are installed and loaded (and if they are not, then they are installed and/or loaded). The second part builds the model in lavaan syntax. The third part calls the semTools function measurementInvariance to actually run the models.

6.3. CALLING R AND GETTING THE RESULTS BACK INTO SAS® FOR DISPLAY

R is called from within Sas® with the X command. Note that this step requires that r.exe is defined in the Windows PATH environment variable.

After R has been executed, the results file is read into Sas® with a simple DATA step. The contents of the file is read and parsed with some basic Sas® text functions such as SCAN.

If the macro parameter MI is set to TRUE, then the macro will determine which model needs to be run in full in R to obtain modification indexes. The R syntax is automatically constructed and submitted to R. The results are read into Sas® as before.

Finally, a PROC REPORT is used to present the results. They are routed to a PDF file by default.

7. FUTURE DEVELOPMENTS

In future versions of this macro, the following features will be added:

1. Addition of a model in which the equivalence of the residual variances is tested.

2. Displaying the model results in terms of factor loadings, factor variances, residuals, etc.

3. Allowing to use ordinal variables.

Suggestions for future developments are welcome at the author's e-mail address (see the section on contact information).

8. REFERENCES

Centers for Medicare (2012). *Medicare Health Outcomes Survey*. *2009-2011 Cohort 12 Analytic Public Use File Data User's Guide*. Technical report prepared by Health Services Advisory Group. Retrieved from http://www.hosonline.org/Content/UsersGuide.aspx on 05-June-2013.

Cheung, Gordon W. & Rensvold, Roger B. (2002). *Evaluation Goodness-of-Fit Indexes for Testing Measurement Invariance*. Structural Equation Modeling, 9(2), 233-255.

FDA (2006). *Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.* Retrieved from http://www.fda.gov/cder/guidance/index.htm on 05-June-2013.

Gu, Fei; Little, Todd, & Kingston, Neal M. (2009). *Using PROC CALIS and PROC CORR to Compare Structural Equation Modeling Based Reliability Estimates and Coefficient Alpha When Assumptions are Violated*. Retrieved from http://www.mwsug.org/proceedings/2009/stats/MWSUG-2009-D04.pdf on 17-July-2013.

Holland, Philip R. (2005) *SAS to R to SAS*. Paper presented at PhUSE, Heidelberg, Germany - 10th October 2005 and at SAS Forum UK Special Interest Groups, Birmingham, UK - 2nd November 2005. Retrieved from http://www.hollandnumerics.com/SASPAPER.HTM on 10-July-2013.

Holzinger, K. J. and Swineford, F. A. (1939). *A study in factor analysis: The stability of a bi-factor solution*. Supplementary Education Monographs, 48. University of Chicago.

Raykov, T (n.d.) *Scale reliability evaluation with LISREL 8.50*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/reliabil.pdf, on 12 June 2013.

Spiro, Avron III, Rogers, William H., Qian, Shirley, and Kazis, Lewis E. (2004). *Imputing physical and mental summary scores (PCS and MCS) for the Veterans SF-12 Health Survey in the context of missing data.* Technical Report prepared by: The Health Outcomes Technologies Program, Health Services Department, Boston University School of Public Health, Boston, MA and The Institute for Health Outcomes and Policy, Center for Health Quality, Outcomes and Economic Research, Veterans Affairs Medical Center, Bedford, MA. 2004. Retrieved from www.hosonline.org/surveys/hos/download/HOS_Veterans_12_Imputation.pdf on 25 July 2013.

Vandenberg, Robert J. & Lance, Charles E. (2000). *A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research*. Organizational Research Methods, *3*(1), 4-70.

Wild, Diane; Eremenco, Sonya; Mear, Isabelle; Martin, Mona; Houchin, Caroline; Gawlicki, Mary; Hareendran, Asha; Wiklund, Ingela; Chong, Lee Yee; von Maltzahn, Robyn; Cohen, Lawrence; and Molsen, Elizabeth (2009). *Multinational Trials—Recommendations on the Translations Required, Approaches to Using the Same Language in Different Countries, and the Approaches to Support Pooling the Data: The ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force Report*. Value in Health (12), 4, pp. 430-440.

SOFTWARE USED

The Sas® macros and programs were developed and tested on Sas® 9.2 on a Windows® XP system.

R version 3.0.1 was used on a Windows® system. Lavaan version 0.5-13 and semTools version 0.4-0 were used.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dirk Heerwegh
Business & Decision Life Sciences
Sint-Lambertusstraat 141 Rue Saint-Lambert
1200 Brussels
Tel +32 2 774 11 00
Fax +32 2 774 11 99
Email: dirk.heerwegh@businessdecision.com
Web: www.businessdecision-lifesciences.com

APPENDIX 1. RESULTS FROM SAS® PROC CALIS TO FIT A CFA MODEL

The CALIS Procedure
Covariance Structure Analysis: Model and Initial Values

Modeling Information

Data Set	WORK.HS
N Records Read	301
N Records Used	301
N Obs	301
Model Type	LINEQS
Analysis	Covariances

Variables in the Model

Endogenous	Manifest	x1 x2 x3 x4 x5 x6 x7 x8 x9
	Latent	
Exogenous	Manifest	
	Latent	F1 F2 F3
	Error	E1 E2 E3 E4 E5 E6 E7 E8 E9

Number of Endogenous Variables = 9
Number of Exogenous Variables = 12

Initial Estimates for Linear Equations

```
x1 =    1.0000 F1  +  1.0000 E1
x2 =         .*F1  +  1.0000 E2
               l21
x3 =         .*F1  +  1.0000 E3
               l31
x4 =    1.0000 F2  +  1.0000 E4
x5 =         .*F2  +  1.0000 E5
               l52
x6 =         .*F2  +  1.0000 E6
               l62
x7 =    1.0000 F3  +  1.0000 E7
x8 =         .*F3  +  1.0000 E8
               l83
x9 =         .*F3  +  1.0000 E9
               l93
```

Initial Estimates for Variances of Exogenous Variables

Variable Type	Variable	Parameter	Estimate
Error	E1	ve1	.
	E2	ve2	.
	E3	ve3	.
	E4	ve4	.
	E5	ve5	.
	E6	ve6	.
	E7	ve7	.
	E8	ve8	.

The CALIS Procedure
Covariance Structure Analysis: Model and Initial Values

Initial Estimates for Variances of Exogenous Variables

Variable Type	Variable	Parameter	Estimate
Error	E9	ve9	.
Latent	F1	phi11	.
	F2	phi22	.
	F3	phi33	.

Initial Estimates for Covariances Among Exogenous Variables

Var1	Var2	Parameter	Estimate
F2	F1	phi21	.
F3	F1	phi31	.
F3	F2	phi32	.

The CALIS Procedure
Covariance Structure Analysis: Descriptive Statistics

Simple Statistics

Variable	Mean	Std Dev
x1	4.93577	1.16743
x2	6.08804	1.17745
x3	2.25042	1.13098
x4	3.06091	1.16412
x5	4.34053	1.29047
x6	2.18557	1.09560
x7	4.18590	1.08953
x8	5.52708	1.01262
x9	5.37412	1.00915

The CALIS Procedure
Covariance Structure Analysis: Optimization

Initial Estimation Methods

1 Instrumental Variables Method
2 McDonald Method

Optimization Start
Parameter Estimates

N	Parameter	Estimate	Gradient
1	l21	0.57517	0.09136
2	l31	0.59311	-0.07898
3	l52	1.11837	-0.00374
4	l62	0.94627	0.02649
5	l83	1.25936	0.01279
6	l93	1.09091	0.00963
7	ve1	0.44352	-0.03362
8	ve2	1.08224	-0.04850
9	ve3	0.95569	0.04909
10	ve4	0.38789	0.02597
11	ve5	0.45550	0.01279
12	ve6	0.33423	-0.07021
13	ve7	0.81928	0.02038
14	ve8	0.44205	-0.04036
15	ve9	0.58066	-0.01979
16	phi11	0.91938	0.03391
17	phi22	0.96727	0.01577
18	phi33	0.36781	0.01986
19	phi21	0.40818	-0.06633
20	phi31	0.24426	-0.05813
21	phi32	0.16849	0.05288

Value of Objective Function = 0.3000928408

The CALIS Procedure
Covariance Structure Analysis: Optimization

Levenberg-Marquardt Optimization

Scaling Update of More (1978)

Parameter Estimates	21
Functions (Observations)	45

Optimization Start

Active Constraints	0	Objective Function	0.3000928408
Max Abs Gradient Element	0.0913584851	Radius	1

Iter	Rest arts	Func Calls	Act Con	Objective Function	Obj Fun Change	Max Abs Gradient Element	Lambda	Actual Over Pred Change
1	0	4	0	0.28524	0.0149	0.0654	0	0.758
2	0	6	0	0.28364	0.00159	0.0120	0	0.718
3	0	8	0	0.28346	0.000184	0.00844	0	0.734
4	0	10	0	0.28342	0.000036	0.00345	0	0.859
5	0	12	0	0.28341	9.083E-6	0.00134	0	1.115
6	0	14	0	0.28341	3.204E-6	0.00132	0	1.403
7	0	16	0	0.28341	1.347E-6	0.000751	0	1.580
8	0	18	0	0.28341	6.073E-7	0.000578	0	1.650
9	0	20	0	0.28341	2.805E-7	0.000371	0	1.673
10	0	22	0	0.28341	1.307E-7	0.000264	0	1.681
11	0	24	0	0.28341	6.107E-8	0.000177	0	1.683
12	0	26	0	0.28341	2.858E-8	0.000123	0	1.684
13	0	28	0	0.28341	1.339E-8	0.000083	0	1.684
14	0	30	0	0.28341	6.275E-9	0.000057	0	1.685
15	0	32	0	0.28341	2.942E-9	0.000039	0	1.685
16	0	34	0	0.28341	1.379E-9	0.000027	0	1.685

Optimization Results

Iterations	16	Function Calls	37
Jacobian Calls	18	Active Constraints	0
Objective Function	0.2834070503	Max Abs Gradient Element	0.0000268271
Lambda	0	Actual Over Pred Change	1.6847948978
Radius	0.000204991		

Convergence criterion (GCONV=1E-8) satisfied.

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Fit Summary

Modeling Info	N Observations	301
	N Variables	9
	N Moments	45
	N Parameters	21
	N Active Constraints	0
	Baseline Model Function Value	3.0527
	Baseline Model Chi-Square	915.7989
	Baseline Model Chi-Square DF	36
	Pr > Baseline Model Chi-Square	<.0001
Absolute Index	Fit Function	0.2834
	Chi-Square	85.0221
	Chi-Square DF	24
	Pr > Chi-Square	<.0001
	Z-Test of Wilson & Hilferty	5.5462
	Hoelter Critical N	129
	Root Mean Square Residual (RMSR)	0.0825
	Standardized RMSR (SRMSR)	0.0652
	Goodness of Fit Index (GFI)	0.9433
Parsimony Index	Adjusted GFI (AGFI)	0.8937
	Parsimonious GFI	0.6289
	RMSEA Estimate	0.0921
	RMSEA Lower 90% Confidence Limit	0.0713
	RMSEA Upper 90% Confidence Limit	0.1137
	Probability of Close Fit	0.0007
	ECVI Estimate	0.4282
	ECVI Lower 90% Confidence Limit	0.3460
	ECVI Upper 90% Confidence Limit	0.5367
	Akaike Information Criterion	127.0221
	Bozdogan CAIC	225.8714
	Schwarz Bayesian Criterion	204.8714
	McDonald Centrality	0.9036
Incremental Index	Bentler Comparative Fit Index	0.9306
	Bentler-Bonett NFI	0.9072
	Bentler-Bonett Non-normed Index	0.8960
	Bollen Normed Index Rho1	0.8607
	Bollen Non-normed Index Delta2	0.9316
	James et al. Parsimonious NFI	0.6048

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Linear Equations

x1	=	1.0000 F1	+	1.0000 E1
x2	=	0.5535*F1	+	1.0000 E2
Std Err		0.0998 l21		
t Value		5.5443		
x3	=	0.7294*F1	+	1.0000 E3
Std Err		0.1093 l31		
t Value		6.6735		
x4	=	1.0000 F2	+	1.0000 E4
x5	=	1.1131*F2	+	1.0000 E5
Std Err		0.0655 l52		
t Value		16.9860		
x6	=	0.9261*F2	+	1.0000 E6
Std Err		0.0555 l62		
t Value		16.6749		
x7	=	1.0000 F3	+	1.0000 E7
x8	=	1.1800*F3	+	1.0000 E8
Std Err		0.1653 l83		
t Value		7.1401		
x9	=	1.0814*F3	+	1.0000 E9
Std Err		0.1514 l93		
t Value		7.1428		

Estimates for Variances of Exogenous Variables

Variable Type	Variable	Parameter	Estimate	Standard Error	t Value
Error	E1	ve1	0.55088	0.11417	4.82499
	E2	ve2	1.13762	0.10223	11.12776
	E3	ve3	0.84714	0.09108	9.30136
	E4	ve4	0.37241	0.04796	7.76557
	E5	ve5	0.44774	0.05869	7.62961
	E6	ve6	0.35739	0.04325	8.26329
	E7	ve7	0.80203	0.08179	9.80614
	E8	ve8	0.48928	0.07457	6.56159
	E9	ve9	0.56806	0.07109	7.99080
Latent	F1	phi11	0.81202	0.14619	5.55444
	F2	phi22	0.98276	0.11267	8.72268
	F3	phi33	0.38506	0.08664	4.44415

Covariances Among Exogenous Variables

Var1	Var2	Parameter	Estimate	Standard Error	t Value
F2	F1	phi21	0.40960	0.07389	5.54316
F3	F1	phi31	0.26309	0.05656	4.65165
F3	F2	phi32	0.17407	0.04956	3.51215

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Squared Multiple Correlations

Variable	Error Variance	Total Variance	R-Square
x1	0.55088	1.36290	0.5958
x2	1.13762	1.38639	0.1794
x3	0.84714	1.27911	0.3377
x4	0.37241	1.35517	0.7252
x5	0.44774	1.66532	0.7311
x6	0.35739	1.20035	0.7023
x7	0.80203	1.18708	0.3244
x8	0.48928	1.02539	0.5228
x9	0.56806	1.01839	0.4422

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Standardized Results for Linear Equations

```
x1      =    0.7719 F1  +  1.0000 E1
Std Err      0.0551
t Value     14.0176
x2      =    0.4236*F1  +  1.0000 E2
Std Err      0.0597 l21
t Value      7.0933
x3      =    0.5811*F1  +  1.0000 E3
Std Err      0.0552 l31
t Value     10.5217
x4      =    0.8516 F2  +  1.0000 E4
Std Err      0.0226
t Value     37.7130
x5      =    0.8551*F2  +  1.0000 E5
Std Err      0.0224 l52
t Value     38.2093
x6      =    0.8380*F2  +  1.0000 E6
Std Err      0.0234 l62
t Value     35.8215
x7      =    0.5695 F3  +  1.0000 E7
Std Err      0.0532
t Value     10.6971
x8      =    0.7231*F3  +  1.0000 E8
Std Err      0.0506 l83
t Value     14.2859
x9      =    0.6650*F3  +  1.0000 E9
Std Err      0.0512 l93
t Value     12.9930
```

Standardized Results for Variances of Exogenous Variables

Variable Type	Variable	Parameter	Estimate	Standard Error	t Value
Error	E1	ve1	0.40420	0.08501	4.75481
	E2	ve2	0.82056	0.05059	16.21870
	E3	ve3	0.66229	0.06419	10.31706
	E4	ve4	0.27481	0.03846	7.14555
	E5	ve5	0.26886	0.03827	7.02542
	E6	ve6	0.29774	0.03921	7.59365
	E7	ve7	0.67563	0.06065	11.14030
	E8	ve8	0.47716	0.07320	6.51900
	E9	ve9	0.55781	0.06807	8.19500
Latent	F1	phi11	1.00000		
	F2	phi22	1.00000		
	F3	phi33	1.00000		

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Standardized Results for Covariances Among Exogenous Variables

Var1	Var2	Parameter	Estimate	Standard Error	t Value
F2	F1	phi21	0.45851	0.06389	7.17706
F3	F1	phi31	0.47050	0.07295	6.44965
F3	F2	phi32	0.28297	0.06884	4.11031

APPENDIX 2. MODIFICATION INDEXES FROM SAS® PROC CALIS

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

NOTE: All parameters in the model are significant. No parameter can be dropped in the Wald tests.

Covariance Structure Analysis: Maximum Likelihood Estimation

Rank Order of the 10 Largest LM Stat for Paths from Endogenous Variables

To	From	LM Stat	Pr > ChiSq	Parm Change
x8	x7	34.02316	<.0001	0.67111
x7	x8	34.02156	<.0001	1.10006
x9	x1	24.49630	<.0001	0.26623
x9	x3	15.78227	<.0001	0.19719
x7	x2	15.45572	<.0001	-0.19437
x8	x9	14.89563	0.0001	-0.74728
x9	x8	14.89309	0.0001	-0.86753
x7	x1	14.20399	0.0002	-0.21204
x3	x5	13.89650	0.0002	-0.21633
x5	x3	11.36748	0.0007	-0.14840

Rank Order of the 10 Largest LM Stat for Paths from Exogenous Variables

To	From	LM Stat	Pr > ChiSq	Parm Change
x9	F1	36.29191	<.0001	0.57698
x7	F1	18.56781	<.0001	-0.42183
x3	F2	9.12046	0.0025	-0.27164
x1	F2	8.87300	0.0029	0.35034
x5	F1	7.41609	0.0065	-0.20990
x9	F2	4.78148	0.0288	0.13841
x8	F1	4.27960	0.0386	-0.21038
x8	F2	3.34789	0.0673	-0.12078
x6	F1	2.83348	0.0923	0.11141
x2	F3	1.57467	0.2095	-0.19831

Rank Order of the 10 Largest LM Stat for Paths with New Endogenous Variables

To	From	LM Stat	Pr > ChiSq	Parm Change
F3	x9	34.03241	<.0001	-0.86852
F1	x9	28.79389	<.0001	0.52731
F1	x7	21.26997	<.0001	-0.33606
F3	x7	14.89274	0.0001	0.41475
F2	x3	9.09606	0.0026	-0.24118
F1	x1	8.50494	0.0035	-0.98470
F1	x5	8.48663	0.0036	-0.31059
F2	x1	8.24107	0.0041	0.45649
F2	x5	6.19997	0.0128	0.63237
F3	x8	5.16356	0.0231	0.41767

NOTE: There is no parameter to free in the default LM tests for the covariances of exogenous variables. Ranking is not displayed.

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Rank Order of the 10 Largest LM Stat for Error Variances and Covariances

Var1	Var2	LM Stat	Pr > ChiSq	Parm Change
E8	E7	34.02398	<.0001	0.53825
E9	E8	14.89464	0.0001	-0.42449
E7	E2	8.88815	0.0029	-0.18333
E3	E2	8.50381	0.0035	0.21897
E5	E3	7.83188	0.0051	-0.13053
E9	E1	7.31148	0.0069	0.13836
E6	E4	6.20014	0.0128	-0.23559
E7	E4	5.90048	0.0151	0.09851
E7	E1	5.40115	0.0201	-0.12954
E9	E7	5.16440	0.0231	-0.18731

APPENDIX 3. RESULTS FROM R (LAVAAN) TO FIT A CFA MODEL

```
lavaan (0.5-13) converged normally after  35 iterations

  Number of observations                           301

  Estimator                                         ML
  Minimum Function Test Statistic               85.306
  Degrees of freedom                                24
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic              918.852
  Degrees of freedom                                36
  P-value                                        0.000

Full model versus baseline model:

  Comparative Fit Index (CFI)                    0.931
  Tucker-Lewis Index (TLI)                       0.896

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3737.745
  Loglikelihood unrestricted model (H1)      -3695.092

  Number of free parameters                         21
  Akaike (AIC)                                7517.490
  Bayesian (BIC)                              7595.339
  Sample-size adjusted Bayesian (BIC)         7528.739

Root Mean Square Error of Approximation:

  RMSEA                                          0.092
  90 Percent Confidence Interval          0.071  0.114
```

```
  P-value RMSEA <= 0.05                          0.001

Standardized Root Mean Square Residual:

  SRMR                                           0.065

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  vis =~
    x1                1.000                               0.900    0.772
    x2                0.554    0.100    5.554    0.000    0.498    0.424
    x3                0.729    0.109    6.685    0.000    0.656    0.581
  txt =~
    x4                1.000                               0.990    0.852
    x5                1.113    0.065   17.014    0.000    1.102    0.855
    x6                0.926    0.055   16.703    0.000    0.917    0.838
  spd =~
    x7                1.000                               0.619    0.570
    x8                1.180    0.165    7.152    0.000    0.731    0.723
    x9                1.082    0.151    7.155    0.000    0.670    0.665

Covariances:
  vis ~~
    txt               0.408    0.074    5.552    0.000    0.459    0.459
    spd               0.262    0.056    4.660    0.000    0.471    0.471
  txt ~~
    spd               0.173    0.049    3.518    0.000    0.283    0.283

Variances:
    x1                0.549    0.114                      0.549    0.404
    x2                1.134    0.102                      1.134    0.821
    x3                0.844    0.091                      0.844    0.662
    x4                0.371    0.048                      0.371    0.275
    x5                0.446    0.058                      0.446    0.269
    x6                0.356    0.043                      0.356    0.298
    x7                0.799    0.081                      0.799    0.676
    x8                0.488    0.074                      0.488    0.477
    x9                0.566    0.071                      0.566    0.558
    vis               0.809    0.145                      1.000    1.000
    txt               0.979    0.112                      1.000    1.000
    spd               0.384    0.086                      1.000    1.000
```

Modification Indices:

	lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
1	vis	=~	x1	NA	NA	NA	NA	NA
2	vis	=~	x2	0.000	0.000	0.000	0.000	0.000
3	vis	=~	x3	0.000	0.000	0.000	0.000	0.000
4	vis	=~	x4	1.211	0.077	0.069	0.059	0.059
5	vis	=~	x5	7.441	-0.210	-0.189	-0.147	-0.147
6	vis	=~	x6	2.843	0.111	0.100	0.092	0.092
7	vis	=~	x7	18.631	-0.422	-0.380	-0.349	-0.349
8	vis	=~	x8	4.295	-0.210	-0.189	-0.187	-0.187
9	vis	=~	x9	36.411	0.577	0.519	0.515	0.515
10	txt	=~	x1	8.903	0.350	0.347	0.297	0.297
11	txt	=~	x2	0.017	-0.011	-0.011	-0.010	-0.010
12	txt	=~	x3	9.151	-0.272	-0.269	-0.238	-0.238
13	txt	=~	x4	NA	NA	NA	NA	NA
14	txt	=~	x5	0.000	0.000	0.000	0.000	0.000
15	txt	=~	x6	0.000	0.000	0.000	0.000	0.000
16	txt	=~	x7	0.098	-0.021	-0.021	-0.019	-0.019
17	txt	=~	x8	3.359	-0.121	-0.120	-0.118	-0.118
18	txt	=~	x9	4.796	0.138	0.137	0.136	0.136
19	spd	=~	x1	0.014	0.024	0.015	0.013	0.013
20	spd	=~	x2	1.580	-0.198	-0.123	-0.105	-0.105
21	spd	=~	x3	0.716	0.136	0.084	0.075	0.075
22	spd	=~	x4	0.003	-0.005	-0.003	-0.003	-0.003
23	spd	=~	x5	0.201	-0.044	-0.027	-0.021	-0.021
24	spd	=~	x6	0.273	0.044	0.027	0.025	0.025
25	spd	=~	x7	NA	NA	NA	NA	NA
26	spd	=~	x8	0.000	0.000	0.000	0.000	0.000
27	spd	=~	x9	0.000	0.000	0.000	0.000	0.000
28	x1	~~	x1	0.000	0.000	0.000	0.000	0.000
29	x1	~~	x2	3.606	-0.184	-0.184	-0.134	-0.134
30	x1	~~	x3	0.935	-0.139	-0.139	-0.105	-0.105
31	x1	~~	x4	3.554	0.078	0.078	0.058	0.058
32	x1	~~	x5	0.522	-0.033	-0.033	-0.022	-0.022
33	x1	~~	x6	0.048	0.009	0.009	0.007	0.007
34	x1	~~	x7	5.420	-0.129	-0.129	-0.102	-0.102
35	x1	~~	x8	0.634	-0.041	-0.041	-0.035	-0.035
36	x1	~~	x9	7.335	0.138	0.138	0.117	0.117
37	x2	~~	x2	0.000	0.000	0.000	0.000	0.000
38	x2	~~	x3	8.532	0.218	0.218	0.164	0.164
39	x2	~~	x4	0.534	-0.034	-0.034	-0.025	-0.025
40	x2	~~	x5	0.023	-0.008	-0.008	-0.005	-0.005
41	x2	~~	x6	0.785	0.039	0.039	0.031	0.031
42	x2	~~	x7	8.918	-0.183	-0.183	-0.143	-0.143
43	x2	~~	x8	0.054	-0.012	-0.012	-0.010	-0.010
44	x2	~~	x9	1.895	0.075	0.075	0.063	0.063
45	x3	~~	x3	0.000	0.000	0.000	0.000	0.000

46	x3	~~	x4	0.142	-0.016	-0.016	-0.012	-0.012
47	x3	~~	x5	7.858	-0.130	-0.130	-0.089	-0.089
48	x3	~~	x6	1.855	0.055	0.055	0.044	0.044
49	x3	~~	x7	0.638	-0.044	-0.044	-0.036	-0.036
50	x3	~~	x8	0.059	-0.012	-0.012	-0.011	-0.011
51	x3	~~	x9	4.126	0.102	0.102	0.089	0.089
52	x4	~~	x4	0.000	0.000	0.000	0.000	0.000
53	x4	~~	x5	2.534	0.186	0.186	0.124	0.124
54	x4	~~	x6	6.220	-0.235	-0.235	-0.185	-0.185
55	x4	~~	x7	5.920	0.098	0.098	0.078	0.078
56	x4	~~	x8	3.805	-0.069	-0.069	-0.059	-0.059
57	x4	~~	x9	0.196	-0.016	-0.016	-0.014	-0.014
58	x5	~~	x5	0.000	0.000	0.000	0.000	0.000
59	x5	~~	x6	0.916	0.101	0.101	0.072	0.072
60	x5	~~	x7	1.233	-0.049	-0.049	-0.035	-0.035
61	x5	~~	x8	0.347	0.023	0.023	0.018	0.018
62	x5	~~	x9	0.999	0.040	0.040	0.031	0.031
63	x6	~~	x6	0.000	0.000	0.000	0.000	0.000
64	x6	~~	x7	0.259	-0.020	-0.020	-0.017	-0.017
65	x6	~~	x8	0.275	0.018	0.018	0.016	0.016
66	x6	~~	x9	0.097	-0.011	-0.011	-0.010	-0.010
67	x7	~~	x7	0.000	0.000	0.000	0.000	0.000
68	x7	~~	x8	34.145	0.536	0.536	0.488	0.488
69	x7	~~	x9	5.183	-0.187	-0.187	-0.170	-0.170
70	x8	~~	x8	0.000	0.000	0.000	0.000	0.000
71	x8	~~	x9	14.946	-0.423	-0.423	-0.415	-0.415
72	x9	~~	x9	0.000	0.000	0.000	0.000	0.000
73	vis	~~	vis	0.000	0.000	0.000	0.000	0.000
74	vis	~~	txt	0.000	0.000	0.000	0.000	0.000
75	vis	~~	spd	0.000	0.000	0.000	0.000	0.000
76	txt	~~	txt	0.000	0.000	0.000	0.000	0.000
77	txt	~~	spd	0.000	0.000	0.000	0.000	0.000
78	spd	~~	spd	0.000	0.000	0.000	0.000	0.000

APPENDIX 4. MEDICARE HOS SAMPLE DATASET: SELECTION OF THE DATA

The dataset was downloaded from http://www.hosonline.org/surveys/hos/download/C12A_PUF.zip and read into Sas® with the program provided by HOS at http://www.hosonline.org/surveys/hos/download/C12A_puf_import_code.sas. Both hyperlinks are accessible via http://www.hosonline.org/Content/DataFiles.aspx (Section PUF data files, Cohort 12, Analytic Data and Analytic Data Import Code).

The below code assumes a library called “d” in which the imported dataset is stored.

```
data frame;
  set d.c12a_puf (where=(p12plregcde=2 AND c12srvdisp = 'M10' AND c12cmpwho
= 1 AND c12srvlang IN (1, 2)));
  rename
c12vrgenhth = q1
c12vrmact = q2a
c12vrstair = q2b
c12vrpaccl = q3a
c12vrpwork = q3b
c12vrmaccl = q4a
c12vrmwork = q4b
c12vrpain = q5
c12vrcalm = q6a
c12vrenergy = q6b
c12vrdown = q6c
c12vrsact = q7;
keep case_id p12plregcde age race educ gender c12srvlang c12vrgenhth
c12vrmact c12vrstair c12vrpaccl c12vrpwork c12vrmaccl c12vrmwork c12vrpain
c12vrcalm c12vrenergy c12vrdown c12vrsact;
run;

proc sort data=frame;
  by c12srvlang;
run;

proc surveyselect data=frame out=sample n=200 seed=8745134;
  strata c12srvlang;
run;
```