**Appendix 1 – Target words for the test battery in González-Fernández and Schmitt (2020)**

|  |  |  |
| --- | --- | --- |
|  | **Target words** | **Frequency band** |
|  | *Mean (v)* | 1K |
|  | *Close (v)* | 1K |
|  | *Hard (a)* | 1K |
|  | *Development (n)* | 2K |
|  | *Season (n)* | 2K |
|  | *Bank (n)* | 1K |
|  | *Challenge (n)* | 2K |
|  | *Character (n)* | 2K |
|  | *Fresh (a)* | 1K |
|  | *Bright (a)* | 1K |
|  | *Broad (a)* | 2K |
|  | *Employ (v)* | 1K |
|  | *Distinction (n)* | 3K |
|  | *Charm (n)* | 2K |
|  | *Terminal (a)* | 4K |
|  | *Fulfil (v)* | 3K |
|  | *Grate (v)* | 5K |
|  | *Redeem (v)* | 6K |
|  | *Draught (n)* | 9K |
|  | *Indent (v)* | 8K |

 *Note*: *v* = verb, *a* = adjective, *n* = noun.

**Appendix 2 – Descriptive statistics by language group**

Prior to fitting the hypothesised second-order model, the means, standard deviations and correlations on the observed indicators were estimated and inspected for the whole sample and by language group (i.e., L1 Chinese and L1 Spanish). Table 1 presents the descriptive statistics and correlations by language group.

The Spearman correlations (data found to be univariate non-normal) on Table 1 below show that the various word-knowledge aspects correlate positively and highly with each other in both language groups. However, while they are all large correlations for the L1-Spanish participants, these correlations are mainly medium-to-large in the L1-Chinese learners. This pattern of somewhat smaller correlations among the word-knowledge aspects for the L1-Chinese participants suggests that the interconnection among word-knowledge components varies slightly as a function of L1. On the contrary, the overall scores on the eight measures do not seem to differ greatly across the two groups. This indicates that the two learner samples were rather homogenous insofar that they both included learners of different proficiency levels, from beginners to advanced, resulting in similar average levels of word knowledge.

**Table 1** Correlations, means (%) and standard deviations for the vocabulary measures as a function of language group (results for L1 Spanish learners below the diagonal and L1 Chinese learners above the diagonal).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | FM Recall | FM Recog | Deriv Recall | Deriv Recog | MM Recall | MM Recog | Collo Recall | Collo Recog | **L1 Chinese** *M* | *SD* |
| FM Recall | - | .440\*\* | .483\*\* | .439\*\* | .611\*\* | .466\*\* | .454\*\* | .433\*\* | 53.15 | 11.99 |
| FM Recog | .773\*\* | - | .579\*\* | .551\*\* | .570\*\* | .620\*\* | .492\*\* | .485\*\* | 79.38 | 9.53 |
| Deriv Recall | .811\*\* | .803\*\* | - | .711\*\* | .721\*\* | .602\*\* | .521\*\* | .561\*\* | 51.57 | 10.26 |
| Deriv Recog | .786\*\* | .828\*\* | .945\*\* | - | .656\*\* | .592\*\* | .533\*\* | .493\*\* | 65.82 | 10.74 |
| MM Recall | .747\*\* | .774\*\* | .845\*\* | .818\*\* | - | .722\*\* | .588\*\* | .603\*\* | 50.29 | 10.00 |
| MM Recog  | .720\*\* | .766\*\* | .835\*\* | .831\*\* | .848\*\* | - | .550\*\* | .497\*\* | 64.90 | 11.62 |
| Collo Recall | .772\*\* | .749\*\* | .827\*\* | .812\*\* | .817\*\* | .771\*\* | - | .701 | 57.21 | 15.38 |
| Collo Recog | .678\*\* | .700\*\* | .763\*\* | .768\*\* | .752\*\* | .737\*\* | .806\*\* | - | 77.32 | 12.50 |
| **L1 Spanish** |  |  |  |  |  |  |  |  |  |  |
| *M* | 53.61 | 82.81 | 51.64 | 61.08 | 49.69 | 70.76 | 60.66 | 79.69 |  |  |
| *SD* | 18.91 | 13.79 | 20.86 | 19.90 | 16.12 | 15.18 | 19.05 | 17.39 |  |  |

Spearman: \*\**p* < .01

*FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition.*

**Appendix 3 – Preliminary CFA analyses**

Prior to fitting the hypothesised multidimensional model (Model 1), a series of preliminary analyses were conducted in order to validate the word-knowledge measures for the sample. These analyses involved assessing the fit of each latent word-knowledge component and its indicators separately by means of individual CFA models, in order to test whether the measurement instruments of the individual word-knowledge components represented the aspects intended to be measured and to validate the word-knowledge factors. The multivariate normality of the data was assessed by means of Mardia’s Multivariate Normality Test (Mardia, 1970), which indicated that the data was multivariate non-normal. Thus, robust estimation methods were employed for parameter estimation (i.e., WLSMV for categorical variables and MLR for continuous ones).

Eight independent CFA were first tested, one for each word-knowledge aspect. In these models, the results of the 20 target words used to measure each of the recall and recognition aspects were loaded as indicators onto their corresponding word- knowledge factor. The scores for the individual target words in each test were categorical, and therefore the WLSMV estimator was employed. Given that the aim of the study is to analyse participants with different proficiencies (from beginners to advanced) and language-learning contexts (L2 immersion, informal or formal instruction), it was expected that performance in some items varied across participants. Yet, all items were kept in the models, as removing worse-performing items would have caused a loss of valuable information for the purpose of analysing vocabulary knowledge across EFL learners in general. The different measures of word-knowledge aspects achieved reliability and a good model fit (see Table 2). This indicates that the measurement models are suitable, and thus, the measures of the various word knowledge aspects are appropriate and valid psychometric constructs.

**Table 2** Fit indices and reliability forthe CFA models with 20 target words as indicators for each word knowledge aspect (L1 Chinese, *n* = 170)

|  |  |
| --- | --- |
|  |  |
|  |  | **Model fit indices** |
|  | χ2 | *df* | *p* value | χ2/*df* | CFI | RMSEA(90% CI) | WRMR | Reliability |
| *Acceptable fit* |  |  | *>.05* | *<3* | *>.95* | *<.05 / <.08* | *<.1.0* | *α >.7* |
| **Form–Meaning Recall model** | 156.299 | 136 | .112 | 1.15 | .94 | .03 (.000-.049)*p* = .959 | .84 | α = .72 |
| **Form–Meaning Recognition model** | 176.010 | 157 | .142 | 1.12 | .96 | .03 (.000-.046)*p* = .981 | .89 | α = .86 |
| **Derivative Recall model** | 224.198 | 167 | .002 | 1.34 | .95 | .05 (.028-.060)*p* = .697 | .84 | α = .85 |
| **Derivative Recognition model** | 219.997 | 164 | .002 | 1.34 | .95 | .05 (.028-.060)*p* = .700 | .85 | α = .84 |
| **Multiple-Meanings Recall model** | 210.372 | 163 | .007 | 1.29 | .95 | .04 (.023-.057)*p* = .808 | .85 | α = .86 |
| **Multiple-Meanings Recognition model** | 209.492 | 168 | .016 | 1.25 | .95 | .04 (.017-.054)*p* = .885 | .82 | α = .85 |
| **Collocation****Recall model** | 168.823 | 160 | .301 | 1.06 | .96 | .02 (.000-.040)*p* = .996 | .84 | α = .81 |
| **Collocation Recognition model** | 166.834 | 160 | .430 | 1.04 | .98 | .01 (.000-.037)*p* = .998 | .83 | α = .76 |

After individually checking the validity of the different word-knowledge aspects, I validated the latent variables as displayed in the hypothesised model (Model 1). In this case, the recall and recognition aspects of each word-knowledge component were specified as the total test score of the 20 target words used to measure them. The recognition and recall composite scores were conceptualized as individual indicators of each word-knowledge component (see Model 1 in the main text). Unidimensionality (factor loadings >.40) and convergent validity (represented by the average variance extracted (AVE >.50)), was attained by every factor: form–meaning AVE = .52, derivatives AVE = .74, multiple-meanings AVE = .71, collocation AVE = .69. This suggests that the composite recall and recognition indicators of each word-knowledge factor belong to their respective constructs. Therefore, the second-order hypothesised model (Model 1) illustrated in the main text was tested. Table 3 shows the path coefficients with their significance and *z* scores for Model 1 in the main text.

**Table 3** Factor loadings, squared multiple correlations, *Z* scores, significance level and residual variance in Model 1 (main text).

|  |  |  |  |
| --- | --- | --- | --- |
| **Paths** | **Factor loadings R** | **R2** | ***Z*** |
| Form–Meaning 🡪 Form–Meaning Recog | .74 | .55 | .749 (*p* =.454) |
| Form–Meaning 🡪 Form–Meaning Recall | .69 | .48 | .752 (*p* =.452) |
| Derivatives 🡪 Derivatives Recog | .82 | .67 | 5.543\*\*\* |
| Derivatives 🡪 Derivatives Recall | .90 | .81 | 5.614\*\*\* |
| Multiple-Meanings 🡪 Multiple-Meanings Recog | .80 | .64 | .312 (*p* =.755) |
| Multiple-Meanings 🡪 Multiple-Meanings Recall | .88 | .77 | .310 (*p* =.756) |
| Collocation 🡪 Collocation Recall | .84 | .71 | 8.419\*\*\* |
| Collocation 🡪 Collocation Recog | .82 | .67 | 9.294\*\*\* |
|  | **Standardised regression weights** | ***Z*** |
| Vocabulary knowledge 🡪 Form–Meaning | .98.931.00.82 | .751 (*p* =.453) |
| Vocabulary knowledge 🡪 Derivatives | 4.895\*\*\* |
| Vocabulary knowledge 🡪 Multiple-Meanings | .309 (*p* =.757) |
| Vocabulary knowledge 🡪 Collocations | 6.877\*\*\* |
|  | **Standardised residual variances** | ***Z*** |
| Form-Meaning Recall (e1) | .52 | 8.150\*\*\* |
| Form-Meaning Recognition (e2) | .45 | 6.518\*\*\* |
| Derivative Recall (e3) | .19 | 4.274\*\*\* |
| Derivative Recognition (e4) | .33 | 6.753\*\*\* |
| Multiple-Meanings Recall (e8) | .22 | 4.294\*\*\* |
| Multiple-Meanings Recognition (e7) | .36 | 8.217\*\*\* |
| Collocation Recall (e6) | .29 | 4.571\*\*\* |
| Collocation Recognition (e5) | .33 | 5.320\*\*\* |
| Form-Meaning (e9) | .04 | .543(*p*=.59) |
| Derivatives (e10) | .14 | 2.526\* |
| Multiple-Meanings (e12) | .01 | .220(*p*=.83) |
| Collocation (e11) | .33 | 4.486\*\*\* |

\* *p* < .05, \*\* *p* < .01, \*\*\* *p* <.001

The model fit shows that the goodness-of-fit statistics reached the threshold to consider Model 1 valid. The overall model exhibited very good construct reliability (Composite Reliability = .94). All factor loadings were above .40, indicating that unidimensionality was achieved. Finally, convergent validity was also attained, as represented by the AVE of the overall vocabulary knowledge construct (AVE = .87) and the four sub-constructs (form–meaning AVE = .52, derivatives AVE = .74, multiple-meanings AVE = .71, collocation AVE = .69).

Despite the fit indices for Model 1 being acceptable, as explained in the main text, the high regressions between vocabulary knowledge and the four word-knowledge components (*β* = .82-1.00) suggest lack of discriminant validity between the word-knowledge components. This indicates that the components behave as inseparable from each other, and thus should be better seen as belonging to a single construct. Therefore, the unidimensional model of vocabulary knowledge was tested (see Table 4 below).

**Table 4** Factor loadings, squared multiple correlations, *Z* scores, significance level and residual variance in Model 2 (main text).

|  |  |  |  |
| --- | --- | --- | --- |
| **Paths** | **Factor loadings*****R*** | ***R2*** |  ***Z*** |
| Vocabulary knowledge 🡪 FM Recall | .68 | .46 | 9.420\*\*\* |
| Vocabulary knowledge 🡪 FM Recog | .73 | .53 | 10.518\*\*\* |
| Vocabulary knowledge 🡪 Deriv Recall | .84 | .71 | 12.476\*\*\* |
| Vocabulary knowledge 🡪 Deriv Recog | .76 | .58 | 10.891\*\*\* |
| Vocabulary knowledge 🡪 MM Recall | .88 | .77 | 13.902\*\*\* |
| Vocabulary knowledge 🡪 MM Recog | .80 | .64 | 11.142\*\*\* |
| Vocabulary knowledge 🡪 Collo Recall | .69 | .48 | 10.488\*\*\* |
| Vocabulary knowledge 🡪 Collo Recog | .67 | .45 | 9.469\*\*\* |
|  | **Error correlations** |  |  ***Z*** |
| FM Recall 🡨🡪 FM Recog | .05 |  | .376 *p*=.71 |
| Derivatives Recall 🡨🡪 Derivatives Recog | .28 |  | 4.564\*\* |
| MM Recall 🡨🡪 MM Recog | .02 |  | .156 *p*=.88 |
| Collocation Recall 🡨🡪 Collocation Recog | .42 |  | 4.962\*\*\* |
|  | **Standardised residual variances** |  ***Z*** |
| Form-Meaning Recall (e1) | .54 | 6.953\*\*\* |
| Form-Meaning Recognition (e2) | .47 | 7.215\*\*\* |
| Derivative Recall (e3) | .30 | 6.494\*\*\* |
| Derivative Recognition (e4) | .42 | 7.239\*\*\* |
| Multiple-Meanings Recall (e5) | .23 | 4.680\*\*\* |
| Multiple-Meanings Recognition (e6) | .37 | 8.355\*\*\* |
| Collocation Recall (e7) | .53 | 8.798\*\*\* |
| Collocation Recognition (e8) | .56 | 7.790\*\*\* |

\*\* *p* < .01, \*\*\* *p* < .001.

**Appendix 4 – Multigroup CFA and measurement invariance procedures**

In order to be able to confidently make comparisons regarding the behaviour of a model across groups it needs to be shown that the latent constructs function the same way across the target groups. This can be checked through multigroup measurement invariance tests. These multigroup analyses can inform of model invariance at different levels. In this study I am interested in evaluating: 1) whether the word-knowledge aspects that comprise vocabulary knowledge operate equivalently across the Spanish and Chinese learners, and 2) whether the theoretical construct of vocabulary knowledge functions similarly across both groups. If these analyses show evidence of measurement invariance in the two groups, the validity for the unidimensional conceptualisation of the vocabulary-knowledge construct across EFL learners of different L1s would be further corroborated.

The application of multigroup CFA invariance involves the comparison of each time more restrictive hierarchical models in a series of steps. The first step is testing for configural invariance, which, if confirmed, allows to test for weak (factor loadings) invariance. If weak invariance is achieved, then strong (intercept) invariance is checked, which, if satisfied, will allow us to continue testing for strict (residual variances) invariance. If all these levels of invariance between groups are fulfilled it can be claimed that the construct being measured has the same meaning and behaves equivalently in the different language groups.

In order to evaluate the various increasingly-restrictive levels of invariance of the unidimensional model across groups I employed the modelling approach (Beaujean, 2014). This approach uses two criteria to establish invariance across groups: an overall adequate fit to the data according to various goodness-of-fit indices, and a negligible change in the CFI value from a less restrictive model to a more restrictive one, derived from the difference between their CFIs (ΔCFI). Following previous research guidance (Byrne, 2016; Kenny, 2016; Putnick & Bornstein, 2016), if the ΔCFI value is smaller than or equal to (≤) -.01, this indicates that the null hypothesis of invariance should not be rejected (Cheung & Rensvold, 2002, p. 251), and is taken as evidence of invariance (equivalence) between the two groups. Conversely, when ΔCFI is larger than -.01, it indicates that the more restrictive models have a significantly worse fit than a less restrictive model, suggesting that some parameters behave differently across groups. This would require fitting new models which allow those parameters one-by-one to differ across groups, in order to control for these differences in the subsequent tests.

The sequence of increasingly-restrictive models fitted to test measurement invariance in this study and their model fit are summarised in Table 5. Following Beaujean (2014), multiple fit indices are reported to decide on the goodness-of-fit and invariance of the models.

First, the configural invariance of the unidimensional vocabulary knowledge model was assessed by fitting the base model for the two learner groups simultaneously. Table 5 shows that, consistent with the results of the single-group analysis, this multigroup CFA model is well fitting (see *Model A*), confirming configural invariance. This level of invariance indicates that the model has the same factor structure in the two groups when assessed concurrently, further demonstrating the validity of the unidimensional structure of vocabulary across the language groups (see Figure 6 in the Main Document for the parameter estimates for the whole sample concurrently, and Figure 1 below for the estimates by language group individually).

Second, weak invariance was assessed by comparing the goodness-of-fit of the configural model (*Model A*) to the fit of the weak invariance model (*Model B*), in which factor loadings were constrained equal across the two groups. The more restrictive weak invariance model shows a good overall fit, and the ΔCFI compared to the less restrictive configural model is lower than -.01, and thus at an appropriate level to establish invariance. This suggests that the factor loadings seem to behave equivalently in the two groups, and is evidence of full weak model invariance across the groups. This level of invariance indicates that the various word-knowledge aspects that comprise vocabulary define and measure the same factor in each group, and, thus, the vocabulary knowledge construct has the same meaning for the two learner groups. This does not mean that the mean score of factor loading parameters needs to be exactly the same across both groups, but that together, in relation to the vocabulary factor, the components perform similarly in both groups of learners. This finding was anticipated in the visual analysis, where the best and worst component contributors to vocabulary were the same in both groups.

Since weak invariance was confirmed, strong invariance was tested next. This strong invariance implies that both the meaning of the vocabulary construct (factor loadings) and the estimated levels of the word-knowledge aspects (intercepts) are equal in both groups. Table 5 shows that the strong invariance model (*Model C*), in which all factor loadings and intercepts were constrained equal, does not have an adequate overall goodness-of-fit, with most fit indices falling outside of the acceptable values. The ΔCFI of this model is also significantly worse than that of the weak invariance model (ΔCFI = -.044), suggesting that some intercepts behave differently across groups. Follow-up analyses identified that the derivative recognition and the multiple-meaning recognition intercepts performed differently, and thus both were freed to differ across groups in *Model C2*. This *Model C2* with two free intercepts had a good fit, and the ΔCFI was not significantly different (i.e., ΔCFI was ≤ -.01) when compared to the weak invariance model (*Model B*). Thus, partial strong invariance was achieved across both groups, indicating that students from these two L1s at the same level of overall vocabulary knowledge would be expected to respond in the same way, and thus have a similar expected score, on most of the word-knowledge components regardless of their L1 background. The analysis of the intercepts for the vocabulary knowledge factor also shows that there is no major difference between the Spanish and Chinese students overall regarding their behaviour on vocabulary knowledge (*z*=-0.66, *p* =0.51).

Given that general measurement invariance can exist even with partial strong invariance (Steinmetz et al., 2009), the final step for establishing measurement invariance of the unidimensional vocabulary model was to check for strict invariance. To do this, *Model D*, in which the residual variances for the indicators were added as equality constraints across groups, was compared to the partial strong invariance model (*Model C2*). The results shows that, while the overall fit of *Model D* is generally adequate, the ΔCFI of the comparison between *Model D* and *Model C2* (ΔCFI = -.015) suggests that the more restrictive model has a slightly worse fit than the less restrictive model. Follow-up analyses indicated that only the residual variance for the multiple-meaning recall test seemed to behave differently, and thus this residual variance was allowed to differ across groups in *Model D1*. The strict invariance model with this residual variance freed (*Model D1*) yielded a good overall fit, and an appropriate ΔCFI value (≤ -.01). The improvement in fit of *Model D1* compared to the partial strong invariance model, while marginal, suggests that partial strict measurement invariance of the vocabulary knowledge construct and its components can be confirmed for both the Chinese and the Spanish learner populations. This partial strict measurement invariance implies that all except one of the word-knowledge aspects seem to measure the vocabulary knowledge dimension similarly and with a comparable degree of reliability in each language group (Deshon, 2004; Kline, 2016).

Together, the results of the multigroup CFA invariance analysis reveal that there is general measurement invariance of the unidimensional vocabulary knowledge model across Chinese and Spanish EFL learners, indicating that the word-knowledge aspects as specified in this study assess the same construct, mean the same and behave similarly across both L1 groups. Thus, we can conclude that there is evidence for the word-knowledge components and the vocabulary-knowledge construct to function equivalently in both language populations. This finding provides further evidence for the construct validity and representation of vocabulary as a unidimensional model.

**Figure 1** Unidimensional model estimates by learner group

|  |  |
| --- | --- |
| L1-Spanish Group (*n* = 144) | Diagram  Description automatically generatedL1-Chinese Group (*n* = 170) |

**Table 5** Summary of tests for multigroup CFA measurement invariance.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model description** | **χ2(df)** | **χ2/*df*** | **CFI** | **RMSEA****(90% CI)** | **SRMR** | **Model comparison** | **ΔCFI** | **ΔRMSEA** | **ΔSRMR** | **Decision** |
| **1. Configural invariance**:*Model A*: No equality constraints imposed | 41.865 (32) | 1.31 | .995 | .04 (.000-.078)*p* = .57 | .016 | \_\_\_\_\_ | \_\_\_\_\_ | \_\_\_\_\_ | \_\_\_\_\_ | \_\_\_\_\_ |
| **2. Weak invariance:***Model B*: equal factor loadings | 66.050 (39) | 1.70 | .987 | .07 (.038-.095)*p* = .15 | .065 | B vs. A | -.008 | -.03 | -.049 | Accept |
| **3. Strong invariance:** *Model C*: equal factor loadings & intercepts  | 168.607 (46) | 3.66 | .943 | .13 (.110-.151)*p* = .00 | .081 | C vs. B | **-.044** | -.06 | -.016 | **Reject** |
| *Model C1*: Model C with Deriv Recog intercept freed across groups | 121.577 (45) | 2.70 | .964 | .10 (.084-.129)*p* = .00 | .073 | C1 vs. B | **-.023** | -.03 | -.008 | **Reject** |
| *Model C2*: Model C with Deriv Recog + MM Recog intercepts freed across groups | 89.068 (44) | 2.02 | .979 | .08 (.057-.107)*p* = .02 | .069 | C2 vs. B | -.008 | -.01 | -.004 | Accept |
| **4. Strict invariance** *Model D*: equal factor loadings, intercepts & error variances  | 124.414 (52) | 2.39 | .966 | .09 (.073-.115)*p* = .00 | .085 | D vs. C2 | **-.013** | -.01 | -.016 | **Reject** |
| *Model D1*: Model D with MM Recall error variance freed across groups | 112.432 (51) | 2.20 | .971 | .08 (.061-.108)*p* = .01 | .084 | D1 vs. C2 | -.008 | .00 | -.015 | Accept |
| *Note*: *N* = 314; Chinese group *n* = 170; Spanish group *n* = 144. Estimation method Maximum Likelihood Robust, missing values method Maximum Likelihood, in lavaan 0.6-5. Δ = difference in fit indices values between models. **Bold numbers** = ΔCFI higher than -.01, and thus null hypothesis and model rejected. |

**References**

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Deshon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137–149.

Kenny, D. A. (2016). *Multiple Latent Variable Models: Confirmatory Factor Analysis*. http://davidakenny.net/cm/mfactor.htm

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 519–530

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), 599–616.