**What to Do When Scalar Invariance Fails: The Extended Alignment
Method for Multi-Group Factor Analysis Comparison of Latent Means Across Many Groups**

Herbert W. Marsh, Australian Catholic University, University of Oxford, United Kingdom
Jiesi Guo, Australian Catholic University,
Philip Parker, Australian Catholic University,
Benjamin Nagengast, University of Tübingen, Germany
Tihomir Asparouhov & Bengt Muthén, Muthén & Muthén
Theresa Dicke, Australian Catholic University

10 August, 2016

Herbert W. Marsh, Jiesi Guo, and Philip Parker, Institute of Positive Psychology and Education,
        Australian Catholic University,
Benjamin Nagengast, Hector Research Institute of Education Sciences and Psychology, University of
        Tübingen, Germany;
Tihomir Asparouhov & Bengt Muthén, Muthén & Muthén, Los Angeles, USA
Theresa Dicke, Institute of Positive Psychology and Education, Australian Catholic University
What to Do When Scalar Invariance Fails: The Extended Alignment

Method for Multi-Group Factor Analysis Comparison of Latent Means Across Many Groups

Abstract

Scalar invariance is an unachievable ideal that in practice can only be approximated; often using potentially questionable approaches such as partial invariance based on a stepwise selection of parameter estimates with large modification indices. Study 1 demonstrates an extension of the power and flexibility of the alignment approach for comparing latent factor means in a large-scale studies (30 OECD countries, 8 factors, 44 items and $N = 249,840$), for which scalar invariance is typically not supported in the traditional confirmatory factor analysis approach to measurement invariance (CFA-MI). Importantly, we introduce an alignment-within-CFA (AwC) approach, transforming alignment from a largely exploratory tool into a confirmatory tool, and enabling analyses that previously have not been possible with alignment (testing the invariance of uniquenesses and factor variances/covariances; multiple-group MIMIC models; contrasts on latent means) and structural equation models more generally. Specifically, it also allowed a comparison of gender difference in a 30-country MIMIC AwC (i.e., a SEM with gender as a covariate) and a 60-group AwC CFA (i.e., 30 countries x 2 genders) analysis. Study 2, a simulation study following up issues raised in Study 1, showed that latent means were more accurately estimated with alignment than with the scalar CFA-MI, and particularly with partial invariance scalar models based on the heavily criticized stepwise selection strategy. In summary, alignment augmented by AwC provides applied researchers from diverse disciplines considerable flexibility to address substantively important issues when the traditional CFA-MI scalar model does not fit the data.

Keywords: measurement invariance; alignment method; stepwise selection strategies, modification indices; MIMIC models; PISA

What to Do When Scalar Invariance Fails: The Extended Alignment Method for

Multi-Group Factor Analysis Comparison of Latent Means Across Many Groups

We begin with the premise that the model of complete scalar invariance based on the confirmatory factor analysis approach to measurement invariance (CFA-MI) is an unachievable ideal that in practice can only be approximated. Furthermore, in relation to current standards of acceptable fit, even acceptable approximations to the complete scalar CFA-MI are rare in large-scale studies based on many factors, items/factors, and groups. Nevertheless, consistently with typical practice, post-hoc adjustments to the a priori scalar CFA-MI model using a traditional stepwise selection strategy based on modification indices will eventually achieve an apparently acceptable fit for a partial scalar CFA-MI (CFA-MI$_{Part}$) model if sufficient adjustments are introduced. However, consistently with severe criticisms of such stepwise procedures in the statistical literature, and related caveats by Byrne, Shavelson, and Muthén (1989; also see Reise, Widaman, & Pugh, 1993) when they first introduced partial invariance, we agree with Asparouhov and Muthén's (2014) supposition that the traditional partial invariance approach to invariance is unlikely to lead the simplest, most interpretable model for large-scale studies, leading them to introduce the CFA-MI$_{AL}$ model.

Based on a large real data demonstration, followed by a simulation study, we extend the usefulness of a new, evolving statistical procedure: multiple group factor analysis alignment (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2013)—hereafter referred to as the CFA-MI$_{AL}$ model. For multiple group data, particularly when the number of groups is large, alignment can be used to compare latent factor means even when there is not support for complete scalar invariance. However, the starting point for alignment is still the typical set of CFA-MI tests of the multigroup invariance of factor loadings, intercepts, and latent means (e.g., Marsh, Muthén, et al., 2009; Meredith, 1993; Millsap, 2011; Vandenberg & Lance, 2000; Widaman & Reise, 1997). Indeed, if there is good support for complete scalar invariance, there is no need to pursue alignment. However, in large-scale studies, scalar CFA-MI models are almost always rejected. In applied research, it is typical to not even test for violations of the underlying invariance assumptions, to simply ignore them, or to explore a potentially large number of alternative, partial invariance models, in which some invariance restrictions are relaxed.

Particularly in large-scale studies, the stepwise selection process of relaxing invariance constraints one parameter at a time is highly cumbersome, idiosyncratic, and likely to capitalize on chance, so that the final solution is not replicable. The alignment (CFA-MI$_{AL}$) model is an easily applied, viable alternative to traditional CFA-MI$_{Part}$ models; it is "based on the configural model and essentially automates and greatly simplifies measurement invariance analysis. The method also provides a detailed account of parameter invariance for every model parameter in every group" (Asparouhov & Muthén, 2014, p. 1). Despite the great promise of CFA-MI$_{AL}$ to address practical problems associated with multiple group tests of invariance, there are important limitations in the currently available version of the CFA-MI$_{AL}$ model that substantially limit its usefulness in applied research and leave a number of unanswered questions about its appropriateness under different situations. Thus, alignment can only be used to test a limited number of CFA models, and cannot incorporate cross-loadings, covariates, or tests of structural equation models (SEMs) more generally. For these reasons it was initially seen primarily as an exploratory tool useful in preliminary analyses.

In the present investigation we introduce what we refer to as Alignment-within-CFA (AwC), which transforms the CFA-MI$_{AL}$ model from an exploratory tool into a confirmatory tool, allows researchers to pursue nearly all issues that can be addressed with traditional CFA and SEM models, and greatly enhances the usefulness of the CFA-MI$_{AL}$ model for applied research. The AwC solution is equivalent to the CFA-MI$_{AL}$ model solution in that it has the same degrees of freedom, same goodness of fit, and same parameter estimates as the CFA-MI$_{AL}$ model. Indeed, the AwC model in its basic form is the same as the alignment model, but reconfigured as a more general CFA model. In this respect, support or lack of support for the alignment model applies to the basic AwC model as well.

With AwC, applied researchers have more flexibility in terms of constraining or further modifying the basic AwC model (as it is a true CFA model) than they would with the CFA-MI$_{AL}$ model upon which it is based (also see Appendices A and B in the online Supplemental Materials). More generally, with the AwC extension of CFA-MI$_{AL}$, it is possible to test SEMs that are more general than CFA measurement models, which are the focus of the CFA-MI$_{AL}$ model. Thus, AwC provides a useful complement to the CFA-MI$_{AL}$ model, overcoming what were thought to be inherent limitations of its usefulness. Here we outline traditional CFA approaches to testing measurement

invariance, describe the CFA-MI$_{AL}$ model, introduce the AwC extension of the CFA-MI$_{AL}$ model, and briefly review the inherent limitations of the traditional stepwise approach to scalar CFA-MI$_{Part}$ models. In Study 1 we then provide an application of the AwC to substantively important issues in testing the measurement invariance of the Programme for International Student Assessment (PISA) motivation and engagement constructs over 30 OECD countries (Nagengast & Marsh, 2013). Finally, in Study 2 we present a simulation study to address questions about alignment raised by Study 1: in particular, a comparison of alignment and CFA-MI$_{Part}$ models in relation to bias in the estimation of latent means, which is a primary focus of scale CFA-MI models.

Multiple Group CFAs and Tests of Measurement Invariance

The Importance of Measurement Invariance

Comparisons of results across multiple groups (e.g., multiple countries) require strong assumptions about the invariance of the factor structure across the groups. Unless the underlying factors really do reflect the same construct, and the measurements themselves are operating in the same way (across groups, over age and time, or across different levels of continuous variables), mean differences and other comparisons might be invalid. If the underlying factors are fundamentally different, then there is no basis for interpreting observed differences (the "apples and oranges" problem). Important issues for applied researchers are the implications of the inevitable failures of these tests of invariance—in relation to the development of measurement instruments and the interpretation of results based on even well-established measures. However, these issues are frequently ignored in applied research. For example, in cross-national studies of motivational differences such as those considered here, interpretations of mean differences—or even relations among different constructs—presuppose that the factors are the same across countries. However, in their review of 48 cross-cultural research studies published in the *Journal of Personality* and *Social Psychology* between 1985 and 2005, Chen (2008; also see Nagengast & Marsh, 2013) reported that less than 17% tested measurement invariance, even though many of the published findings suggested violations of measurement invariance.

General Multigroup CFA Model

In its most general form, the CFA-MI model for p indicators, m latent variables and g groups is defined by the following equations (Sörbom, 1974; also see Nagengast & Marsh, 2013):

$$x^{(g)} = \tau_x^{(g)} + \Lambda_x^{(g)} \xi^{(g)} + \varepsilon^{(g)} \tag{1}$$

$$E\left(\xi^{(g)}\right) = \nu^{(g)} \tag{2}$$

$$Var\left(\xi^{(g)}\right) = \Phi^{(g)} \tag{3}$$

$$Var\left(\varepsilon^{(g)}\right) = \Theta_\varepsilon^{(g)} \tag{4}$$

In the CFA-MI model, the p-dimensional group-specific response vectors $x^{(g)}$ that are typically of high-dimensionality (indicators of the latent factors) for each of g groups are explained by an underlying set of latent variables $\xi^{(g)}$ of lower dimensionality: an m-dimensional vector. The $p \times m$ —dimensional factor loading matrix $\Lambda_x^{(g)}$ specifies the relations of the latent variables and the indicators. The p-dimensional vector $\tau_x^{(g)}$ contains the group-specific intercepts, one for each indicator, and the p-dimensional vector $\varepsilon^{(g)}$ contains the residuals with a $p \times p$-dimensional variance-covariance matrix $\Theta_\varepsilon^{(g)}$ that is typically assumed to be a diagonal matrix, implying that residuals associated with different indicators are uncorrelated. The m-dimensional mean vector of the latent variables is given by $\nu^{(g)}$, the $m \times m$-dimensional variance-covariance matrix of relations among the multiple latent factors by $\Phi^{(g)}$. Both the latent variables $\xi^{(g)}$ and the residuals $\varepsilon^{(g)}$ are assumed to be normally distributed. The superscripts (g) indicate that the corresponding vectors and matrices can vary across the multiple groups. The model implies group-specific $p \times p$-dimensional variance-covariance matrices $\Sigma_{xx}^{(g)}$ and p-dimensional mean vectors $\mu_x^{(g)}$ for the observed variables

$$\mu_x^{(g)} = \tau_x^{(g)} + \Lambda_x^{(g)} \nu^{(g)}, \tag{5}$$

$$\Sigma_{xx}^{(g)} = \Lambda_x^{(g)} \Phi^{(g)} \Lambda_x^{(g)} + \Theta_\varepsilon^{(g)}. \tag{6}$$

Thus, individual responses (y_{ipg}) are defined as:

$$y_{ipg} = \nu_{pg} + \lambda_{pg}\,\eta_{ig} + \varepsilon_{ipg} \tag{7}$$

where $p = 1, \ldots, P$ and P is the number of observed indicator variables, $g = 1, \ldots, G$ and G is the number of groups, $i = 1, \ldots, N_g$ where N_g is the number of independent observations in group g, and

η_{ig} is a latent variable. The discrepancy between the model implied and the observed mean vectors and covariance matrices constitutes the basis for global tests of model fit.

Traditional Multigroup CFA Tests of Measurement Invariance

Typically, CFA-MI tests (see Marsh, Muthén et al., 2009; Meredith, 1993; Vandenberg & Lance, 2000; Widaman & Reise, 1997) begin with a *configural invariance model* in which the factor loading matrices $\mathbf{\Lambda}_y^{(g)}$ are restricted to have the same pattern of fixed and freed elements across the groups. In this model, none of the estimated parameters are constrained to be invariant over groups (except for those constrained to fixed values—typically 0 or 1—used to identify the factor structure in each group). If this model does not fit the data, there are fundamental differences in the dimensionality of assessed constructs across the multiple groups, and cross-country comparisons on common scales are fraught with difficulty (see Marsh & Grayson's 1994 discussion of a hierarchy of invariances and of what interpretations might be justified without at least partial configural invariance). The configural invariance model also serves as a reference model against which to compare the fit of more restrictive invariance models that impose further constraints, setting parameters to be invariant across the multiple groups.

The second CFA-MI model is usually the *metric invariance model* (Vandenberg & Lance, 2000, or the *weak measurement invariance model*, Meredith, 1993). In this model, the factor loading matrices are set to be invariant across the multiple groups (i.e. $\mathbf{\Lambda}_y^{(g)} = \mathbf{\Lambda}_y$). When metric invariance holds, the indicators are related to the latent variables in the same way in all groups. Differences in the latent variables get translated into differences in the indicators in a similar way across the groups. Metric invariance is the precondition for meaningful comparisons of the variance-covariance matrices of the latent variables $\mathbf{\Phi}^{(g)}$ across the groups, as they are defined by similar measurement models (Marsh, Hau, Artelt, Baumert, & Peschar, 2006; Meredith, 1993; Widaman & Reise, 1997).

After establishing metric invariance, there is no universal agreement on what restrictions to test next (Nagengast & Marsh, 2013). Marsh, Muthén, et al. (2009) presented a 13-model taxonomy of measurement invariance that systematically incorporates many combinations of invariance tests. The configural invariance model and the metric invariance model are included as the first two models

in this taxonomy. All further models are built on the metric invariance model, and further restrict parameters to be invariant across multiple groups. However, for the present purposes, and in many other applications, the main focus is on the *scalar invariance model* (Vandenberg & Lance, 2000; also referred to as *strong measurement invariance;* Meredith, 1993), which is usually included in tests of measurement invariance. In this model, the item intercepts are set to be invariant across the multiple groups (i.e. $\tau_x^{(g)} = \tau_x$). Scalar invariance is a precondition for comparing latent factor means across the multiple groups (Marsh, Muthén et al., 2009; Meredith, 1993; Widaman & Reise, 1997). If this restriction holds, there are no systematic differences in the average item responses between groups that are not due to differences in the mean level of latent variables. Although this is not always a focus of measurement invariance studies, further tests might include uniquenesses, factor variances, factor covariances, path coefficients, latent means, or various combinations of these different sets of parameters (e.g., Marsh, Muthén et al., 2009).

Criticisms of the Traditional Approach to Partial Scalar CFA-MI Solutions

All statistical models are false. An overarching premise of the present investigation is the now widely accepted truism all that statistical models—including CFA and SEMs—only reflect approximations to reality that are always wrong (e.g., MacCallum, 2003; Marsh, Lüdtke, et al., 2013; McDonald, 2010; but also see Box, 1979; Thurstone, 1930; Tukey, 1961). As emphasized by MacCallum (2003, p. 114) in his presidential address:

> Regardless of their form or function, or the area in which they are used, it is safe to say that these models all have one thing in common: They are all wrong. Simply put, our models are implausible if taken as exact or literal representations of real world phenomena.

From this perspective, it is essential for applied researchers to evaluate how model misspecification influences their interpretations and conclusions. As applied here, the complete scalar CFA-MI model, based on the assumption that a large number of parameters have exactly the same values in a large number of groups, is highly implausible if based on real data. Indeed, in the same way that from a philosophical perspective all statistical models are wrong, even the assumption that any two parameters are exactly the same is always wrong, and will be shown to be false from a statistical perspective when based on a sufficiently large *N*. From this statistical perspective the critical question

becomes whether the approximation provided by the complete scalar CFA-MI is acceptable and, if not, whether an appropriate approximation to this model can be found.

Large-scale application of CFA-MI models. Classic demonstrations of support for complete scalar CFA-MI models typically are based on small-scale studies in which the numbers of factors, groups, and indicators are all small (e.g., Byrne et al., 1989; Reise, Widaman, & Pugh, 1993). In contrast, in large-scale studies like the cross-national PISA research with many groups, factors, items/factor and participants like the Nagengast and Marsh (20130 study), which was the starting point of the present investigation, an acceptable fit of the complete scalar CFA-MI model is rarely achieved (e.g., Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; He & Kubacka, 2015; Rutkowski & Svetina, 2014; Zercher, Schmidt, Cieciuch, & Davidov, 2015).

Thus, Rutkowski and Svetina (2014) contended that most studies in support of measurement invariance were based on a few groups and relatively small sample sizes, and that there were relatively few published studies that even tested scalar invariance in large-scale applications with typical numbers of groups and sample sizes in cross-national surveys such as the Trends in International Mathematics Study (TIMSS), PISA, the Teaching and Learning International Survey (TALIS), and in many surveys not focused on education outcomes—such as those administered by the World Health Organization and UNICEF. Rutkowski (semnet@listserv.ua.edu, 6 June, 2016) also chaired an expert group for the 2013 TALIS survey that conducted multiple group analyses on dozens of scales across some 32 countries, but found scalar invariance untenable in nearly all cases (also see He & Kubacka, 2015).

Similarly, in apparently the largest published study of scalar CFA-MI ever conducted, Zercher et al. (2015) evaluated the invariance of responses to a total of 90 groups (15 countries in each of 6 waves of the European Social Survey (ESS; Total N = 173,071) for a single 3-item scale designed to measure universalism. Demonstrating that the scalar CFA-MI model was unacceptable for analyses across the 15 countries in each wave considered separately, as well as for the analysis of 90 groups across the six waves, they then deleted groups that were not at least partially invariant, eliminating all but 37 of 90 groups. Had they considered the multiple factors from instruments from which this scale was selected, or included more than just three items, support for even partial

invariance would probably have been much worse. Noting the limitations of the scalar CFA-MI approach in large-scale studies, they recommended further research using recent developments in invariance testing, including the alignment approach considered here.

Problems with stepwise approaches to partial invariance. Byrne, Shavelson, and Muthén (1989) introduced and popularized the CFA-MI$_{Part}$ model, particularly in relation to testing differences in latent means. Based on a small-scale application (two groups, 11 indicators designed to measure four factors), they relied heavily on modification indices supplemented by substantive knowledge and intuition to make post-hoc corrections to achieve a scalar CFA-MI$_{Part}$ model. Emphasizing that post-hoc adjustments are problematic, rendering probability values meaningless, they lamented that applied researchers "must await the research efforts of statisticians in resolving this important psychometric obstacle" (p. 465). As an interim remedy they recommended cross-validation, but noted that "the relaxation of many parameters is likely to yield an unsuccessful cross-validation" (p. 465) and stressed the need for the use of sound judgment. However, more than a quarter of a century after this seminal work, the common practice is to produce scalar CFA-MI$_{Part}$ models based substantially on forward stepwise application of modification indices (Schmitt & Kuljanin, 2008), with some post-hoc justification for the reasonableness of adjustments that had not been hypothesized a priori. Thus, Asparouhov and Muthén (2014) proposed alignment as a potentially more useful alternative to the potentially dubious scalar CFA-MI$_{Part}$ model.

It is worthwhile briefly reviewing well-known problems with the application of forward stepwise selection procedures to achieve an acceptable fitting model. Although these issues are more widely recognized in relation to stepwise multiple regression, a similar logic applies to the use of modification indices to achieve an acceptable fit in scalar CFA-MI$_{Part}$ models. The starting point is a complete scalar invariance model that does not provide an acceptable fit to the data, and that typically is predicated on acceptable fit of the corresponding configural model. In traditional CFA-MI$_{Part}$ models the applied researcher selects the estimated parameter with the largest modification index and frees this parameter. This process is repeated until an acceptable fit is achieved and the freeing of additional parameters does not substantially improve the fit. In large-scale studies with many groups,

factors, and measured variables it is entirely possible that acceptable fit in relation to current standards of goodness of fit will require hundreds or even thousands of adjustments.

In scathing attacks on stepwise strategies, statistician Frank Harrell (2011), along with many others (e.g., Davison, 2003; Judd & McClelland, 1989; MacCallum, Roznowski & Necowitz, 1992) underlined the weaknesses of stepwise strategies, particularly forward stepwise strategies, as used in the typical CFA-MI$_{Part}$ model. Harrell (2001, p. 56) emphasizes that:

> Stepwise variable selection has been a very popular technique for many years, but if this procedure had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing.

Davison (2003, p. 400) notes that "These three procedures [forward selection, backward elimination, and stepwise regression] have been shown to fit complicated models to completely random data, and although widely used they have no theoretical basis". Similarly, Judd and McClelland (1989, p. 204) note that "An unfocused search through many possible models (sometimes referred to as a 'fishing expedition') increases the likelihood of capitalizing on chance and finding a model which represents only a spurious relationship".

More specifically, the typical forward stepwise selection procedure based on modification indices to achieve partial scalar CFA-MI in large-scale studies is dubious in that: goodness of fit and related indices are positively biased. For the selected parameters that are freed, modification indices (and extent of non-invariance) are positively biased. In contrast, for unselected parameters for which non-invariance is assumed to be zero, modification indices are negatively biased. Furthermore, because there is extreme multicollinearity in the modification indices, variable selection becomes arbitrary. CFA-MI$_{Part}$ models identified by stepwise methods have an inflated risk of capitalizing on chance features of the data, such that the final scalar CFA-MI$_{Part}$ that is the end result of this stepwise process is not optimal when cross-validated with new data, but the final model is rarely tested in this way. Indeed, the iterative stepwise selection process will sometimes find a local minimum and get stuck in a suboptimal region of model space, so that potentially better models are not even considered. Even among alternative stepwise procedures, the forward approach is generally not the preferred method, in that it results in suppression effects, such that adjustments are only significant after an

earlier adjustment has been made. Thus, in their review of model modifications in CFAs, MacCallum et al. (1992) demonstrated that stepwise procedures produced inconsistent and erratic cross-validation, bringing them to "a position of considerable skepticism with regard to the validity of the model modification process as it is often used in practice" (p. 502).

There are, of course, some issues that are idiosyncratic to the application of stepwise procedures for partial scalar invariance. In particular, predicated on the finding of a well-fitting configural model, adjustments are made primarily (or exclusively) in relation to factor loadings and intercepts, in order to achieve an acceptable goodness of fit for the scalar CFA-MI model compared to the corresponding configural model. However, goodness of fit provides a dubious basis for evaluating the model, as freeing enough parameters ultimately will achieve a fit that approaches that of the configural model, which has already been shown to provide an acceptable fit. Also, as emphasized earlier, the main purpose of the scalar invariant model (partial or complete) is to provide a justification for the evaluation of latent means and related statistical models, but there is no guarantee that the stepwise selection process, based on freeing factor loadings and intercepts to achieve an acceptable goodness of fit, will facilitate this objective of providing unbiased means.

Alignment Method to Test Measurement Invariance

In the typical test of scalar invariance, the intercepts v_{pg} and loading parameters λ_{pg} are held equal across groups; the factor mean and variance in one group are fixed to 0 and 1 respectively. As already emphasized, this fully invariant scalar model will frequently not provide an acceptable fit to the data, particularly when the numbers of items, latent factors, and groups are large, as is the case in the PISA 2006 data considered here. In contrast, the CFA-MI$_{AL}$ model (Asparouhov and Muthén, 2014) does not assume measurement invariance, but seeks an optimal measurement invariance pattern based on a simplicity function that is similar to the rotation criteria used with exploratory factor analysis. With this approach it is possible to estimate all of the parameters (v_{pg}, λ_{pg}, α_g, ψ_g), while holding non-invariance to a minimum.

The alignment approach begins with a traditional CFA-MI analysis and is predicated on the assumption that the fit of the configural model is acceptable and substantially better than the scalar model. Following these initial analyses, the first step of the CFA-MI$_{AL}$ model is a configural model

(M0) in which all factor means and variances are constrained across all groups to be 0 and 1 respectively, but all factor loadings and intercepts are freely estimated. The final alignment model (CFA-MI$_{AL}$) has the same fit as M0. Asparouhov and Muthén (2014) describe the relation between M0 and CFA-MI$_{AL}$ as parallel to unrotated and rotated EFA models in which the rotated model simplifies the interpretation without compromising model fit. This is accomplished by minimizing a total loss/simplicity function that accumulates measurement non-invariance across G groups with respect to α_g and ψ_g, based on the component loss function, which has been used in EFA (see, e.g., Jennrich, 2006). In this way, a non-identified model where factor means and factor variances are added to the configural model is made identified, by adding a simplicity requirement. This loss function is minimized when there are a few large non-invariant measurement parameters and many approximately invariant measurement parameters, rather than many medium-sized non-invariant measurement parameters. This is akin to EFA rotation functions, which aim for either large or small loadings, but not midsized loadings (see Asparouhov & Muthén, 2014 for more details).

Following the selection of the CFA-MI$_{AL}$ model, alignment offers a detailed analysis to determine which parameters are approximately invariant and which are not. For each measurement parameter, the largest set of groups is found such that the estimate for each group is not significantly different from the average parameter estimate across all groups in the invariant set using a multiple comparison process, with additional rules to ensure that the process stabilizes. The relative contribution of each parameter to the simplicity/loss function provides an indication of the degree of non-invariance (differential item functioning) associated with parameter estimates that can be useful in the refinement of items in future applications.

Asparouhov and Muthén (2014) demonstrated support for the CFA-MI$_{AL}$ model on the basis of a simulation and also on analysis of real data. In their simulation study, they varied the sample size (100 or 1,000 per group), number of groups (2, 3, 15 or 60), and extent of non-invariance (0%, 10%, 20%). Results showed that known population parameters were accurately estimated even when there was substantial non-invariance, particularly when sample sizes were large. Even in the worst case (substantial non-invariance, small Ns, large number of groups), biases tended to be small. In the real-data example from the European Social Survey (49,894 subjects in 26 European countries),

Asparouhov and Muthén (2014) tested the cross-country invariance on the basis of four items designed to measure a single factor in which the CFA scalar model showed a poor fit to the data. The alignment approach also showed considerable non-invariance for three of the four items, but relatively little non-invariance in the fourth item. Although the authors highlighted some differences between the CFA-MI$_{AL}$ and traditional scalar models in terms of latent means, the relative ranking of the 26 countries was very similar in respect of the traditional scalar CFA-MI and alignment models. In concluding remarks, Asparouhov and Muthén argued that alignment provides many advantages over the traditional CFA-MI approach to complete or partial scalar invariance; tests of some of these assertions are the focus of the present investigation.

Despite the great promise of the CFA-MI$_{AL}$ model to address practical problems associated with multiple group tests of measurement invariance, there are important limitations in the currently available version that substantially limit its usefulness in applied research (Asparouhov & Muthén, 2014; see earlier discussion of AwC) in relation to the full range of tests of measurement invariance in CFAs and SEMs more generally. On this basis, the early CFA-MI$_{AL}$ was characterized as primarily an exploratory tool. Largely overcoming these current limitations, here we introduce the Alignment-within-CFA (AwC) approach, which transforms alignment from an exploratory to a confirmatory tool, allowing the researcher to pursue nearly all issues that can be addressed with traditional CFA and SEM models, and greatly enhancing its usefulness for applied research.

The AwC approach is based on a similar logic to the exploratory structural equation model (ESEM) within CFA (EwC), which similarly transformed the usefulness of ESEM (Asparouhov & Muthén, 2009; Marsh, Muthén, et al., 2009; Marsh, Nagengast, & Morin, 2013; Marsh, Morin, Parker, & Kaur, 2014). In AwC the first step is to test a standard CFA-M$_{AL}$, as described by Asparouhov and Muthén (2014). However, the next step is to reconfigure this model as a standard CFA model, using as starting values the final CFA-MI$_{AL}$ estimates with appropriate fixed and free parameter estimates, such that the AwC solution is equivalent to the CFA-MI$_{AL}$ solution in terms of number of estimated parameters, goodness of fit, and definition of the factor structure (see subsequent discussion in Methods section and a detailed description in Appendix B of the online Supplemental Materials). Indeed, the AwC model in its basic form is the same as the alignment model, only being

reconfigured as a more general CFA model, so that support for the CFA-MI$_{AL}$ solution necessarily implies support for the AwC. However, the AwC provides the flexibility to test this solution within a broader range of CFA and SEMs as demonstrated in Study 1

Study 1: An Overview of the Substantive and Methodological Focus

The data considered here are based on the student background questionnaire of PISA 2006, which contains eight scales measuring a variety of motivational and engagement constructs in science (e.g. academic self-concept, self-efficacy, and value; see Supplemental Materials, Appendix 1 for further discussion). Here we apply the CFA-MI and CFA-MI$_{AL}$ models to evaluate the measurement properties of these scales for nationally representative samples of 15-year-old students from 30 OECD countries (total $N = 249,840$). Using these data, Nagengast and Marsh (2013) applied traditional CFA-MI models to demonstrate that the a priori scales showed a well-defined eight factor solution. There was reasonable support for the invariance of factor loadings across countries (metric invariance), but not for the invariance of item intercepts (scalar invariance), making mean comparisons across countries dubious. Hence, these data provide an ideal application of the CFA-MI$_{AL}$ model, which is specifically designed for such purposes. In this respect the substantive orientation of this investigation is to evaluate cross-cultural differences in latent means of science-related motivational constructs as well as relations between these motivational factors and important covariates: gender, science achievement, and socioeconomic status (SES).

In pursuing our methodological aims, we demonstrated the flexibility of AwC and its applicability to substantively important issues. We began by comparing the estimated factor means based on the new CFA-MI$_{AL}$ model and the traditional scalar invariance model, and then introduced AwC. Based on AwC, we extended alignment to include tests of the invariance of factor variance-covariance and item uniqueness. We then integrated the MG-CFA models and the multiple indicators multiple cause (MIMIC) models with AwC, to test the invariance of relations between the motivational constructs and the covariate variables, particularly for gender. In an alternative approach to the evaluation of gender differences in factor means, we conducted a 60-group (30 countries x 2 genders) AwC with a priori contrasts within each country, comparing these results with those based on the corresponding 30-group MIMIC analysis.

Methods

Data

Our data are nationally representative responses by 15-year-old students from all 30 OECD countries in PISA 2006 (*N* = 249,840). These raw data are readily available through the OECD-PISA website (https://www.oecd.org/pisa/pisaproducts/) as well as in the extensive documentation, manuals, and technical reports. The samples were collected using a complex two-stage sampling design and were, after using the appropriate survey weights, representative of the national population (OECD, 2009). Although academic achievements in reading, mathematics, and science were assessed in PISA 2006, only science-related motivation items were included in the questionnaire (see OECD, 2009). Overall, 44 motivation items were used to measure eight motivational constructs on a 4-point Likert scale, with 1 indicating that the participants "*strongly agree*" and 4 indicating "*strongly disagree*", with two exceptions: *science self-efficacy* (ranging from "*do easily*" to "*could not do it*" on a 4-point Likert scale) and *extracurricular activities* (ranging from "*very often*" to "*hardly ever*" on a 4-point Likert scale). For the present purposes, responses were reverse-scored, so that higher values represent more favorable responses and thus, higher levels of motivation.

Eight motivational constructs. The *Science self-concept* scale assessed students' self-perceptions of their ability in science (e.g., "I learn science topics quickly"). The *science self-efficacy* scale assesses students' confidence in performing real world science-related tasks (e.g., "Identify the science question associated with the disposal of garbage"). The *Enjoyment of science learning* scale assessed the enjoyment a student gains from performing a science-related activity (e.g., "I am interested in learning about science"). The *instrumental motivation* scale assesses how well science achievement relates to current and future goals (e.g., "I study science because I know it is useful for me"). The *future-oriented science motivation* scale assessed students' expectations about tertiary science studies and working in science-related careers ("I would like to work in a career involving science"). The scales that assessed students' perceptions of *general value of science* (e.g., "science is valuable to society"; henceforth referred to as "general value") and *personal values of science (*e.g., "Science is very relevant to me") were also included. Finally, *extracurricular activities in science* assessed the frequency of students engaging in out-of-school activities related to science (e.g.,

"Borrow or buy books on science topics"). Scale reliabilities for the eight motivational factors were acceptable (see Table 1).

Covariates. Gender (0 = male, 1 = female), SES (Economic, Social and Cultural Index [ESCS]; see OECD, 2009) and science achievement were treated as covariates in MIMIC models. To prevent biased population estimates, PISA measured science abilities using five plausible values for each subject (with a mean of 500 and a standard deviation of 100). Hence, to be able to correct the measurement error appropriately, these sets of plausible values were used to measure students' achievement (see OECD, 2009).

Statistical Analyses

All analyses were conducted with Mplus (Version 7.11; Muthén & Muthén, 1998–2015). A main focus in the present investigation is the application of AwC to MIMIC and MG-CFA models based on the robust maximum likelihood estimator (MLR), with standard errors and tests of fit that are robust in relation to non-normality and non-independence of observations. In addition, we applied corrected standard errors and model fit statistics to control for the nesting of students within schools, based on the TYPE = COMPLEX option in Mplus. The HOUWGT weighting variable was also taken into account in data analysis, in order to correct the computation of standard errors and tests of statistical significance (see Nagengast & Marsh, 2013 for more discussion). For the present purposes we used the FIXED option available in the Mplus CFA-MI$_{AL}$ model, in which the latent factor mean and variance of one arbitrarily selected group (in this case the first group, Australia) were fixed to 0 and 1 respectively (see Appendix B in the online Supplemental Materials for the Annotated Mplus syntax; also see Asparouhov & Muthén, 2014 for more discussion).

As discussed earlier, if the invariance of item intercepts (or even factor loadings) is not supported and the scalar model provides a poor model fit, an alignment analysis can be employed to evaluate latent mean comparisons. AwC can be applied when there is a need to conduct additional analysis that cannot be easily implemented within the alignment framework but that can be estimated with CFA and SEM models. All parameter estimates from the alignment solution should be used as starting values to estimate the AwC model. For purposes of identification, one item from each factor is arbitrarily selected (e.g., the first indicator) as a referent indicator, and the factor loading and

intercept of this indicator are fixed to the estimated values from the alignment solution (using starting

values supplied by the Mplus package). However, it is also possible to achieve identification using

other traditional approaches (e.g., fixing factor variances). The alignment solution (as well as the

AwC solution, which is equivalent to the alignment solution), has the same degrees of freedom, the

same chi-square and goodness of fit statistics as the configural MG-CFA model (see Supplemental

Materials, Appendix 2, for further discussion).

This process of constructing the AwC model from the M_{AL} solution is demonstrated in

Appendix 2 (Supplemental Materials). The output file for the M_{AL} solution contains the start values–

parameter estimates based on the final M_{AL} solution, which are then used to construct the AwC

syntax. For each latent factor loading, the first indicator of that factor and the corresponding indicator

intercept is fixed, and this process is repeated for each of the multiple groups. Output from the M_{AL}

and AwC demonstrates that all parameter estimates are the same for the M_{AL} and AwC solutions (this

is shown in Appendix 2 for one country, USA; this was also the case for all 30 countries).[1] However,

because the AwC is merely a CFA model, it is possible to conduct other CFA and SEM models that

cannot be tested with the CFA-MI$_{AL}$ model.

Missing data. In order to account for the five plausible values for each achievement score, all

data analyses involving achievement were run separately for each of the five plausible values. For

each of the five data sets, each based on different plausible values, we used full information maximum

likelihood (FIML) estimation (Enders, 2010) to handle missing data on the remaining items, given the

relatively small amount of missing data (mean coverage rates across the 44 items being .974). This

approach is similar to using FIML within each of the five data sets and treating achievement as an

auxiliary variable (Enders, 2010). Final parameter estimates, standard errors and goodness-of-fit

statistics were obtained with the automatic aggregation procedure implemented in Mplus for multiple

imputation, to properly handle plausible values (Rubin, 1987).

Goodness of fit. A number of traditional indices that are relatively independent of sample

size were utilized to assess model fit (Hu & Bentler, 1999; Marsh, Balla, & McDonald, 1988; Marsh,

Hau & Wen, 2004): the comparative fit index (CFI), the root-mean-square error of approximation

(RMSEA) and the Tucker-Lewis Index (TLI). Values greater than .95 and .90 for CFI and TLI

typically indicate excellent and acceptable levels of fit to the data. RMSEA values of less than .06 and .08 are considered to reflect good and acceptable levels of fit to the data. However, these cutoff values constitute only rough guidelines, rather than golden rules (Marsh, Hau & Wen, 2004). Typically it is more useful to compare the relative fit of different models in a nested taxonomy of measurement invariance models than to compare the relative fit of single models (Marsh, Muthén, et al., 2009). Cheung and Rensvold (2002) and Chen (2007) have suggested that if the decrease in CFI is not more than .01, and the RMSEA increases by less than .015 for the more parsimonious model, then invariance assumptions are tenable. Again, all these proposals should be considered as rough guidelines only, or rules of thumb.

Results

Factor Structure: Preliminary CFA

In preliminary analyses, we evaluated the factor structure and relations with covariates on the basis of the total group. The total group CFA model provided a good fit to the data (CFI = .963, TLI = .960, RMSEA = .012, see Model TG1CFA, Table 2) and the factor loadings of the eight scales range from .564 to .869 (see Table 1). We then added the three covariates (gender, SES, and achievement) to the total group CFA model (TG2CFA), which also provided a good fit to the data (CFI = .955, TLI = .950, RMSEA = .013).

The correlations among the 8 factors and three covariates (Table 3; see Appendix 4 in the Supplementary Materials for a more detailed summary) are of substantive interest, and serve as an advance organizer for subsequent analyses. Not surprisingly, all 28 correlations among the eight motivational constructs were positive (M r = .547, .370 to .785), and all were statistically significant, due in part to the large sample size. Boys had somewhat higher scores than girls (r = .024 to .094) for these science constructs. Science achievement was significantly positively correlated with all the motivational constructs (r = .081 to .372) except for extracurricular activities (r = .012, ns), whilst correlations with SES (r = -.030 to .114) were smaller.

Traditional CFA Test of Measurement Invariance of Factor Structure Over Countries

Next we conducted a series of increasingly stringent tests of measurement invariance across the 30 countries. The configural invariance model (MG1 in Table 2) fitted the data well (CFI = .952,

TLI = .948, RMSEA = .027) and served as a baseline model that was later used for comparison

purposes with more restrictive invariance models. We then tested metric invariance (Model MG2,

Table 2) by constraining the factor loadings to be invariant across the 30 countries. This more

parsimonious model resulted in a small decrease in fit indices compared to the configural model (CFI

= .006, TLI = .005, RMSEA = .001). In support of metric invariance, these differences were less than

the recommended cutoff values typically used to argue for the less-parsimonious model.

In Model MG3, we tested the scalar invariance model in which the 44 item intercepts, as well

as the factor loadings, were constrained to be invariant across countries. The fit of the scalar model

might be seen as minimally acceptable (e.g., CFI = .906, TLI = .906; RMSEA = .058) by some

standards, but compared to the metric invariance model (MG2), the decrease in fit indices (CFI = .040,

TLI =.037, RMSEA = .020) was substantially greater than the recommended cutoff values for MG3.

These results demonstrate a lack of support for scalar invariance.

When scalar invariance is rejected, alternative tests of partial invariance based on

modification indices are suggested (Byrne, Shavelson, & Muthén, 1989). However, there are many

large modification indices based on the MG3—thousands of which are statistically significant; for

intercepts alone, 201 in the range of 100 to 200, 159 in the range of 200 to 500, and 59 in the range of

500 to 2928. Hence, the process of freeing parameter estimates one at a time until an acceptable fit is

obtained would be very laborious. More importantly, as noted earlier, the stepwise approach to partial

invariance has been severely criticized on the grounds of being biased, capitalizing on chance, and not

resulting in an optimal model (e.g., Davison, 2003; Harrell, 2011; Judd & McClelland, 1989;

MacCallum, Roznowski, & Necowitz, 1992) leading Asparouhov & Muthén (2014) to suggest that

this approach is unlikely to result in the most useful model (see earlier discussion on stepwise

strategies). In summary, these results suggest a lack of support for the scalar measurement model;

such support is prerequisite for comparing the means of the latent motivational constructs across 30

countries (similar conclusions are reached in Nagengast & Marsh, 2013).

The MG-CFA Model With the Alignment Method

In pursuit of the comparison of latent means, we applied the CFA-MI$_{AL}$ model to evaluate a

MG-CFA model of the eight motivational constructs. Although alignment attempts to minimize the

amount of non-invariance, it does not compromise the model fit. Thus, the MG-CFA with alignment (MG4$_{AL}$) has the same fit as the configural model (MG1), indicating that the alignment model fits the data well. More importantly, alignment allows us to compare mean differences of latent factors, and greatly simplifies measurement invariance analyses.

A potentially important contribution of the CFA-MI$_{AL}$ model is to provide a detailed account of parameter invariance for every model parameter in every group. For example, inspection of Table 1 shows that, on average, there is more non-invariance associated with item intercepts than there is with factor loadings. This is of course consistent with the CFA-MI results, which showed that there was reasonable support for the invariance of factor loadings, but not item intercepts. However, even within the set of items designed to measure the same construct there were substantial differences. For example, in the General Value factor the item intercepts of the item "Advances in science usually improve people's living conditions" were invariant across 22 of 30 groups, whereas the intercept was only invariant across 10 groups for the item "Science is valuable to society". Such information is especially useful for developing or revising a scale for future research (see Table 4 for more details about the invariance status of item loadings and intercepts involving self-concept and general value).

Although it is useful, diagnostic information about the extent of violation of invariance based on the CFA-MI$_{AL}$ model is based on tests of statistical significance that are highly influenced by sample size. However, such values can easily be transformed into standardized differences in the metric of Cohen's D that provide a potentially more meaningful summary of practical significance. For example, we present the difference between the alignment and scalar models for each of the eight factors (Table 1)—the mean and standard deviation across items within each scale and the 30 countries. Although the mean differences are consistently small, the standard deviations these differences are larger in size, particularly for the intercepts, which previous results have shown to be more non-invariant. Similarly, we show differences between the alignment and scalar model in relation to Cohen's D for individual items in the self-concept and general value scales (Table 2). Alternatively, these values can be represented as box plots, which provide a more heuristic representation of the distribution of differences in relation to Cohen's D values (see the boxplots in Appendix 6, Supplemental Materials). Although traditional modification indices and expected change

parameters are not included in the alignment output at this time, these values can be easily obtained

from the equivalent AwC model (see Appendix 7, Supplemental Materials).

Latent Means Comparisons: Alignment vs. Scalar Methods

In an attempt to look more closely at latent mean differences based on the CFA-MI$_{AL}$ model

and the traditional scalar invariance method, we focus on two motivational constructs: self-concept,

and general value of science. Graphs of the latent means (Figure 1) for self-concept based on the

alignment model (MG4$_{AL}$) and the scalar invariance model (MG2) demonstrate that latent mean

differences are highly similar (i.e., factor means are close to the diagonal). Both methods show that

Mexico (MEX) has the highest level of self-concept and Japan (JPN) the lowest level. For general

value, the similarity in the pattern of means for the two approaches is somewhat lower than for self-

concept. For example, the scalar method indicates that Iceland (ISL) has a substantially different

mean from Greece (GRC), whereas the CFA-MI$_{AL}$ model indicates essentially no difference between

these two countries. In contrast, for general value the CFA-MI$_{AL}$ model indicates a substantial mean

difference between Norway (NOR) and Austria (AUT), whereas the factor means of these two

countries are similar for the scalar method. In summary, the pattern of factor means based on the

CFA-MI$_{AL}$ model was more closely related to those based on scalar invariance for self-concept than

for general value. This is also consistent with our findings that the self-concept scale fitted the data

better than the general value scale, when the two constructs were considered separately.

Tests of the Invariance of the Latent Factor Variance–Covariance Matrix

Subsequently, we tested invariance constraints on various combinations of uniquenesses,

factor variances, and factor covariances, using the AwC extension of the CFA-MI$_{AL}$ model. Although

there is no a priori rationale for the ordering of these models, they are all nested under the alignment

model (MG4$_{AL}$ in Table 2). In this respect, the results are informative about the nature of invariance,

but also demonstrate the usefulness of AwC.

Inspection of the fit indices suggests that constraining factor variances and covariances to be

equal across the 30 OECD countries is reasonable (e.g., MG4 vs MG7$_{AwC}$ in Table 2; ΔCFI = .005,

ΔTLI = .003, ΔRMSEA = .001), whereas constraints associated with the invariance of

uniquenesses are not acceptable (e.g., MG4 vs MG5$_{AwC}$ in Table 2; ΔCFI = .029, ΔTLI = .028, Δ

RMSEA = .006; also see MG8$_{AwC,}$ which constrains uniquenesses as well as factor variance and

covariances). Relations among the eight factors are essentially the same as those observed with the

total group Model (i.e., TG2CFA; see Table 2 and Appendix 3 in the Supplemental Materials) and so

are not considered further. The lack of support for uniqueness invariance suggests that comparison of

the manifest means of the constructs across countries is inappropriate. Although the implications of

these results are not critical for the evaluation of latent means, as in the present investigation, they do

dictate caution in the evaluation of manifest means, which is a focus of many PISA studies.

Relations to Achievement, Gender, and SES: Integration of Multiple-Group and MIMIC

Approaches

Here we used a multiple-group MIMIC (MG-MIMIC) model to evaluate country-to-country

variation in how the three covariates (achievement, gender, and SES) are related to each of the

motivational constructs. More specifically, the eight motivational constructs were regressed on each

of the three covariates, and we evaluated differences across the 30 OECD countries. However, to

make the presentation manageable, we focus on the effects of the MIMIC variables on self-concept

and general value, but note that the same approach was used for each of the eight motivation factors

(also see Table 2). Again, we note that this SEM analysis is one that could not be evaluated with the

standard alignment model, but is possible with the AwC extension introduced here.

The configural MIMIC with no invariance constraints provided a reasonable fit to the data

(MG-MIMIC1 in Table 2; CFI = .942, TLI = .937, RMSEA = .028). Constraining factor loadings to

be invariant over the 30 groups led to a small decrease in fit indices (MG-MIMIC1 vs. MG-MIMIC2

in Table 2; ΔCFI = .006, ΔTLI = .005, RMSEA = .001). However, the fit of the scalar model with

the invariance of item intercepts (CFI = .898, TLI = .895, RMSEA = .036) was unsatisfactory,

compared to Model MG-MIMIC2, in that the decrement in fit was substantial (ΔCFI = .038, ΔTLI

= .037). Hence, these results based on MIMIC models largely parallel those based on the

corresponding models without MIMIC variables, in which the scalar invariance model did not fit the

data. In models without MIMIC variables, this problem was circumvented by the use of the CFA-

MI_{AL} model. However, covariates and SEMs more generally cannot be accommodated by the CFA-MI_{AL} model; this limitation is overcome by the AwC extension of the CFA-MI_{AL} model.

As discussed earlier, the AwC solution is equivalent to the configural MG-CFA model in that it has the same degree of freedom, goodness of fit, and measurement parameter estimates. The MG-MIMIC model with AwC (MG-MIMIC4$_{AwC}$)[2] and the configural MG-MIMIC1 model provide a reasonable fit to the data (i.e., CFI = .942, TLI = .937, RMSEA = .028).

The pattern of path coefficients across countries is graphed in Figure 2. On average, achievement and SES positively predict self-concept and general value; achievement had stronger predictive effects than did SES. The pattern of path coefficients involving achievement varied substantially over countries for general value (.189 to .472, median = .285) and, in particular, self-concept (.033 to .496, median = .236), whereas the pattern involving SES was smaller and more consistent for both constructs (self-concept: .000 to .169, median = .051; general value: -.007 to .147, median = .086). Thus, students with high science ability and from higher SES backgrounds were more likely to have high self-concept and general value of science.

Alternative Approaches to Gender Differences: AwC Extensions of the Alignment Method

The MIMIC model provides a parsimonious summary of the effects of covariates and the motivation factors, but is based on scalar invariance assumptions that the factor loadings and intercepts of the PISA factors are invariant over gender. Although the assumption of invariant intercepts is testable in the MIMIC model, the assumption of invariant factor loadings is not. Here, the fit of MIMIC models does not differ substantially from that of the corresponding models without MIMIC variables. However, particularly if this were not the case, it would be useful to fit less parsimonious but potentially more appropriate models in which MIMIC variables are represented as multiple group variables. We build on an early example (Little 1997) of juxtaposing MIMIC and multiple-group approaches to evaluate gender differences, in four countries. We illustrate how this approach can be adapted and extended to alignment and AwC models (also see Marsh, Nagengast & Morin, 2013, who extended this approach and adapted it to ESEM).

Here, we are particularly interested in how the patterns of gender differences in the motivational constructs vary across countries. Because we already know that the scalar invariance

model does not fit the data, we evaluated gender differences with two alternative approaches, both based on the CFA-MI$_{AL}$ model using AwC—again focusing on self-concept and general value to make the presentation more manageable (but also see the pattern of gender differences for all eight motivation factors for the total sample in Table 2). The first approach is an extension of the traditional MIMIC model, to evaluate the consistency of gender differences across the 30 countries. In the second approach we transformed the 30-group analysis into a 60-group analysis in which responses by boys and girls within each country were used to form separate groups.

MIMIC model gender differences: AwC extensions of the CFA-MI$_{AL}$ model. Because the scalar invariance model does not fit the data, we instead evaluate gender differences based on the MG-MIMIC model with AwC (MG-MIMIC4$_{AwC}$; see Appendix 2 for syntax in an annotated example). Gender differences (and confidence intervals) in each of the 30 countries are graphed in Figure 3 for self-concept and general value. Controlling for SES and achievement, boys tend to have higher self-concepts across all 30 countries (β = .010 to .243, median = .135). Although boys are also favored in general value, the differences are smaller (-.085 to .157, median = .041), and in some countries the differences favor girls. Consistent with these observations, the result of the WALD test applied to gender difference shows highly significant country-to-country variability in the size of gender differences in self-concept (Wald $\chi^2(29)$ = 494.630, $p < .001$) and, to a lesser extent, general value (Wald $\chi^2(29)$ = 225.015, $p < .001$)

60-group CFA model of gender differences: AwC extensions of alignment. In an alternative approach to testing gender differences, we began with 60 (30 countries x 2 genders) groups rather than 30. This approach is less parsimonious than the MIMIC approach but more flexible in terms of testing the scalar invariance assumption over gender, which is not easily tested with the MG-MIMIC model. The configural 60-group CFA model with the eight motivational constructs (MCG1 in Table 2) provided a good fit to the data (CFI = .950, TLI = .946, RMSEA = .036). As in earlier analyses there was only a small decrease in fit indices for the metric model in which factor loadings were constrained to be equal over the 60 country-gender groups (MCG2 in Table 2; ΔCFI = .007, ΔTLI = .005, ΔRMSEA = .002). However, again the scalar invariance of intercepts (MCG3) was

not supported in relation to the substantial decreases in fit indices compared with Model MCG2 (Δ

CFI = .049, Δ TLI = .045, Δ RMSEA = .013), leading us to pursue the 60 country-gender groups

CFA model with alignment and AwC.

It should be noted that the MG-CFA approach used here relies heavily on the flexibility of the

"model constraint" command in Mplus to calculate gender differences, with the delta method being

utilized to estimate the standard errors. The AwC alignment model (MCG4$_{AwC}$) has the same degrees

of freedom, same chi-square and model fit as the configural CFA model (MCG1). For the purposes of

this investigation, a graphical depiction of the patterns of gender differences in self-concept and

general value is presented in Figure 3. There is clear evidence that gender differences in self-concept

and general value vary substantially by countries (Wald $\chi^2(29)$ = 264.292, $p < .001$, Wald $\chi^2(29)$ =

194.702, $p < .001$, respectively).

Furthermore, to explore the sizes of latent mean differences in motivational constructs across

countries and gender, we decomposed variance estimates into contrast tests of differences associated

with the 30 countries, the two gender groups, and their interactions; and estimated variance

components for each of these differences (sums of squares and variance components in Table 5) using

the "model constraint" command in Mplus. Thus, we used these constraints to obtain analysis-of-

variance-like estimates of the statistical significance and proportion of variance in latent mean

differences explained by the 30 countries, the two gender groups, and the 30 Country X Gender

interactions (see Marsh et al., 2013 for a related approach). Comparison of the variance components

shows that first-order gender differences (.040 and .006 for self-concept and general value

respectively) are much smaller than those associated with either the gender-by-country interactions

(.135 & .116) or, and in particular, the first-order effects of country (2.404 & 1.366). However, due to

the large sample sizes, all these effects are highly significant.

Of particular interest to the application of the CFA-MI$_{AL}$ model and the AwC, we compared

the pattern of gender differences based on 60 group alignment models with those based on the 30-

group MIMIC model (see Figure 4). Inspection of the caterpillar plots for the two approaches

demonstrates that they highly similar, particularly for self-concept, but to a lesser extent also for general value.

Discussion, and Implications: Study 1

Study 1 is apparently the first large-scale applications of the CFA-MI$_{AL}$ model proposed by Asparouhov and Muthén (2014) in which there were multiple factors as well as large numbers of items, factors, groups, individuals, and estimated parameters. Particularly the 60-group analysis used to assess gender differences is one of the largest published CFA-MI studies. In accomplishing this goal, we introduced the new AwC approach and demonstrated its usefulness, substantially enhancing the flexibility of the CFA-MI$_{AL}$ model in relation to substantively important issues that could not be evaluated appropriately using traditional MG-CFA methods. Of particular interest, invariance of item intercepts was not supported, and the scalar model provided poor model fit on the basis of the traditional scalar CFA-MI model; this implied the incomparability of factor means. However, we found that the CFA-MI$_{AL}$ model provided a much better fitting model that allowed us to compare means across the 30 countries. We also demonstrated how alignment was useful for developing or revising a scale measuring science motivation, in terms of cross-cultural generalizability.

In demonstrating the substantive usefulness of the CFA-MI$_{AL}$ model and the AwC extension, we evaluated the consistency over 30 OECD countries of latent means of the motivational constructs, as well as relations between the motivation constructs and the three criterion variables (gender, achievement, and SES). The associations between the motivational constructs and the criterion variables varied substantially over countries. On average, science achievement was positively associated with the motivational constructs, whereas associations of gender and SES to the motivational constructs were mostly small. Of particular interest, we evaluated gender differences in self-concept and general value on the basis of the 30-group MIMIC model (i.e., gender as a MIMC variable) and the 60-group AwC model (60 = 30 countries x 2 genders). Both models resulted in highly similar patterns of results, indicating that boys tended to have high self-concept in science, whereas the gender difference favoring boys in general value was relatively small. There was, however, country-to-country variation in the results, which necessitated the AwC extension of the CFA-MI$_{AL}$ model. In pursuing the methodological aims of this investigation, we demonstrated the

flexibility of the AwC extension of the CFA-MI$_{AL}$ model and its applicability to a wide variety of different situations that are likely to be useful for applied researchers, given that the CFA-MI$_{AL}$ model as currently operationalized can only be used to test a limited number of CFA models.

In summary, the results of Study 1 are supportive of alignment, particularly when extended to include the AwC transformation. Nevertheless, as alignment is a new statistical approach, "best practice" will evolve with experience. In particular, there are key questions arising from the results of Study 1 that we address in Study 2, which is based on simulated data, to provide a stronger basis for evaluating alignment in relation to viable alternatives.

Study 2: An Overview of the Substantive and Methodological Focus

Study 2 a simulation study, allowed us to evaluate the appropriateness of alignment in relation to known population parameters under a variety of different conditions. Of particular relevance to our earlier discussion of problems with the stepwise approach in the traditional partial invariance model, we compare the known parameter values from the population generating model with estimated values based on the alignment model and both the complete and the partial invariance scalar models. In order to enhance comparability, we then built on the simulation design that Asparouhov and Muthén (2014) used to introduce alignment, and address several critical issues left unanswered by Study 1 and the Asparouhov and Muthén demonstration—particularly in relation to estimates of latent means, which were the primary focus of Study 1, as they are in studies of scalar invariance more generally. More specifically we addressed the following issues that followed from limitations of Study 1, which relied on "real" data and a limited amount of alignment research to test the following a priori hypotheses:

1. When scalar invariance does not hold, bias in estimation of known latent means is consistently smaller for alignment than for either the complete or partial scalar approaches. (We leave as a research question the difference in bias between the complete and partial scalar CFA-MI solutions, and whether this difference is consistent over different conditions.)

2. When scalar invariance does hold, bias in estimates will be small and similar in size for both the scalar invariance and alignment models.

3. Cross-validation using new data from the same population generating model will support the superiority of alignment in relation to Hypothesis 1.

In addition to these a priori hypotheses, in Study 2 and subsequent discussion we address the unresolved question of how to evaluate the suitability of the alignment and AwC models.

Methods

In Study 2 we extended the original Asparouhov and Muthén (2014) simulation study, providing an overview of the quality of the alignment estimation in comparison with configural, scalar, and partial invariance models. An apparently unique feature of our simulation is that we rejected the typical assumption of CFA-MI models (and most simulation studies) that some parameter estimates are completely invariant across all groups. Instead, none of the parameter values in our population generating model were all non-invariant (i.e., none were exactly the same in the multiple groups), as would be the case in practice with real data. In this respect, we explore how well the alignment optimization functioned under the complete non-invariance condition with different patterns of large and small non-invariant parameters. As in study 1, the fixed alignment estimation method was used with 500 replications and maximum Likelihood in this simulation study.

Data Generation

On the basis of the Asparouhov and Muthén (2014) simulation study, we generated data using a one-factor model with five indicator variables and 15 groups. In all groups the residual variances of indicator variables were set to 1. The simulation design factors manipulated in the study included: (1) group size ($N = 100$ and 1000), (2) magnitude and percentage of non-invariance (10% large + 90% small; 20% large + 80% small); (3) approaches to invariance testing (alignment, configural, complete scalar, and partial scalar). Using the same method as Asparouhov and Muthén (2014), we generated three group types, and then repeated those types to create 15 groups. Each group type had the same parameter values. For example, the first, fourth, and seventh groups were simulated in the same manner. For group type 1 the distribution of the factor was $\alpha = 0$, $\psi = 1$; for group type 2, $\alpha = .3$, $\psi = 1.5$; and for group type 3, $\alpha = 1$, $\psi = 1.2$: this is consistent with the group types in Asparouhov and Muthén (2014). The alignment factor mean and factor variance were fixed to 0 and 1 respectively in

the first group; this matches the metric used to generate the data (see Asparouhov & Muthén, 2014 for more details).

Magnitude and percentage of non-invariance. For the pattern of loading and intercept non-invariance, two misfit conditions (small and large) were simulated. In each group there was one large non-invariant intercept parameter (e.g., $\upsilon = .50$ or $-.50$) and one large non-invariant loading parameter (e.g., $\lambda = 1.40, .50,$ or $.30$). The rest of the intercept and loading parameters were set to reflect a small extent of non-invariance ($\upsilon = 0 \pm .05 \; and \; \pm .10$ and $\lambda = 1 \pm .05 \; and \; \pm .1$). On the basis of this design (see Table 6), the ranges of the standard deviation of each loading and intercept across groups were from .04 to .23 and from .08 to .25 (see Table 6). In order to vary the percentage of large non-invariant parameters, we replaced large non-invariant loading values with small non-invariant values from each odd-numbered group and replaced large non-invariant intercept values with small non-invariant values from each even-numbered group. Also, to provide a test of the alignment model when there was complete scalar invariance, we simulated two additional groups with all non-invariant loading and intercept values ($\upsilon = 0, \lambda = 1$).

Approaches to Invariance Testing

We compared the alignment estimation with configural, scalar, and partial invariance models across all conditions (number of groups, magnitude and percentage of non-invariance, and approaches to invariance testing), totaling 12 conditions. This is apparently one of the few simulation studies to test the traditional stepwise adjustments to the scalar invariance model, and the first to juxtapose it with alignment. We suspect that this is due at least in part to the unique complications of applying this stepwise approach across a large number of replicates, even when the population generating model is known. In particular, the final solution for each replication can differ substantially terms of the number of post-hoc adjustments that are made, as well as which parameter estimates that are actually freed. In our operationalization of the stepwise approach to partial invariance, we first compared relative model fits for the configural and scalar invariance models based on each replication. At each step of the stepwise procedure within a given replication, if the ΔCFI was greater than .01 (Cheung and Rensvold, 2002 and Chen, 2007), an additional parameter, that having the largest modification

index, was freed. We repeated this procedure until ΔCFI ≤ .010, at which point we terminated the

iterative process and started again with the next replicate. Note that this procedure was used for each

replication, so that although the CFI for each of the 500 replicates was necessarily similar to the

configural model, the number and choice of invariance constraints that were freed varied across the

different replicates.

Measurement estimate analyses. To explore how well alignment estimated the group-

specific measurement models, we considered a variety of measures of accuracy and precision. Our

emphasis was on the latent means that are the focus of the present investigation, as they are in most

scalar invariance studies. However, across the 500 replicates we also report the mean, SD and average

mean square error (MSE) of bias (difference between the estimated and the true value) for factor

means, factor variances, loadings, and intercepts. The MSE captures the bias and variability of the

estimates by summing the square of the bias and the variance of the estimate. In addition, for every

replicate solution in each condition, we cross-validated the parameter estimates to test Hypothesis 3.

This was accomplished separately for each replicate by using the fixed values based on the solution

for that replicate applied to a new sample of cases generated from the same population generating

model (i.e., same sample size and values for large and small non-invariant parameters).

<div align="center">Results</div>

The goodness of fit measures (Table 7) merely confirm the design of the simulation study.

The goodness of fit measures (Table 7) merely confirm the design of the simulation study. The fit of

the configural and partial invariance models were similar and extremely high (e.g., CFIs ≥ .989) for

all conditions (i.e., small vs. large *N*; 10% vs 20% large misfit). The fit of the metric model was

marginal (CFIs = .905 to .942) and the fit of the scalar model was clearly unacceptable (CFIs = .819

to .876). For both the scalar and metric models, the fit was noticeably worse when the number of large

non-variant parameters was larger. Also of note, the number of post-hoc estimates freed in the partial

invariance models (i.e., the number of parameters in the partial solution less the number of parameters

in the scalar solution) varied systematically across replicates within each condition; on average, the

number of adjustments was greater when the amount of misfit was greater, but also when the sample

size was larger.

Latent Mean Bias When Scalar Invariance is Violated (Hypothesis 1)

The central results of the simulation study (Table 8) were designed to test Hypothesis 1. For the present purposes we focus on bias in the estimation of the latent factor means (α in the column labeled "average bias" in Table 8), but also present values for other parameter estimates as well. Consistent with a priori predictions, across all conditions average bias in latent means was systematically smaller for the alignment solutions than for either the complete or partial scalar solutions. Although bias was left as a research question, it is important to note that the average bias was also consistently larger for the partial than for the complete scalar condition. This pattern of differences (alignment better than scalar; complete scalar better than partial scalar) is consistent across all sample sizes and non-invariant conditions.

The pattern of results was somewhat more complicated for the variation in bias estimates across the different conditions (α in the column labeled "SD of bias" in Table 8). Again, consistent with Hypothesis 1, the variation in bias in latent mean estimates is consistently smaller than variation for the complete and partial scalar solutions. However, when the number of large non-invariant parameters is small (10% vs. 20%), the variation in bias is greater for the complete scalar than the partial scalar solutions, whereas when the number of large non-invariant parameters is large variation in bias estimates is greater for the partial solutions than the scalar solutions. Not surprisingly, the variation in bias estimates is systematically smaller when sample size is larger (1,000 vs. 100).

Average mean square error (MSE in Table 1) integrates average bias and variation in bias into a single index. Hence, it is not surprising that the alignment solutions performed systematically better than either the complete or partial scalar solutions. Consistently with the average bias results, the complete scalar solutions performed better than did the partial scalar solutions. However, consistently with the SD of bias results, the difference between complete and partial scalar conditions was larger when the number of large non-invariant parameters was small.

In the final columns in Table 8, we have translated the size of bias estimation in the latent means into an effect-size metric—average bias divided by the pooled standard deviation of the latent mean estimates. However, these values closely mirror those based on the average bias.

In summary, there is clear support for Hypothesis 1. At least in terms of the conditions in our simulation, alignment outperformed both the complete and partial scalar approaches when there was no support for complete scalar invariance. Although it was not predicted a priori, the surprisingly poor performance of the partial scalar solution in relation to the complete scalar solution was consistent with negative reviews of the stepwise approach used to make adjustments in the partial scalar model.

Latent Mean Bias When There is Support for Scalar Invariance (Hypothesis 2)

Although this was not a major focus of the present investigation, it is relevant to evaluate the alignment solution in relation to the complete scalar solution when there was support for scalar invariance (in Table 8, rows with percentage of non-invariance = 0%). Importantly, for both the alignment and complete scalar solutions, there was almost no bias in estimation of the latent means. Also, the variation in the bias estimates was nearly the same for the two sets of solutions. Indeed, the mean square errors (MSEs) that take into account both systematic bias and variation are also very small and identical (to three decimal places) for the complete scalar and alignment solutions. Again, the SDs of the bias in estimates (and MSEs) are smaller when the sample size is larger.

It is also interesting to compare these SDs of bias with those based on solutions where scalar invariance does not hold. These SDs are clearly smaller when there is complete scalar invariance, but the sizes of these differences vary substantially for complete scalar and alignment solutions. In particular, variation in alignment solutions is only modestly smaller, whereas the variation in the complete scalar solutions is substantially smaller. These results are also consistent with the a priori hypothesis that even when the scalar solution is viable, alignment is still appropriate. In summary, there is clear support for Hypothesis 2. At least in terms of the conditions in our simulation, nothing is lost by applying alignment, even when there is support for complete scalar invariance.

Cross-Validation Support for the Results (Hypothesis 3)

Consistent with a priori Hypothesis 3, there is good cross-validation support for the results in support of Hypothesis 1. Indeed, the cross-validation indices in relation to bias in the latent means in Table 9 are nearly identical to those in Table 7. Although this finding is tangential to the main focus of the present investigation, the reason why the cross-validation indices are so good is that both the alignment and, in particular, the partial invariance approaches, were designed to optimize the

goodness of fit of solutions in relation to factor loadings and intercepts, not the latent means. Hence, the inevitable deterioration due to capitalization on chance in cross-validation is not large for the bias in estimation of latent means.

Discussion and Implications: Study 2

Building on the original Asparouhov and Muthén (2014) simulation study, the results of Study 2 provide important new support for alignment and thus for AwC, which was the major focus of our study. Whilst, in support of alignment, Asparouhov and Muthén presented results based on a few indicative parameter estimates from just one group, we have provided a more comprehensive evaluation of results across all parameter estimates and all groups. More importantly, we expanded the simulation study to include partial scalar invariance estimates. This is particularly important because the stepwise strategy continues to be used widely with partial scalar invariance, even though this has been criticized severely by statisticians and quantitative psychologists alike. Indeed, even the important caveats offered by Byrne, Shavelson, and Muthén (1989) when they first introduced the partial invariance strategy, have tended to be ignored in subsequent research. Critically, consistent with a priori predictions in relation to latent means, the results of Study 2 support the a priori hypotheses that alignment outperforms both the complete and partial scalar approaches when the fit of the complete scalar model is unacceptable, and performs no worse than the complete scalar solution even when there is complete scalar invariance.

Overall Discussion, Limitations and Directions for Future Research

Study 1 is apparently the first large-scale application of the CFA-MI$_{AL}$ model, and one of the largest applications of the CFA-MI approach, with so many factors, items, and estimated parameters. Indeed, most CFA-MI demonstrations focus on a small number of groups (e.g., Byrne, et al., 1989; Reise, Widaman & Pugh, 1993), whilst the relatively few studies based on a large number of groups often consider a single factor based on a relatively small number of items (e.g., Zercher et al., 2015). In Study 1 we could have considered a single factor or each of the eight factors in isolation. However, the initial focus was to follow up the Marsh and Nagengast (2013) study, where the focus was on the fit of the multidimensional factor structure across all eight factors. Obviously this was only possible through considering all eight factors within the same models. Indeed, even if there were good support

for the fit of each factor considered separately, there is no guarantee that a model with all the factors

in the same model would fit. Furthermore, because of the moderate to large correlations among the

different factors, not even the estimated factor loadings and intercepts would have been the same in

models of each factor considered separately. Although models of each factor considered separately

might provide supplemental information, this information and more is already available through the

alignment model of all eight factors. In summary, the scale of data in Study 1 provided a realistically

complex demonstration of the CFA-MI$_{AL}$ model in relation to actual practice.

Introduction of AwC and Parallels With Exploratory Structural Equation Models (ESEM)

A critical feature of Study 1 was the introduction of the AwC extension, which transforms

alignment into a confirmatory tool rather than being largely exploratory. The AwC extension greatly

enhances the usefulness and flexibility of alignment to address substantively important issues in

further CFA and SEM analyses that would not otherwise be possible with alignment. It is also

interesting to explore some of the similarities between the development of alignment and ESEM. In

both cases, development came about because of the typically overly restrictive assumptions of the

traditional CFA model; requiring cross-loadings to be zero (ESEM); the scalar invariance constraints

in CFA-MI models (alignment). In both cases, the apparently inherent limitations of ESEM and

alignment were mostly overcome by the introduction of EwC and AwC, transforming exploratory

tools to confirmatory, and greatly expanding the range of models that could be considered. Indeed,

because the EwC approach has been widely applied, some of the novel applications of the EwC

extension to ESEM (Marsh, Morin et al., 2014) are likely to be valuable to the application of AwC, as

well as to future developments of Mplus to facilitate these applications.

The juxtaposition of the ESEM and alignment also identifies potentially serious limitations of

alignment as currently specified, in that it begins with an implicit assumption that the configural CFA-

MI model is able to fit the data. However, as presently operationalized, the CFA-MI$_{AL}$ model is

limited to independent cluster factor structures in which indicators are not allowed to cross-load on

multiple factors. However, this factor structure, which underpins most CFA studies, is overly

restrictive in many applications (Marsh, Lüdtke, et al., 2013), leading to a growing body of research

suggesting that the cross-loadings in ESEM often provide a more appropriate, better-fitting solution.

The introduction of AwC allows limited scope in testing and perhaps, relaxing this requirement of no cross-loadings—but is limited in that substantial cross-loadings would call into question the alignment structure that is the basis of AwC. Similarly, although multigroup tests of invariance are possible with ESEM, they suffer the same limitations with CFA-MI models as have been highlighted in the present investigation. We also note that tests of longitudinal measurement invariance over multiple occasions is not possible with alignment in its current form, but is possible with ESEM. Recognizing the potential synergy between the ESEM and alignment, Asparouhov and Muthén (2014) mooted the combination of ESEM and alignment into a single model as a useful development in future versions of Mplus. This development would also enable applied researchers to use both the AwC and the EwC transformations in the same analysis.

Comparison of Alignment and Partial Invariance Approximations to Scalar Invariance

An obvious limitation of Study 1 is that it left unanswered the question of how alignment would compare with the traditional stepwise approach used to achieve partial invariance. This could not be easily addressed with real data in which the true population parameters are unknown. Thus, we undertook a simulation study (Study 2) to evaluate the extent of bias in estimation of latent means based on alignment, compared with complete and partial invariance models under a variety of different conditions. In relation to the degree of non-invariance associated with our design in Study 2, the fit of the configural model was obviously better than that typically found in practice. However, even for the condition where the number of large non-invariant parameters is small, the fit of the metric and scalar models is somewhat poorer than that observed in Study 1, suggesting that the extent of non-invariance in the simulated data is greater than that in Study 1.

The Study 2 results are unambiguous, in that alignment consistently outperformed partial invariance in particular, as well as the complete scalar invariance models. Of course, as is always the case with simulation studies, the generalizability of these conclusions is limited by the design of the study (see discussion of limitations below). However, our simulation study should have been ideally suited to the partial invariance strategy, in that there were only a few large non-invariant parameter estimates, in combination with many small ones. Nevertheless, given the scathing reviews of stepwise procedures generally (see earlier discussion of problems with stepwise approaches), perhaps it is not

surprising that stepwise approaches perform so poorly. From this perspective, it is somewhat

surprising that applied SEM/CFA researchers have persevered so long with a procedure that is so

dubious. Indeed, such issues were recognized by Byrne, Shavelson, and Muthén (1989) when they

first introduced the partial invariance more than 25 years ago, and the failure to resolve these long-

standing issues was a primary motivation for Asparouhov and Muthén (2014) introducing alignment

as a viable alternative to partial invariance. The results of the present investigation, the first empirical

test of this implicit assumption, provide clear support for alignment and further call into question the

traditional partial invariance approach.

We also note that the partial invariance model does not have to be driven purely by a stepwise

empirical approach, even though this is the typical approach (Schmitt & Kuljanin, 2008). Indeed,

defenses of the procedure, starting with the original Shavelson et al. (1989) demonstration, note the

need to evaluate the selection of parameters to be freed in relation to theory and substantive

knowledge. However, this tends to be done in a strictly post-hoc manner to justify the results of the

stepwise empirical selection process (Schmitt & Kuljanin, 2008). A more appropriate use of theory

and substantive knowledge might be to develop truly a priori models that could then be empirically

tested in relation to goodness of fit and evaluation of parameter estimates (MacCallum, et al., 1992).

Here we have pitted the alignment and partial invariance approaches against each other,

treating them as antithetical. However, this perspective might be too simplistic, and we speculate that

a synergistic combination of both approaches could be advantageous. Modification indices are the

critical feature of the typical partial invariance model. Although modification indices and expected

change parameters are not currently available with the alignment model, they are readily available for

the equivalent AwC model. However, indices based on the final AwC model are fundamentally

different from the modification indices used in the partial invariance model, particularly in relation to

identifying parameters that cause the most stress to scalar invariance. Thus, the modification indices

that these are based on in the final and "best" AwC model can, and should, be added as a single step

rather than one at a time in the potentially many steps of the forward stepwise approach. In this sense,

the adjustments identified by the AwC model are more like the "all possible combinations" approach

to stepwise selection, which has important advantages over (in particular) the forward stepwise

strategy typically used, and also backward elimination and bidirectional elimination (a combination of forward and backward approaches). Thus, a potential synergy between alignment and partial invariance models could be to use the modification indices based on the AwC model to identify parameters to be freed in the partial invariance model.

How to Evaluate the Appropriateness of the Alignment and AwC Models

   Limitations in the AwC model. For demonstrations based on new statistical procedures, typically there are potentially important limitations and a need for further research. Particularly in relation to the limitations identified by Asparouhov and Muthén (2014), the introduction of the AwC extension of the CFA-MI$_{AL}$ model is an important development, greatly expanding the range of models that can be considered with alignment, as illustrated in Study 1. There are, nevertheless, limits to the generalizability of results based on the AwC transformation of the original alignment solution to new models. In particular, there is an implicit assumption that the alignment factor structure continues to be appropriate when it is incorporated into new models that take advantage of the flexibility of AwC. However, we suggest that there is a hierarchy of models in relation to how reasonable this assumption is likely to be. At the top of the hierarchy, this assumption is entirely reasonable for the basic AwC model, which does not introduce new constraints or additional variables, as it is merely an equivalent transformation of the alignment model. The assumption is likely to be more reasonable when the new models are nested under the original model (e.g., more constraints are added) than when new variables are added. When new variables are added, the assumption is likely to be more reasonable when new variables are merely correlated with the alignment factors, or alignment factors are used to predict new variables, than in MIMIC models that impose additional invariance assumptions. For example, if the fit of the MIMIC alignment model in Study 1 had been much worse than that in the basic alignment model (or, equivalently, the configural model), then the results would have to be interpreted with caution. However, this concern is not specific to the alignment model, but also applies to MIMIC models in conjunction with the scalar CFA-MI models and single-group models. Indeed, under these circumstances it might be more appropriate to forgo the MIMIC model altogether and resort to an appropriate multiple group model. In Study 1 we demonstrated how this was possible in relation to gender differences, comparing

models with gender added to the AwC model as a MIMIC variable, and gender treated as a multiple group variable (i.e., creating separate male and female groups for each country).

Focus on latent means. Our focus was primarily on estimation of latent means, rather than on factor loadings, intercepts, and factor variance/covariance estimates. This focus is justified, in that the main purpose of tests of scalar invariance is to provide a justification for the evaluation of latent means. This also has some interesting implications in relation to the results. In particular, the stepwise strategy in the partial invariance model is designed to maximize goodness of fit in relation to adjustments to the factor loadings and intercepts in the complete scalar model, rather than latent means. From this perspective, it is not surprising that the results based on latent means cross-validated so well, in that the adjustments did not capitalize on chance in relation to latent means. Nevertheless, in other applications of alignment it might be important to evaluate the extent of bias in the estimation of other parameter estimates.

How large is a large non-invariant parameter? An ongoing, unresolved issue with alignment is how to evaluate the appropriateness of the solution when true population parameters are unknown. In particular, because the fit of the alignment model is necessarily the same as the configural model, its appropriateness cannot be evaluated by goodness of fit. We note, however, that this limitation also exists with the partial invariance model, in which a sufficient number of invariance constraints are freed so that its fit does not differ substantially from that of the configural model. Hence, the partial invariance model cannot be evaluated in relation to goodness of fit. On the basis of preliminary results, Asparouhov and Muthén (2014) suggested that alignment studies should be interpreted cautiously if more than 20% of the parameter estimates are non-invariant. However, this suggestion is, perhaps, overly simplistic. As shown here, alignment works well even when all of the parameters are non-invariant, as long as the deviations are small. Asparouhov and Muthén implicitly recognized this in that they focused on deviations that were statistically significant, and used a conservative criterion of $p < .001$. Nevertheless, because this criterion is highly sample-size dependent, guidelines based upon it are unlikely to be generalizable. Hence, what is needed is a more absolute index of what constitutes "large" that is relatively independent of sample size and practically useful.

The alignment solution routinely provides additional insights into the quality of alignment solutions in terms of the largest deviations in relation to individual indicators and groups. Although this information is clearly useful from a diagnostic perspective, it is based largely on tests of statistical significance that are highly dependent on sample size and thus, idiosyncratic to a particular data set. However, as illustrated here, these tests are easily supplemented with measures of practical significance by transforming the differences into a standardized effect size metric (Cohen's D) that is more comparable in relation to external comparisons with different studies, as well as internal comparisons within the same study. We also note that presenting these Cohen's D statistics in terms of box plots provides a useful summary of the distribution of values across different groups and items, particularly since the CFA-MI$_{AL}$ loss function is minimized when there are a few large non-invariant parameters and many approximately invariant parameters. Because our study is apparently the first application of standardized effect sizes (ESs) to evaluate the results from the alignment model, it is premature to provide guidelines about what constitutes large, medium and small ESs, but such intuitions should evolve with further application.

We also note that output from the alignment program currently does not include modification indices (which are highly influenced by sample size) or related measures of expected parameter change (raw and standardized), which provide a more practical, "absolute" index (that is sample size independent) of how much a fixed or constrained parameter would change if freed. However, with the basic AwC model these additional indices are readily available and likely to be useful in evaluating the extent of non-invariance for different parameter estimates. Whittaker's (2012) simulation study suggested that expected change parameters were somewhat better at identifying misspecified parameter estimates, but recommended using them in combination with modification indices. However, the potential value of the expected change indices is to provide a generalizable index of what constitutes a "large" misspecification—Whittaker suggested standardized values greater than .2.

Following Whittaker's suggestion, we evaluated the estimated parameters with the largest modification indices for the PISA data, along with standardized and unstandardized indices of expected change (see Appendix 7, Supplemental Materials). Although it is probably premature to propose cutoff values for alignment and AwC models that are based on Whittaker's results, which

emerged in a different context, it is interesting to note that less than 3% of the parameter estimates had completely standardized, expected parameter change values (STDYC_EPC in Appendix 7) greater than .2 in absolute value—far lower than the 20% cutoff suggested by Asparouhov and Muthén (2014). Consistently with suggestions by Whittaker, inspection of Appendix 7 indicates that modification and expected parameter change indices provide different perspectives, so that some combination of both might also provide a useful starting point for identifying parameters to free in partial invariance models that do not rely on apparently dubious, forward stepwise selection.

Alternative Approaches to Measurement Invariance

Recently there has been considerable development of alternative approaches to the evaluation of latent means in large-scale studies when there is a lack of support for scalar invariance. A number of studies have used multilevel modeling, treating the multiple groups as level 2 and the cases nested within each group as level 1 (see Jak, Oort & Dolan, 2013). However, implicit in the multilevel approach is the assumption that the groups are a random sample from a well-defined population in which the focus is on the population from which the groups have been sampled; group-specific values are assumed to represent random variation from this population value. In contrast, the MG-CFA approach treats groups as fixed effects, with inferences that focus on specific groups. Consistently with this distinction, alignment provides considerable information about the source of non-invariance that is generally not available with the multilevel approach. Muthén and Asparouhov (2013) also demonstrated that the multilevel approach is better suited to situations in which there is a very large number of indicators (e.g., items on an achievement test, as opposed to the relatively few items used to measure psychological constructs on most surveys). In addition to the multilevel approach, there are important developments in other evolving approaches, including Bayesian structural equation modeling (e.g., Zercher, et al., 2015), multilevel mixture modeling (Muthén & Muthén, 2011-2015). Also, perhaps, partial invariance models that do not rely on stepwise strategies will prove critical to the development of measurement invariance models.

In summary, alignment augmented by AwC provides applied researchers with considerable flexibility to address substantively important issues when the traditional CFA scalar model does not fit the data. Both our review of the literature condemning stepwise selection strategies, and our

empirical results, suggest that alignment is more appropriate than the typical practice of stepwise partial invariance. The introduction of AwC transforms alignment from being largely exploratory into a confirmatory tool, and substantially increases the range of situations in which it can be used. Although alignment and AwC provide a wealth of information to evaluate the quality of the alignment solution, an unresolved issue is how to evaluate whether the alignment solution is trustworthy in relation to evaluating latent means from multiple groups. This is, perhaps, not surprising, because essentially the same problem was identified by Byrne, et al (1989) when they first introduced partial invariant models and the problem has not been resolved in the subsequent 25 years, in terms of evaluating partial invariant models.

We offer some tentative solutions to this issue as directions for further applied and simulation research. Despite their limitations we are confident that, given that these are new statistical tools, "best practice" will evolve with experience. Other potentially important directions for further research include synergistic combinations of the advantages of alignment with other approaches, such as ESEM (particularly in relation to cross-loadings, but also longitudinal invariance), partial invariance models (based on adjustments identified by alignment and AwC, rather than stepwise strategies), multilevel modeling, mixture models, and Bayesian structural equation models.

Footnotes

1. We also note that the standard errors for all parameter estimates were very similar in the M_{AL} and AwC solutions, but not exactly identical. This is necessarily the case, in that some parameters in the M_{AL} solution are freely estimated, while they are fixed in the AwC solution (e.g., one factor loading for each factor; see Appendix 1 in the Supplemental Materials for further discussion).

2. In the AwC extension of the MIMIC model, the parameter estimates from the alignment solution based on 30 groups were used as starting values. For model identification, the first loading and intercept for each factor was fixed to its estimated values from the alignment solution, and latent factor variance (residual variance) and means were freely estimated in each group (see Appendices for more detail).

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, 16, 397–438.

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, (September), 1–14. doi:10.1080/10705511.2014.919210

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association, 74,* 1–4. doi:10.1080/ 01621459.1979.10481600

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. doi:10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. doi:10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018.http://dx.doi.org/10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. doi:10.1207/S15328007SEM0902_5

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40,* 55–75. http://dx.doi.org/10.1146/annurev-soc-071913-043137

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge, United Kingdom

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: Guilford.

Harrell, F. E. 2001. *Regression Modeling Strategies*. Springer-Verlag, New York.

He, J. & Kubacka, K. (2015). *Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013.* OECD Education Working Papers, No. 124, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrp6fwtmhf2-en

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55. doi:10.1080/10705519909540118.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: the oblique case. *Psychometrika*, *71*(1), 173–191. doi:10.1007/s11336-003-1136-B

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling, 20*(2), 265-282. http://dx.doi.org/10.1080/10705511.2013.769392

Judd, C. M. and McClelland, G. H. 1989. Data *Analysis: A Model-Comparison Approach.* Harcourt Brace Jovanovich, San Diego, CA.

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504. http://dx.doi.org/10.1037/0033-2909.111.3.490

MacCallum, R. C. (2003). 2001 Presidential Address: Working with imperfect models. *Multivariate Behavioral Research, 38,* 1, 113–139. doi:10.1207/S15327906MBR3801_5

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504. http://dx.doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 391–410.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu and

Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341. http://dx.doi.org/ 10.1207/s15328007sem1103_2

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling*, *1*, 317-359.

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6, 311–360. doi:10.1207/s15327574ijt0604_1

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18*(3), 257-284. http://dx.doi.org/10.1037/a0032773,

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85-110. doi:10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, 16(3), 439-476. doi:10.1080/10705510903008220

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

Marsh, H.W., Morin, A. J .S., Parker, P., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the best Features of Exploratory and Confirmatory Factor Analysis. *Annual Review of Clinical Psychology*, 10. 85-110

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science, 5,* 675– 686. doi:10.1177/ 1745691610388766

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Taylor & Francis Group.

Muthén, B., and Asparouhov, T. (2013). *New Methods for the Study of Measurement Invariance with Many Groups. Technical report.* Available online at: http:// statmodel2.com/download/PolAn.pdf.

Muthén, L. K., & Muthén, B. O. *(1998–2015). Mplus user's guide (version 7.11).* Los Angeles, CA: Muthén & Muthén.

Nagengast, B., & Marsh, H. W. (2013). Motivation and engagement in science around the globe: testing measurement invariance with multigroup SEMs across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier, D. Rutkowski (Eds.). *A Handbook of International Large-Scale Assessment Data Analysis* (pp. 317-344). Chapman & Hall, CRC Press

OECD. (2009). *PISA 2006. Technical Report.* Paris, France: OECD.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley. doi:10.1002/9780470316696

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566. http://dx.doi.org/10.1037/0033-2909.114.3.552

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57.http://dx.doi.org/10.1177/0013164413498257

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210-222. http://dx.doi.org/10.1016/j.hrmr.2008.03.003

Sörbom (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology*, 27, 229-239.

Thurstone, L. L. (1930). The learning function. *Journal of General Psychology, 3,* 469–493. doi:10.1080/00221309.1930.9918225

Tukey, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and

spectrum analysis. *Technometrics, 3,* 191–219. doi:10.1080/00401706.1961.10489940

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance

literature: suggestions, practices, and recommendations for organizational research.

Organizational Research Methods, *3*(1), 4–70. doi:10.1177/109442810031002

Whittaker, T. A. (2012) Using the modification index and standardized expected parameter change for

model modification, The Journal of Experimental Education, 80, 26-44, doi:

10.1080/00220973.2010.531299

Widaman, K. F., & Reise, S.P. (1997). Exploring the measurement invariance of psychological

instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West

(Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse*

research (pp. 281–324). Washington, DC: American Psychological Association.

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism

value over time and across countries in the European Social Survey: Exact vs. approximate

measurement invariance. *Frontiers in Psychology, 6,* Article ID 733.

doi:10.3389/fpsyg.2015.00733

Table 1

Information on the Eight Motivational Constructs in This Study

Motivational constructs	Number of Items	Median reliability α over countries	Median factor loadings (total sample)	Parameter invariance status (percentage of invariant parameters based on the alignment method)[a]		Difference of alignment and scalar model standardized to Cohen's D[b] (Mean[SD])[b]		
				Loadings	Intercepts	Loadings	Intercepts	Mean
Enjoyment	5	.92	.844	82.7%	49.3%	.004(.037)	.003(.069)	-.002(.017)
Instrumental motivation	5	.92	.833	77.3%	61.3%	-.001(.039)	.027(.046)	-.020(.011)
Future-oriented motivation	4	.92	.887	62.5%	55.8%	.045(.069)	-.089(.078)	.057(.036)
Self-efficacy	8	.83	.630	85.4%	47.9%	.003(.038)	.020(.087)	-.017(.017)
Self-concept	6	.92	.843	58.9%	58.9%	-.003(.043)	.029(.058)	-.026(.010)
General value	5	.75	.615	90.0%	50.7%	-.005(.029)	-.010(.101)	.001(.020)
Personal value	5	.80	.715	72.7%	52.7%	.007(.039)	.006(.080)	-.002(.021)
Extracurricular	6	.78	.642	81.1%	62.2%	.013(.057)	-.190(.115)	.140(.050)

Note. [a] Total number of approximate measurement invariance groups across indicators divided by total number of groups across indicators. [b] Cohen's D is computed by the differences of unstandardized loadings/intercepts between alignment and scalar models, divided by pooled standard deviation.

Table 2
Model Fit Statistics for Multiple-Group and MIMIC Models Based on 30 Countries

Models	Description	χ^2	df	Params	CFI	TLI	RMSEA
	Total group (TG) models						
TG1CFA	Total group CFA	32752	874	160	.963	.960	.012
TG2CFA	Total group CFA with covariates	40598	982	193	.955	.950	.013
TG2SEM	Total group MIMIC	40598	982	193	.955	.950	.013
	Multiple-group (MG) models (30 groups)						
MG1	Configural	183577	26220	4800	.952	.948	.027
MG2	IN = FL	205325	27264	3756	.946	.943	.041
MG3	IN = FL, INT	334112	28308	2712	.906	.906	.036
MG4$_{AL}$	Alignment	183577	26220	4800	.952	.948	.027
MG4$_{AwC}$	Alignment with AwC	183577	26220	4800	.952	.948	.027
MG5$_{AwC}$	Align, IN = Uniq	279428	27496	3524	.923	.920	.033
MG6$_{AwC}$	Align, IN = FV	190730	26452	4568	.950	.946	.027
MG7$_{AwC}$	Align, IN = FV, CV	199820	27264	3756	.947	.945	.028
MG8$_{AwC}$	Align, IN = FV, CV, Uniq	295985	28539	2481	.918	.919	.034
	Multiple-group MIMIC (gender, SES & ACH as covariates)						
MG-MIMIC1	Configural	225544	29460	5790	.942	.937	.028
MG-MIMIC2	IN = FL	247807	30504	4746	.936	.932	.029
MG-MIMIC3	IN = FL,INT	378195	31548	3702	.898	.895	.036
MG-MIMIC4$_{AwC}$	Alignment	225544	29460	5790	.942	.937	.028
	Multiple-group (MG) models (60 groups: 30 Countries x 2 Gender)						
MCG1	Configural	337910	52440	9600	.950	.946	.036
MCG2	IN = FL	380127	54564	7476	.943	.941	.038
MCG3	IN = FL,INT	626837	56688	5352	.901	.901	.049
MCG4$_{AwC}$	Alignment	337910	52440	9600	.950	.946	.036

Note. [AwC] alignment-within-CFA approach (AwC); CFI = comparative fit index; TLI = Tucker–Lewis Index; Params = number of free parameters; ACH = Science achievement; RMSEA = root mean squared error of approximation; For multiple group invariance models, IN = the sets of parameters constrained to be invariant across the multiple groups: FL = factor loadings; INT = item intercepts; FV = factor variance; CV = factor variance–covariances.

Table 3

Latent Correlations Among the Eight Motivational Constructs and the Three Covariates, Based on the Total Group CFA

	1	2	3	4	5	6	7	8	9	10
Enjoyment (1)	–									
Instrumental motivation (2)	.590	–								
Future-oriented motivation (3)	.661	.713	–							
Self-efficacy (4)	.486	.370	.375	–						
Self-concept (5)	.611	.572	.558	.551	–					
General value (6)	.518	.437	.386	.491	.399	–				
Personal value (7)	.705	.674	.666	.522	.560	.785	–			
Extracurricular activities (8)	.639	.464	.569	.452	.497	.411	.592	–		
Gender (9)	.068	.024	.074	.060	.136	.056	.053	.094	–	
SES (10)	.025	.019	(-.011)	.241	.083	.114	.058	-.030	(.004)	–
Science achievement (11)	.198	.081	.095	.372	.153	.262	.149	(.012)	(.012)	.449

Note. The correlation matrix is based on Model MG6$_{AwC}$. All correlation coefficients are statistically significant ($p < .001$), except for those in parentheses.

Table 4

Parameter Invariance Status of Factor Loadings and Intercepts Across Groups for Self-Concept and General Value Scales

Items	Descriptions	Measurement invariance status across countries[1]		Difference of alignment and scalar model standardized to Cohen's D^2 (Mean[SD])	
		Loadings	Intercepts	Loadings	Intercepts
Self-concept					
ST37Q01	Learning advanced science topics would be easy for me.	19	10	-.029(.060)	.003(.106)
ST37Q02	I can usually give good answers to test questions on science topics.	6	10	.013(.060)	.036(.052)
ST37Q03	I learn science topics quickly.	30	24	.011(.015)	.043(.029)
ST37Q04	Science topics are easy for me.	16	19	-.015(.022)	.020(.039)
ST37Q05	When I am being taught science I can understand the concepts very well.	9	18	.005(.033)	.039(.052)
ST37Q06	I can easily understand new ideas in science.	26	25	-.003(.030)	.033(.027)
General value					
ST18Q01	Advances in science usually improve people's living conditions.	26	22	-.008(.027)	-.02(.068)
ST18Q02	Science is important for helping us to understand the natural world.	25	21	.013(.033)	-.00(.107)
ST18Q04	Advances in science usually help improve the economy.	28	12	.001(.018)	-.00(.111)
ST18Q06	Science is valuable to society.	28	10	-.001(.024)	-.00(.074)
ST18Q09	Advances in science usually bring social benefit.	29	11	-.030(.025)	-.01(.136)

Note. [1]Number of approximate measurement invariance groups for each indicator divided by total number of groups (e.g., 30). [2] Cohen's D is computed by the differences of unstandardized loadings/intercepts between alignment and scalar models, divided by pooled standard deviation.

Table 5

Latent Mean Differences in Self-Concept and General Value of Science Across (30 Countries x 2 Gender) Groups

	Self-concept		General value	
	SS	VC	SS	VC
Gender	.040(.002)	0.05%	.006(.001)	0.06%
Countries	2.404(.079)	28.7%	1.366(.075)	13.1%
Interaction	.135(.017)	1.61%	.116(.017)	1.11%

Note. SS = sums of squares; VC = variance components. Latent mean differences in self-concept and general value were decomposed to assess the main effects of differences due to the 30 countries, the two gender groups, and their interaction.

Table 6.

Non-invariance Pattern Based on 20% Large Non-invariance

Parameter	Group 1	Group 2	Group 3	Mean across 15 groups	SD across 15 groups
Y1 loading	1.00	1.05	.95	1.00	.04
Y2 loading	1.00	1.10	.90	1.00	.08
Y3 loading	**1.40**	.90	1.10	1.13	.21
Y4 loading	1.00	.95	**.30**	.75	.33
Y5 loading	1.00	**.50**	1.05	.85	.26
Y1 intercept	.00	**-.50**	-.05	-.18	.23
Y2 intercept	.00	.05	**.50**	.18	.23
Y3 intercept	.00	-.10	.10	.00	.08
Y4 intercept	.00	.10	-.05	.02	.06
Y5 intercept	**.50**	-.05	.05	.17	.25
Factor mean	.00	.30	1.00	.43	.43
Factor variance	1.00	1.50	1.20	1.17	.21

Note. Large non-invariant parameters are shaded and bolded.

Table 7

Model Fit Statistics for Invariance Models

Models	N	% large noninvariance	χ^2	df	Params	CFI	TLI	RMSEA
configural	100	10%	77	75	225	.998	.999	.020
metric	100	10%	298	131	169	.942	.933	.112
scalar	100	10%	542	187	113	.876	.900	.138
partial	100	10%	200	170	130	.990	.991	.041
configural	100	20%	77	75	225	.998	.999	.020
metric	100	20%	386	131	169	.905	.891	.139
scalar	100	20%	672	187	113	.819	.855	.161
partial	100	20%	191	161	139	.989	.990	.043
configural	1000	10%	76	75	225	1.000	1.000	.006
metric	1000	10%	1770	131	169	.943	.934	.112
scalar	1000	10%	3734	187	113	.876	.900	.138
partial	1000	10%	425	164	136	.991	.992	.040
configural	1000	20%	75	75	225	1.000	1.000	.005
metric	1000	20%	2635	131	169	.907	.893	.138
scalar	1000	20%	4995	187	113	.821	.856	.160
partial	1000	20%	400	155	145	.991	.991	.040
continual	100	0%	77	75	225	.998	.998	.019
scalar	100	0%	192	187	113	.996	.998	.017
continual	1000	0%	75	75	225	1.000	1.000	.005
scalar	1000	0%	187	187	113	1.000	1.000	.004

Table 8

Average Bias, SD of Bias, and MSE for the FIXED Alignment Estimates Using Maximum Likelihood

Models	N	% large Non-invariance	Average bias				SD of bias				Average MSE				ES_within	ES_total
			α	ψ	λ	υ	α	ψ	λ	υ	α	ψ	λ	υ	α	α
Small *N* and Small Non-Invariance																
Align	100	10%	-.010	.030	-.007	.011	.164	.329	.133	.142	.027	.109	.018	.020	-.009(.148)	-.008
Scalar	100	10%	-.067	-.039	.007	.049	.195	.351	.188	.192	.042	.125	.035	.039	-.060(.149)	-.056
Partial	100	10%	.126	.057	-.015	-.107	.179	.363	.133	.149	.048	.135	.018	.034	.114(.162)	.106
Small *N* and Large Non-Invariance																
Align	100	20%	-.047	-.114	.047	.006	.170	.308	.148	.167	.031	.108	.024	.028	-.043(.153)	-.040
Scalar	100	20%	-.119	-.283	.084	.041	.189	.322	.244	.226	.050	.184	.066	.053	-.107(.171)	-.100
Partial	100	20%	.134	-.085	.035	-.145	.219	.358	.162	.214	.066	.136	.027	.067	.121(.198)	.112
Large *N* and Small Non-Invariance																
Align	1000	10%	.039	.008	-.004	-.033	.059	.109	.045	.051	.005	.012	.002	.004	.035(.054)	.033
Scalar	1000	10%	-.077	-.066	.014	.054	.106	.187	.167	.161	.017	.039	.028	.029	-.069(.096)	-.064
Partial	1000	10%	.158	.084	-.025	-.134	.067	.165	.071	.073	.029	.034	.006	.023	.143(.060)	.133
Large *N* and Large Non-Invariance																
Align	1000	20%	.000	-.091	.034	-.033	.072	.136	.063	.083	.005	.027	.005	.008	.000(.065)	.000
Scalar	1000	20%	-.129	-.305	.090	.047	.111	.218	.226	.196	.029	.140	.059	.041	-.117(.100)	-.108
Partial	1000	20%	.225	-.010	.004	-.200	.157	.133	.069	.168	.075	.018	.005	.068	.203(.142)	.189
Small *N* and No Large Non-Invariance																
Align[a]	100	0%	.016	.024	.007	.020	.156	.249	.134	.148	.027	.063	.018	.024	-.007(.141)	-.007
Scalar[a]	100	0%	.008	.031	.009	.001	.165	.250	.086	.111	.027	.063	.007	.012	.005(.149)	.004
Small *N* and No Large Non-Invariance																
Align	1000	0%	.002	.003	.001	.003	.052	.078	.043	.048	.003	.006	.002	.002	-.005(.047)	-.005
Scalar	1000	0%	.001	.004	.001	.000	.052	.078	.028	.035	.003	.006	.001	.001	.000(.047)	.000

Note. Align = Alignment. MSE = mean square error.

[a] in complete invariance models all factor loadings are set as 1, and all intercepts are set as 0.

Table 9.

Average Bias, SD of Bias and MSE for the Alignment Based on Cross-Validation Data

Models	*N*	Percentage of large non-invariance	Average bias	SD of bias	Average MSE
Align	100	10%	-.010	.167	.028
Scalar	100	10%	-.067	.197	.043
Partial	100	10%	.126	.179	.048
Align	100	20%	-.047	.173	.032
Scalar	100	20%	-.119	.192	.051
Partial	100	20%	.134	.220	.066
Align	1000	10%	.039	.060	.005
Scalar	1000	10%	-.077	.107	.017
Partial	1000	10%	.158	.066	.029
Align	1000	20%	-.001	.072	.005
Scalar	1000	20%	-.129	.111	.029
Partial	1000	20%	.225	.157	.075

Note. Align = Alignment
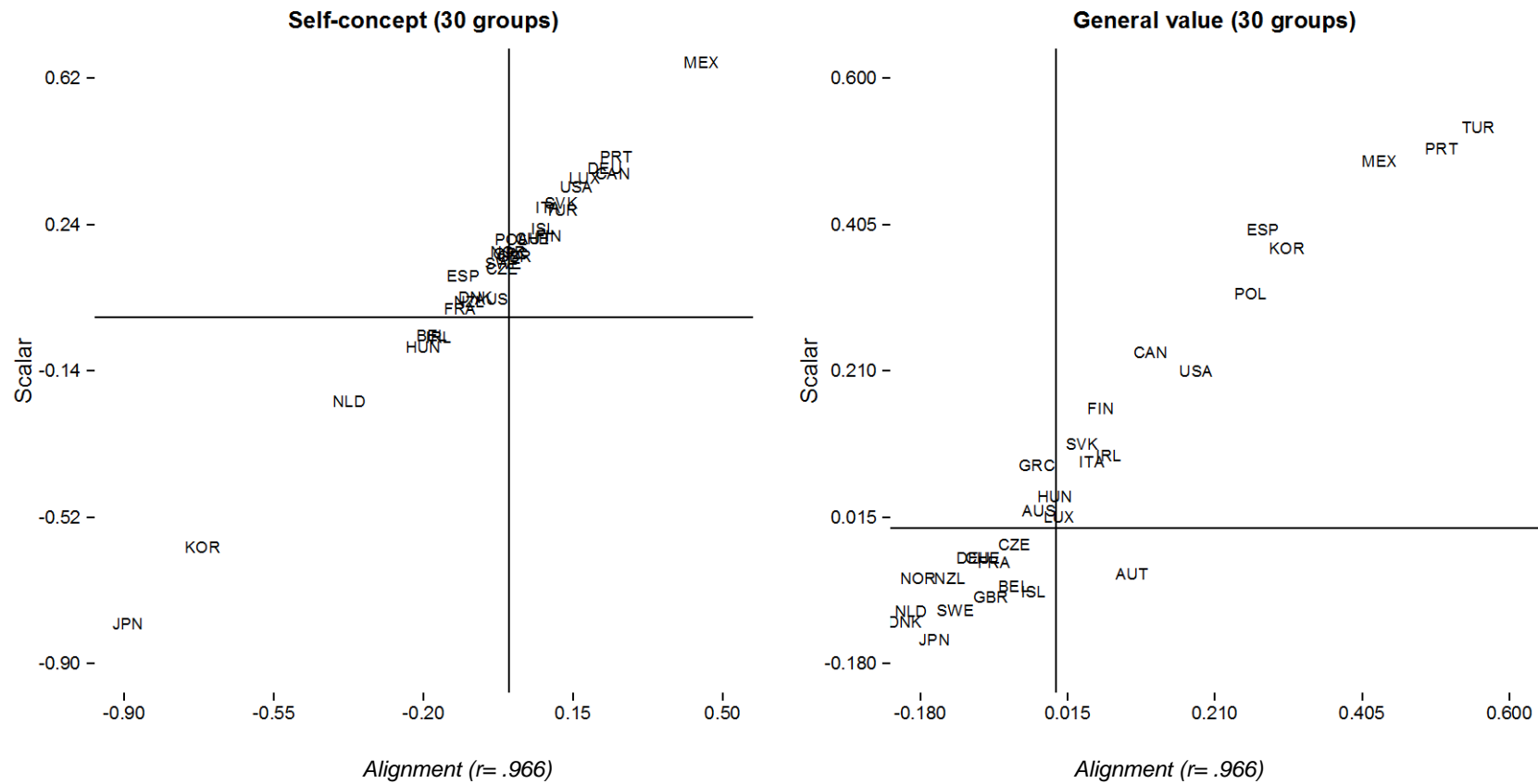
Figure 1. Factor means of self-concept and general value of science for 30 and 60 groups: alignment method and scalar model.
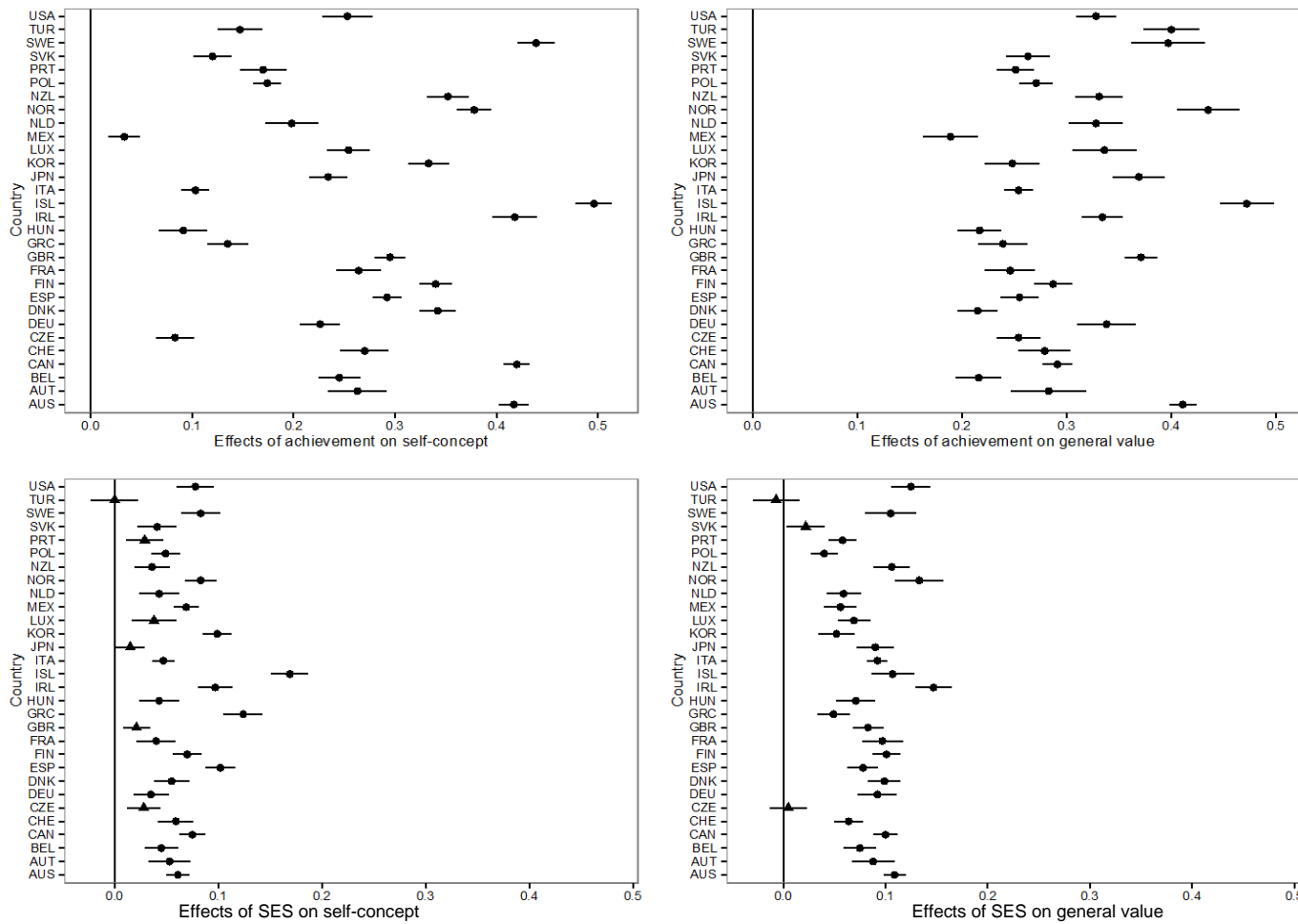
Figure 2. The effects of science achievement and SES on general value and self-concept based on 30 groups MIMIC model with alignment. *Note*. Circles indicate statistical significant (p < .001), whereas triangle indicates coefficients are not significant at the .001 level of confidence. The bar indicates +1/-1 standard error.
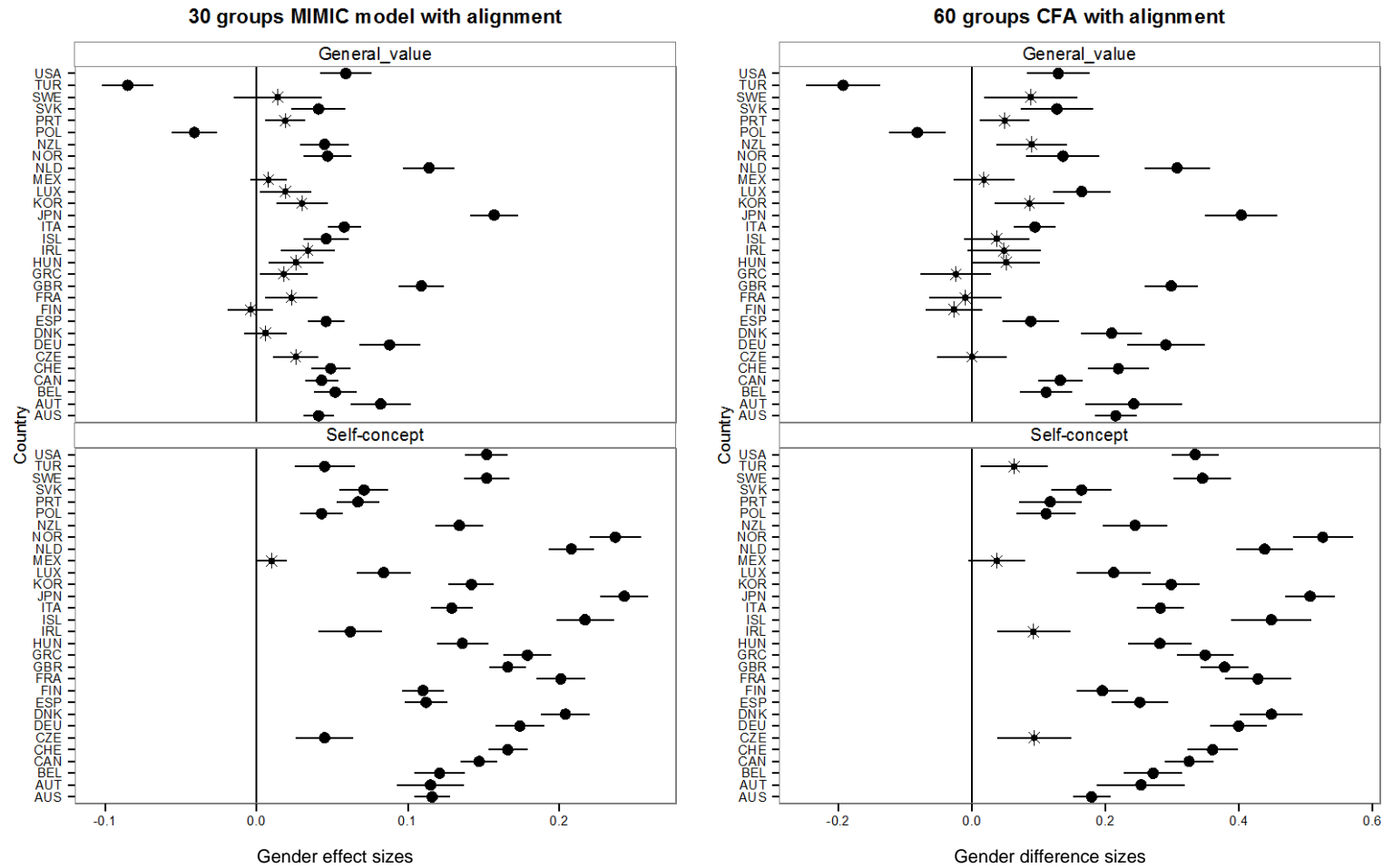
Figure 3. Patterns of gender differences: general value and self-concept based on two different models.
Note. Large circles indicate statistically significant coefficients (p < .001), whereas small circles (☼)indicate coefficients are not significant at p < .001. The bars indicates +1/-1 standard error.
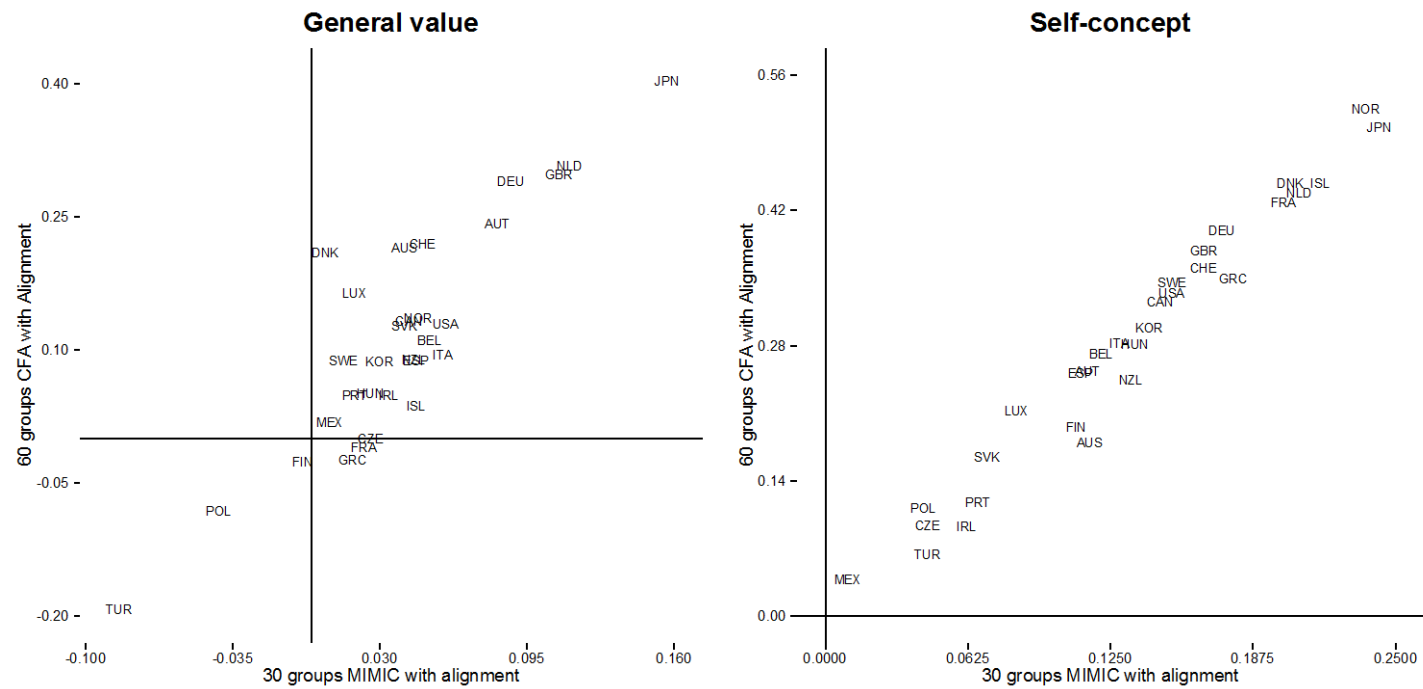
Figure 4. Patterns of gender differences: general value and self-concept, based on two different models.

Supplemental Materials

Table of Contents

Appendix 1:

Quality of PISA Scales

Much effort went into the development of internationally agreed-upon and comparable scales in the PISA data collections. The PISA assessment framework and items were developed through discussions between international groups of substantive and psychometric experts, followed by rigorous translation, verification, and national adaptions, which were implemented to balance the comparability and ecological validity of measures. Lastly, these measures were piloted and field tested, necessary changes were made, and the surveys were conducted with standardized administration procedures. In an evaluation of earlier PISA scales, where this development was described in more detail, Marsh, Hau, et al. (2006) concluded that the PISA scales were the strongest measure of educational psychology's most useful constructs and demonstrated good support for their psychometric properties, for metric – but not scalar – invariance across 25 countries, and cross-cultural validity in relation to consistent support for convergent and discriminant validity across 25 countries. They argued that the scales were useful for a broad range of educational psychology research: as a set of outcome measures; as a powerful set of intervening variables that facilitates the attainment of many long-term, desirable educational outcomes, and as a basis for mapping other educational psychology constructs in relation to their convergence and divergence with the PISA constructs. Similar to results by Marsh, Hau, et al. (2006) as well as Nagengast and Marsh (2013), the PISA results based on the present investigation show good support for the configural model and even the metric model. We also note that in cross-cultural research (as in other areas also) there are few if any published large-scale studies (i.e., many groups, factors, and items) that support scalar invariance (see earlier discussion).

Appendix 2:

Guidelines for Configuring the Alignment-With-CFA (AwC) Approach

The AwC (Alignment-with-CFA) approach can be amended for greater precision. Its model is estimated according to the following steps:

(1) Testing measurement invariance in relation to factor loadings and item intercepts is a precondition for comparing latent factor means across multiple groups. If the scalar model provides good model fit and support for item intercept invariance, alignment or AwC models should not be pursued, and the more parsimonious scalar model can be used for latent mean comparisons.

(2) If item intercept invariance (or even factor loadings) is not supported, and the scalar model provides poor model fit, an alignment analysis should be employed for latent mean comparisons.

(3) If there is a need to conduct an additional analysis that cannot be easily implemented within the alignment framework, but can be estimated with CFA models, then all parameter estimates from the alignment solution should be used as starting values to estimate the AwC model.

(4) Since a total of $2m$ constraints need to be added for the AwC model to be identified, selected parameter estimates are fixed to the values obtained from the alignment solution:

(i)   The m factor variances and m factor means are freely estimated. Then, for purposes of identification, a referent indicator (e.g., the first indicator) is selected for each factor, and the factor loading and intercept of this indicator are fixed to its estimated values from the alignment solution (i.e., these values are not allowed to be freely estimated).

(ii)   All other parameter estimates are free in AwC, the same as in the alignment solution.

(iii) The AwC solution will have the same degrees of freedom, the same chi-square, and goodness of fit statistics as the configural CFA model. However, the AwC solution will have the same parameter estimates as the alignment solution. Standard errors will also be highly similar, but might be slightly inflated, suggesting that caution still needs to be exerted in the interpretation of marginally non-significant results. In this sense, the AwC model is equivalent to the alignment solution. Importantly, the researcher has more flexibility in terms of how to constrain or further modify the AwC model (as it is a true CFA model) than with the alignment model upon which it is based.

Appendix 3:

Annotated Mplus Input and Output for the Alignment and Alignment-Within-CFA (AwC) Approaches

TITLE: The alignment model with 30 groups (Model MG4$_{AL}$ see Table 2);

DATA: FILE = "PISA06_data.dat";
VARIABLE:
NAMES = SCHOOLID STIDSTD
COUNTRY OECD W_FSTUWT
ST16Q01 ST16Q02 ST16Q03 ST16Q04
ST37Q06 Gender SES PV1SCIE;
MISSING=.;
USEVARIABLES ARE ST16Q01-
ST19Q06 Gender SES PV1SCIE
CONTSCHL;
CLUSTER = CONTSCHL;
`! cluster by school;`

WEIGHT = W_FSTUWT;
`! W_FSTUWT is the student-`
`level weighting variable in`
`the PISA database.`

classes = c(30);
knownclass = c(COUNTRY=36 40 56 124
203 208 246 250 276 300 348 352 372 380
392 410 442 484 528 554 578 616 620 703
724 752 756 792 826 840);
`!Define the 30 multiple`
`groups (countries)`

DEFINE: CONTSCHL =
(country*10000) + SCHOOLID;
`! Define group to be a`
`unique combination of`
`country (country ID code`
`multiplied by 1000) and`
`school ID;`

ANALYSIS: TYPE=MIXTURE
COMPLEX;
`!Complex: Analysis takes`
`nesting of students with`
`schools into account;`
           ESTIMATOR = ML;
           PROCESSORS = 6;

TITLE: Model MG-MIMIC$_{AwC}$ (see Figure 2) 30 groups MIMIC model with the alignment model;

DATA: FILE = "PISA06_data.dat";
VARIABLE:
NAMES = SCHOOLID STIDSTD
COUNTRY OECD W_FSTUWT
ST16Q01 ST16Q02 ST16Q03 ST16Q04
ST37Q06 Gender SES PV1SCIE;
MISSING=.;
USEVARIABLES ARE ST16Q01-
ST19Q06 Gender SES PV1SCIE
CONTSCHL;
CLUSTER = CONTSCHL;
`! cluster by school;`

WEIGHT = W_FSTUWT;
`! W_FSTUWT is the student-`
`level weighting variable in`
`the PISA database.`

GROUPING = country (36 40 56 124 203
208 246 250 276 300 348 352 372 380 392
410 442 484 528 554 578 616 620 703 724
752 756 792 826 840);
`!Define the 30 multiple`
`groups (countries)`

DEFINE: CONTSCHL =
(country*10000) + SCHOOLID;
`! Define group to be a`
`unique combination of`
`country (country ID code`
`multiplied by 1000) and`
`school ID;`

ANALYSIS: TYPE= COMPLEX;
`!Complex: Analysis takes`
`nesting of students with`
`schools into account;`
           ESTIMATOR = MLR;
           PROCESSORS = 6;

MODEL:

joyscie BY ST16Q01 ST16Q02 ST16Q03
ST16Q04 ST16Q05;
instscie BY ST35Q01 ST35Q02 ST35Q03
ST35Q04 ST35Q05;
scifut   BY ST29Q01 ST29Q02 ST29Q03
ST29Q04;
scieeff BY ST17Q01 ST17Q02 ST17Q03
ST17Q04
        ST17Q05 ST17Q06 ST17Q07
ST17Q08;
scscie   BY ST37Q01 ST37Q02 ST37Q03
ST37Q04 ST37Q05 ST37Q06;
genscie BY ST18Q01 ST18Q02
ST18Q04 ST18Q06 ST18Q09;
perscie BY ST18Q03 ST18Q05 ST18Q07
ST18Q08 ST18Q10;
sciact   BY ST19Q01 ST19Q02 ST19Q03
ST19Q04 ST19Q05 ST19Q06;

output:
tech1 tech8 align SVALUES;
! SVALUES: ask for starting
values for subsequent AwC
model.

<<< Here an example of
starting values for the USA
group in the output of the
alignment model above. >>

%C#30%

   joyscie BY st16q01*0.73192;

---

   joyscie BY st16q02*0.71737;
MODEL:

joyscie BY ST16Q01 ST16Q02 ST16Q03
ST16Q04 ST16Q05;
instscie BY ST35Q01 ST35Q02 ST35Q03
ST35Q04 ST35Q05;
scifut   BY ST29Q01 ST29Q02 ST29Q03
ST29Q04;
scieeff BY ST17Q01 ST17Q02 ST17Q03
ST17Q04
        ST17Q05 ST17Q06 ST17Q07
ST17Q08;
scscie   BY ST37Q01 ST37Q02 ST37Q03
ST37Q04 ST37Q05 ST37Q06;
genscie BY ST18Q01 ST18Q02
ST18Q04 ST18Q06 ST18Q09;
perscie BY ST18Q03 ST18Q05 ST18Q07
ST18Q08 ST18Q10;
sciact   BY ST19Q01 ST19Q02 ST19Q03
ST19Q04 ST19Q05 ST19Q06;

! It is easy to incorporate
 covariates in AwC model.
(e.g., add
scscie-sciact on Gender SES
PV1SCIE
for the MIMIC model (MG-
MIMIC4$_{AwC}$)

! for all parameters, the
exact values from an
alignment model including
all the eight constructs (44
items) as starts values
(using *)

! For identification
purposes, the first item per
factor is constrained to its
estimated values from the
alignment solution, and
factor variances and means
are free
<<< Model specifications are
shown for the USA groups. All
other groups are defined in a
similar manner>>
Model 840:
  joyscie BY st16q01@0.73192;
  joyscie BY st16q02*0.71737;

```
                                             st19q03*0.25967;
joyscie BY st16q03*0.68005;          joyscie BY st16q03*0.68005;
joyscie BY st16q04*0.73858;          joyscie BY st16q04*0.73858;
joyscie BY st16q05*0.78418;          joyscie BY st16q05*0.78418;
…..                                     …..
…..                                     …..
sciact BY st19q01*0.44008;           sciact BY st19q01@0.44008;
sciact BY st19q02*0.46243;           sciact BY st19q02*0.46243;
sciact BY st19q03*0.53099;           sciact BY st19q03*0.53099;
sciact BY st19q04*0.32932;           sciact BY st19q04*0.32932;
sciact BY st19q05*0.54562;           sciact BY st19q05*0.54562;
sciact BY st19q06*0.18266;           sciact BY st19q06*0.18266;
instscie WITH joyscie*0.47751;       instscie WITH joyscie*0.47751;
…..                                     …..
…..                                     …..
sciact WITH scscie*0.47891;          sciact WITH scscie*0.47891;
sciact WITH genscie*0.43133;         sciact WITH genscie*0.43133;
sciact WITH perscie*0.51505;         sciact WITH perscie*0.51505;

[ st16q01*2.63270 ];                 [ st16q01@2.63270 ];
[ st16q02*2.41076 ];                 [ st16q02*2.41076 ];
[ st16q03*2.33300 ];                 [ st16q03*2.33300 ];
[ st16q04*2.72070 ];                 [ st16q04*2.72070 ];
[ st16q05*2.70186 ];                 [ st16q05*2.70186 ];
…..                                     …..
…..                                     …..
[ st19q01*1.62389 ];                 [ st19q01@1.62389 ];
[ st19q02*1.12605 ];                 [ st19q02*1.12605 ];
[ st19q03*1.29471 ];                 [ st19q03*1.29471 ];
[ st19q04*1.06963 ];                 [ st19q04*1.06963 ];
[ st19q05*1.33893 ];                 [ st19q05*1.33893 ];
[ st19q06*1.02782 ];                 [ st19q06*1.02782 ];
[ joyscie*0.04410 ];                 [ joyscie*0.04410 ];
[ instscie*0.24782 ];                [ instscie*0.24782 ];
[ scifut*0.32264 ];                  [ scifut*0.32264 ];
[ scieeff*0.09824 ];                 [ scieeff*0.09824 ];
[ scscie*0.19773 ];                  [ scscie*0.19773 ];
[ genscie*0.20793 ];                 [ genscie*0.20793 ];
[ perscie*0.25161 ];                 [ perscie*0.25161 ];
[ sciact*0.75671 ];                  [ sciact*0.75671 ];

st16q01*0.17500;                     st16q01*0.17500;
st16q02*0.19479;                     st16q02*0.19479;
st16q03*0.22249;                     st16q03*0.22249;
st16q04*0.16762;                     st16q04*0.16762;
st16q05*0.16281;                     st16q05*0.16281;
…..                                     …..
…..                                     …..
st19q01*0.44733;                     st19q01*0.44733;
st19q02*0.19859;                     st19q02*0.19859;
```

```
    st19q03*0.25967;
    st19q04*0.24697;
    st19q05*0.31817;
    st19q06*0.23871;


  joyscie*0.93175;
  instscie*0.70558;
  scifut*0.86945;
  scieeff*1.01327;
  scscie*0.96378;
  genscie*1.08514;
  perscie*0.78113;
  sciact*1.18903;
```

```
<<Mplus output>>
```
MODEL RESULTS:
```
!unstandardized results for
the alignment model above.
```

JOYSCIE BY

ST16Q01	0.732	0.008	90.750	0.000
ST16Q02	0.717	0.009	79.000	0.000
ST16Q03	0.680	0.008	80.764	0.000
ST16Q04	0.739	0.008	94.243	0.000
ST16Q05	0.784	0.008	102.351	0.000

…..
…..

SCIACT BY

ST19Q01	0.440	0.016	26.978	0.000
ST19Q02	0.462	0.015	31.327	0.000
ST19Q03	0.531	0.024	22.318	0.000
ST19Q04	0.329	0.010	32.620	0.000
ST19Q05	0.546	0.019	28.814	0.000
ST19Q06	0.183	0.012	14.648	0.000

INSTSCIE WITH

JOYSCIE	0.478	0.017	28.129	0.000

…..
…..

SCIACT WITH

JOYSCIE	0.654	0.026	24.893	0.000
INSTSCIE	0.389	0.021	18.376	0.000
SCIFUT	0.545	0.023	23.308	0.000
SCIEEFF	0.474	0.026	18.443	0.000
SCSCIE	0.479	0.025	19.093	0.000
GENSCIE	0.431	0.024	17.732	0.000
PERSCIE	0.515	0.023	22.134	0.000

Means

JOYSCIE	0.044	0.027	1.646	0.100
INSTSCIE	0.248	0.023	10.804	0.000
SCIFUT	0.323	0.022	14.655	0.000
SCIEEFF	0.098	0.037	2.652	0.008

SCSCIE	0.198	0.039	5.054	0.000

```
st19q04*0.24697;
  st19q05*0.31817;
  st19q06*0.23871;


  joyscie*0.93175;
  instscie*0.70558;
  scifut*0.86945;
  scieeff*1.01327;
  scscie*0.96378;
  genscie*1.08514;
  perscie*0.78113;
  sciact*1.18903;
```

```
<<Mplus output>>
```
MODEL RESULTS:
```
!unstandardized results for
the AwC model above.
```

JOYSCIE BY

ST16Q01	0.732	0.000	999.00	999.00
ST16Q02	0.717	0.008	84.657	0.000
ST16Q03	0.680	0.009	75.183	0.000
ST16Q04	0.739	0.009	83.899	0.000
ST16Q05	0.784	0.008	95.966	0.000

…..
…..

SCIACT BY

ST19Q01	0.440	0.000	999.00	999.00
ST19Q02	0.462	0.013	34.882	0.000
ST19Q03	0.531	0.019	27.514	0.000
ST19Q04	0.329	0.014	24.028	0.000
ST19Q05	0.546	0.015	36.337	0.000
ST19Q06	0.183	0.014	12.637	0.000

INSTSCIE WITH

JOYSCIE	0.478	0.017	27.613	0.000

…..
…..

SCIACT WITH

JOYSCIE	0.654	0.024	27.053	0.000
INSTSCIE	0.389	0.020	19.881	0.000
SCIFUT	0.545	0.021	26.262	0.000
SCIEEFF	0.474	0.026	18.162	0.000
SCSCIE	0.479	0.022	21.776	0.000
GENSCI	0.431	0.025	17.379	0.000
PERSCIE	0.515	0.026	19.866	0.000

Means

JOYSCIE	0.044	0.023	1.941	0.052
INSTSCIE	0.248	0.017	14.483	0.000
SCIFUT	0.323	0.014	23.811	0.000
SCIEEFF	0.098	0.030	3.285	0.001

SCSCIE	0.198	0.025	7.926	0.000
GENSCIE	0.208	0.036	5.854	0.000
PERSCIE	0.252	0.028	8.836	0.000
SCIACT	0.757	0.036	20.759	0.000

Intercepts

ST16Q01	2.633	0.000	999.00	999.00
ST16Q02	2.411	0.009	254.053	0.000
ST16Q03	2.333	0.015	153.671	0.000
ST16Q04	2.721	0.013	205.973	0.000
ST16Q05	2.702	0.016	172.449	0.000
.....				
.....				
ST19Q01	1.624	0.017	96.907	0.000
ST19Q02	1.126	0.013	85.931	0.000
ST19Q03	1.295	0.016	78.744	0.000
ST19Q04	1.070	0.009	115.829	0.000
ST19Q05	1.339	0.018	72.948	0.000
ST19Q06	1.028	0.013	78.160	0.000

Variances

JOYSCIE	0.932	0.025	36.572	0.000
INSTSCIE	0.706	0.023	31.103	0.000
SCIFUT	0.869	0.021	41.597	0.000
SCIEEFF	1.013	0.040	25.155	0.000
SCSCIE	0.964	0.038	25.124	0.000
GENSCIE	1.085	0.054	20.231	0.000
PERSCIE	0.781	0.031	25.446	0.000
SCIACT	1.189	0.085	13.936	0.000

Residual Variances

ST16Q01	0.175	0.006	27.283	0.000
ST16Q02	0.195	0.006	33.802	0.000
ST16Q03	0.222	0.007	33.757	0.000
ST16Q04	0.168	0.006	25.923	0.000
ST16Q05	0.163	0.008	21.616	0.000
.....				
.....				
ST19Q01	0.447	0.013	33.577	0.000
ST19Q02	0.199	0.010	19.488	0.000
ST19Q03	0.260	0.011	23.291	0.000
ST19Q04	0.247	0.013	19.398	0.000
ST19Q05	0.318	0.011	27.897	0.000
ST19Q06	0.239	0.022	10.982	0.000

SCIACT	0.757	0.043	17.650	0.000

Intercepts

ST16Q01	2.633	0.015	172.239	0.000
ST16Q02	2.411	0.016	153.681	0.000
ST16Q03	2.333	0.009	252.438	0.000
ST16Q04	2.721	0.009	293.153	0.000
ST16Q05	2.702	0.010	281.415	0.000
.....				
.....				
ST19Q01	1.624	0.000	999.00	999.00
ST19Q02	1.126	0.015	77.440	0.000
ST19Q03	1.295	0.019	69.870	0.000
ST19Q04	1.070	0.012	88.178	0.000
ST19Q05	1.339	0.018	74.436	0.000
ST19Q06	1.028	0.014	73.530	0.000

Variances

JOYSCIE	0.932	0.025	37.180	0.000
INSTSCIE	0.706	0.024	29.056	0.000
SCIFUT	0.869	0.019	46.779	0.000
SCIEEFF	1.013	0.055	18.373	0.000
SCSCIE	0.964	0.024	40.116	0.000
GENSCIE	1.085	0.059	18.468	0.000
PERSCIE	0.781	0.046	16.931	0.000
SCIACT	1.189	0.065	18.232	0.000

Residual Variances

ST16Q01	0.175	0.006	27.283	0.000
ST16Q02	0.195	0.006	33.802	0.000
ST16Q03	0.222	0.007	33.757	0.000
ST16Q04	0.168	0.006	25.923	0.000
ST16Q05	0.163	0.008	21.616	0.000
.....				
.....				
ST19Q01	0.447	0.013	33.577	0.000
ST19Q02	0.199	0.010	19.488	0.000
ST19Q03	0.260	0.011	23.291	0.000
ST19Q04	0.247	0.013	19.398	0.000
ST19Q05	0.318	0.011	27.897	0.000
ST19Q06	0.239	0.022	10.982	0.000

GENSCIE	0.208	0.044	4.710	0.000
PERSCIE	0.252	0.028	8.883	0.000

Appendix 4:

Correlations Among the Eight Motivational Constructs in PISA 2006

Not surprisingly, all 28 correlations among the eight motivational constructs were positive (Mean[M] $r = .547$) and, due in part to the large sample size, all were statistically significant. The largest correlation was between general value and personal value of science ($r = .785$). Correlations among enjoyment, instrumental motivation and future-oriented motivation in science were substantial ($r = .590$ to $.713$), and they were all highly correlated with personal value ($r = .666$ to $.705$). Correlations of self-concept in science with other motivational constructs were moderate ($r = .399$ to $.611$) and slightly larger than those for self-efficacy in science ($r = .370$ to $.518$). Correlations involving general value were comparatively smaller ($r = .386$ to $.511$), except for the aforementioned substantial correlations with personal value. Finally, engagement in extracurricular activities in science was highly correlated with enjoyment ($r = .639$) and personal value ($r = .592$), but less correlated with the other motivation constructs ($r = .452$ to $.569$).

With respect to correlations relating the motivational constructs to the three covariates, correlations of gender to the eight motivational variables were statistically significant but small ($r = .024$ to $.094$), and favored boys. This pattern of correlations was similar with those of SES to motivational constructs ($r = -.030$ to $.114$). However, science achievement was more highly correlated with motivational constructs ($r = .081$ to $.372$), with the exception that correlation between achievement and extracurricular activities was non-significant ($r = .012$). The strongest relations emerged with science self-efficacy ($r = .372$) and general value of science ($r = .262$). Students who had higher science achievement values tended to report that they would be able to solve a range of scientific problems and ascribe a higher societal value to science. Correlations of achievement to self-concept and general value were somewhat smaller (rs = $.153$ and $.149$ respectively).

Appendix 5:

Annotated Input Files Used in Study 2 (Simulated Data)

Title: Population Model - Input for the Data Generation (20% large non-invariance)

! In all input files, statements preceded by ! are annotations.

! The Monte Carlo facility is used to generate the data.

MONTECARLO:

NGROUPS = 15; *! This statement indicates the number of groups.*

NOBSERVATIONS = 15(100); *! This statement indicates the sample size in each group.*

NREPS = 500; *! 500 replications are requested.*

SEED = 4533; ! set seed

REPSAVE = ALL; !*saver all of the data sets generated in a Monte Carlo simulation study*

SAVE = cfa-G15-N100-20-*.dat; *! This statement identifies the data set to be created.*

! The following section defines the population model based on the parameters described in Table S2
*! and Figure 2. The * symbol precedes specific parameter values.*
! Factor loadings are noted with BY, regressions with ON, correlations with, means and
! intercepts are noted between brackets []; variances and residuals are noted without brackets.
ANALYSIS: ESTIMATOR = ML;
Model population: *!Group 1*
f by y1*1.00 y2*1.00 y3*1.40 y4*1.00 y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];

Title: Population Model - Input for the Data Generation (10% large non-invariance)

MONTECARLO:

NGROUPS = 15; *! This statement indicates the number of groups.*

NOBSERVATIONS = 15(100); *! This statement indicates the sample size in each group.*

NREPS = 500; *! 500 replications are requested.*

SEED = 4533; ! set seed

REPSAVE = ALL; !*saver all of the data sets generated in a Monte Carlo simulation study*

SAVE = cfa-G15-N100-10-*.dat; *! This statement identifies the data set to be created.*

! The following section defines the population model based on the parameters described in Table S2
*! and Figure 2. The * symbol precedes specific parameter values.*
! Factor loadings are noted with BY, regressions with ON, correlations with, means and
! intercepts are noted between brackets []; variances and residuals are noted without brackets.
ANALYSIS: ESTIMATOR = ML;

Model population: *!Group 1*
f by y1*1.00 y2*1.00 y3*1.00 y4*1.00 y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];
ODEL POPULATION-g2:

f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;
[y1*-.50 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g3:
f by y1*0.95 y2*0.90 y3*1.10 y4*0.30
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];
MODEL POPULATION-g4:
f by y1*1.00 y2*1.00 y3*1.40 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g5:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;
[y1*-.50 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g6:
f by y1*0.95 y2*0.90 y3*1.10 y4*0.30
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];
MODEL POPULATION-g7:
f by y1*1.00 y2*1.00 y3*1.40 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g8:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;
[y1*-.50 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];

MODEL POPULATION-g2:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;

[y1*0.05 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g3:
f by y1*0.95 y2*0.90 y3*1.10 y4*1.05
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];
MODEL POPULATION-g4:
f by y1*1.00 y2*1.00 y3*1.40 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.00];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g5:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.95;
[y1*-.50 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g6:
f by y1*0.95 y2*0.90 y3*1.10 y4*0.30
y5*1.05;
[y1*-.05 y2*-.05 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];
MODEL POPULATION-g7:
f by y1*1.00 y2*1.00 y3*1.00 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g8:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;
[y1*0.05 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];

MODEL POPULATION-g9:
f by y1*0.95 y2*0.90 y3*1.10 y4*0.30
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;

f*1.2; [f*1];
MODEL POPULATION-g10:
f by y1*1.00 y2*1.00 y3*1.40 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g11:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;
[y1*-.50 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g12:
f by y1*0.95 y2*0.90 y3*1.10 y4*0.30
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];
MODEL POPULATION-g13:
f by y1*1.00 y2*1.00 y3*1.40 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g14:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;
[y1*-.50 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g15:
f by y1*0.95 y2*0.90 y3*1.10 y4*0.30
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];

OUTPUT: tech9;
MODEL POPULATION-g9:
f by y1*0.95 y2*0.90 y3*1.10 y4*1.05
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];
MODEL POPULATION-g10:

f by y1*1.00 y2*1.00 y3*1.40 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.00];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g11:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.95;
[y1*-.50 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g12:
f by y1*0.95 y2*0.90 y3*1.10 y4*0.30
y5*1.05;
[y1*-.05 y2*-.05 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];
MODEL POPULATION-g13:
f by y1*1.00 y2*1.00 y3*1.00 y4*1.00
y5*1.00;
[y1*0.00 y2*0.00 y3*0.00 y4*0.00 y5*0.50];
y1-y5*1;
f*1; [f*0];
MODEL POPULATION-g14:
f by y1*1.05 y2*1.10 y3*0.90 y4*0.95
y5*0.50;
[y1*0.05 y2*0.05 y3*-.10 y4*0.10 y5*-.05];
y1-y5*1;
f*1.5; [f*.3];
MODEL POPULATION-g15:
f by y1*0.95 y2*0.90 y3*1.10 y4*1.05
y5*1.05;
[y1*-.05 y2*0.50 y3*0.10 y4*-.05 y5*0.05];
y1-y5*1;
f*1.2; [f*1];

OUTPUT: tech9;

TITLE: Alignment Model
! The following statement is used to identify the data file.
 DATA: FILE = cfa-G15-N100-20-22.dat ; *Here, the data file is labeled* cfa-G15-N100-20-
        199.dat, which is the 199th replication for 100 of sample size with 20% large non-
        invariance

 VARIABLE: NAMES = y1-y5 GROUP;
USEVARIABLES = y1-y5;
Classes = c(15); ! *number of groups*
knownclass = c(GROUP = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15);

ANALYSIS:
TYPE = MIXTURE;
ESTIMATOR = MLR;
alignment = fixed;
ALGORITHM = INTEGRATION;
PROCESSORS = 2;

MODEL: %OVERALL%
F by Y1-Y5;
! Specific sections of output are requested.
OUTPUT: SVALUES align;

Title: ICM-CFA (Configural, metric, and scalar invariance model)
! The following statement is used to identify the data file.
DATA: FILE = cfa-G15-N100-20-199.dat ; *Here, the data file is labeled* cfa-G15-N100-20-
        199.dat, which is the 199th replication for 100 of sample size with 20% large non-
        invariance

 VARIABLE: NAMES = y1-y5 GROUP;
USEVARIABLES = y1-y5;
GROUPING = GROUP (1 2 3 4 5 6 7 8 9 10 11 12 13 14 15);

ANALYSIS:
ESTIMATOR = ML; *!Maximum Likelihood (ML) estimation is used.*
Model = configural metric, and scalar;
MODEL:
f by y1* y2 y3 y4 y5; ! freely estimate the first factor loading
[y1 y2 y3 y4 y5];
f@1; [f@0];! *For identification purposes, factor variance and latent mean are fixed to be 1 and
        0 in the first group.*
! Specific sections of output are requested.
Output: sampstat standardized SVALUES stdyx tech4;

Title: ICM-CFA (Partial invariance model)
! The following statement is used to identify the data file.
DATA: FILE = cfa-G15-N100-20-199.dat ; *Here, the data file is labeled* cfa-G15-N100-20-199.dat, which is the 199th replication for 100 of sample size with 20% large non-invariance
VARIABLE: NAMES = y1-y5 GROUP;
USEVARIABLES = y1-y5;
GROUPING = GROUP (1 2 3 4 5 6 7 8 9 10 11 12 13 14 15);
ANALYSIS:
ESTIMATOR = ML;
MODEL:
f by y1* y2 y3 y4 y5;
[y1 y2 y3 y4 y5];
y1-y5*1;
f@1;
[f@0];
MODEL 1:
 [Y5]; [Y1]; *! freely estimate intercept Y1 and Y5 according to the strategy of partial invariance described in the main text*
f@1; [f@0]; *For identification purposes, factor variance and latent mean are fixed to be 1 and 0 in the first group.*
MODEL 2:
f; [f];*! Freely estimate factor variance and latent mean from Group 2 to Group 15*
MODEL 3:
F BY Y4;
f; [f];
MODEL 4:
[Y2];
[Y4];
[Y3];
F BY Y2;
f; [f];
MODEL 5:
F BY Y5;
[Y1];
F BY Y3;
f; [f];
MODEL 6:
F BY Y4;
f; [f];
MODEL 7:
[Y2];
[Y4];

f; [f];
MODEL 8:
F BY Y5;
f; [f];
MODEL 9:
F BY Y4;
f; [f];
MODEL 10:
[Y2];
[Y4];
f; [f];
MODEL 11:
F BY Y5;
F BY Y3;
f; [f];
MODEL 12:
F BY Y4;
f; [f];
MODEL 13:
[Y2];
[Y4];
[Y3];
F BY Y3;
f; [f];
MODEL 14:
f; [f];
MODEL 15:
F BY Y4;
f; [f];
OUTPUT: MODINDICES(5) tech9
SVALUES;

The previous population model is used to generate Validation data, except for changing the seed.
Title: Cross-validation Alignment Model
DATA: FILE = cfa-G15-N100-20-V199.dat ; *Here, the data file is labeled* cfa-G15-N100-20-V199.dat, which is the 199th replication of the validation data for 100 of sample size with 20% large non-invariance
VARIABLE: NAMES = y1-y5 GROUP;
USEVARIABLES = y1-y5;
GROUPING = GROUP (1 2 3 4 5 6 7 8 9 10 11 12 13 14 15);
ANALYSIS: ESTIMATOR = ML;
! The previous Alignment model is re-expressed using CFA based on validation data.
! The model section uses the exact values of the non-standardized loadings, intercepts, residual
! variances, and factor variances estimated from the previous model (using @).
!Only latent mean are freely estimated.
!But for identification purposes, latent mean are fixed to be 0 in the first group.

MODEL:
f by y1* y2-y5;
f@1;
MODEL 1:
F BY Y1@1.135;
F BY Y2@0.941;
F BY Y3@1.398;
F BY Y4@0.908;
F BY Y5@1.092;
[F@0];
[Y1@0.124];
[Y2@0.183];
[Y3@0.113];
[Y4@0.1];
[Y5@0.598];
F@1;
Y1@0.914;
Y2@1.249;
Y3@0.916;
Y4@1.171;
Y5@0.988;

MODEL 2:
F BY Y1@1.108;
F BY Y2@1.111;
F BY Y3@1.086;
F BY Y4@0.993;
F BY Y5@0.66;
[F]; *! freely estimate latent mean*
[Y1@-0.353];
[Y2@0.442];

[Y3@0.044];
[Y4@0.355];
[Y5@0.098];
F@0.925;
Y1@0.997;
Y2@1.106;
Y3@0.895;
Y4@0.966;
Y5@1.17;
………..
………..
…………
MODEL 15:
 F BY Y1@1.192;
 F BY Y2@1.073;
 F BY Y3@1.104;
 F BY Y4@0.498;
 F BY Y5@1.032;
 [F]; *! freely estimate latent mean*
 [Y1@-0.199];
 [Y2@0.463];
 [Y3@0.196];
 [Y4@-0.148];
 [Y5@0.048];
 F@0.632;
 Y1@0.716;
 Y2@0.82;
 Y3@1.251;
 Y4@0.915;
 Y5@1.223;

Title: Cross-validation Scalar Model

DATA: FILE = cfa-G15-N100-20-V199.dat ; *Here, the data file is labeled* cfa-G15-N100-20-V199.dat, which is the 199th replication of the validation data for 100 of sample size with 20% large non-invariance

VARIABLE: NAMES = y1-y5 GROUP;
USEVARIABLES = y1-y5;
GROUPING = GROUP (1 2 3 4 5 6 7 8 9 10 11 12 13 14 15);
ANALYSIS:
ESTIMATOR = ML;
! The previous scalar model is re-expressed using CFA based on validation data.
! The model section uses the exact values of the non-standardized loadings, intercepts, residual
! variances, and factor variances estimated from the previous model (using @).
!Only latent mean are freely estimated.
!But for identification purposes, latent mean are fixed to be 0 in the first group.

MODEL:
f by y1* y2-y5;
f@1;
MODEL 1:
F BY Y1@1.215;
F BY Y2@1.22;
F BY Y3@1.285;
F BY Y4@0.684;
F BY Y5@0.905;
[F@0];
[Y1@0.013];
[Y2@0.396];
[Y3@0.232];
[Y4@0.049];
[Y5@0.385];
F@1;
Y1@0.842;
Y2@1.25;
Y3@1.065;
Y4@1.265;
Y5@1.171;

MODEL 2:
[F]; *! freely estimate latent mean*
F@0.744;
Y1@1.011;
Y2@1.176;
Y3@0.882;
Y4@1.32;

Y5@1.193;
………..
………..
…………

MODEL 14:
[F]; *! freely estimate latent mean*
F@0.848;
Y1@1.081;
Y2@1.187;
Y3@0.998;
Y4@1.845;
Y5@1.069;

MODEL 15:
[F]; *! freely estimate latent mean*
F@0.532;
Y1@0.763;
Y2@0.847;
Y3@1.192;
Y4@0.964;
Y5@1.311;

Title: Cross-validation partial Model
DATA: FILE = cfa-G15-N100-20-V199.dat ; *Here, the data file is labeled* cfa-G15-N100-20-V199.dat, which is the 199th replication of the validation data for 100 of sample size with 20% large non-invariance
VARIABLE: NAMES = y1-y5 GROUP;
USEVARIABLES = y1-y5;
GROUPING = GROUP (1 2 3 4 5 6 7 8 9 10 11 12 13 14 15);
ANALYSIS: ESTIMATOR = ML;
! The previous partial solution specific to the 199th replication is re-expressed using CFA based on validation data.
! The model section uses the exact values of the non-standardized loadings, intercepts, residual
! variances, and factor variances estimated from the previous model (using @).
!Only latent mean are freely estimated.
!But for identification purposes, latent mean are fixed to be 0 in the first group.

MODEL:
f by y1* y2-y5;
f@1;
MODEL 1:
F BY Y1@1.167;
F BY Y2@1.12;
F BY Y3@1.209;
F BY Y4@1.024;
F BY Y5@1.012;
[F@0];
[Y1@-0.013];
[Y2@0.094];
[Y3@-0.212];
[Y4@0.013];
[Y5@0.479];
F@1;
Y1@0.875;
Y2@1.168;
Y3@1.167;
Y4@1.112;
Y5@1.062;

MODEL 2:
F BY Y1@1.167;
F BY Y2@1.12;
F BY Y3@1.209;
F BY Y4@1.024;
F BY Y5@1.012;
[F]; ! *freely estimate latent mean*
[Y1@-0.509];
[Y2@0.094];
[Y3@-0.212];
[Y4@0.013];
[Y5@-0.064];
F@0.748;
Y1@1.048;
Y2@1.197;
Y3@0.893;
Y4@0.988;
Y5@1.141;
………
………
………
MODEL 15:
F BY Y1@1.167;
F BY Y2@1.12;
F BY Y3@1.209;
F BY Y4@0.32;
F BY Y5@1.012;
[F]; ! *freely estimate latent mean*

[Y1@-0.509];
[Y2@0.094];
[Y3@-0.212];
[Y4@0.013];
[Y5@-0.064];
F@0.619;
Y1@0.745;
Y2@0.791;
Y3@1.234;
Y4@0.958;
Y5@1.25;

OUTPUT: MODINDICES(5) SVALUES;

Appendix 6:

Boxplots of Deviations Factor loading and Intercepts for Items from Selected PISA Scales Based on Alignment and scalar Models

Self-Concept: Factor Loadings   General Value: Factor Loadings
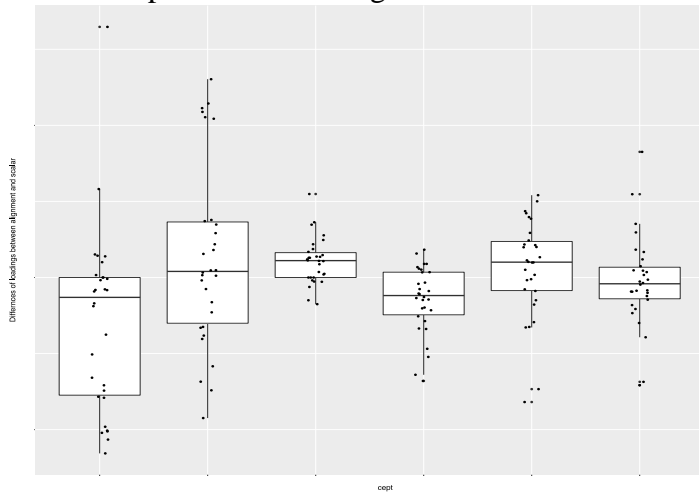
Self-Concept: Intercepts   General Value: Intercepts

Appendix 7:

Largest Modification Indices (MI) and Expected Parameter Change (EPC) for the Basic AwC

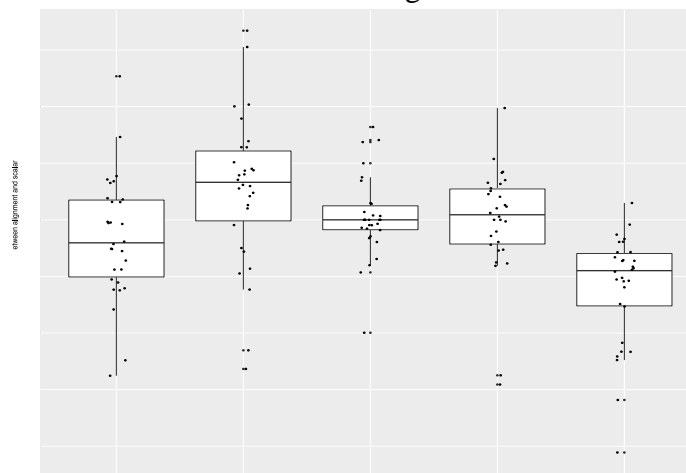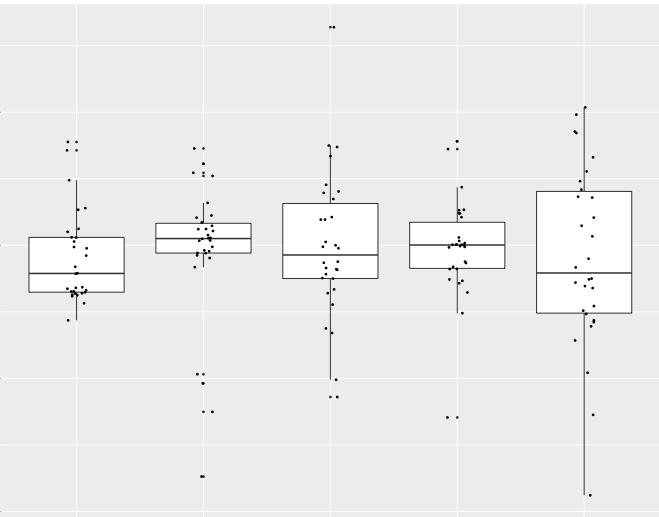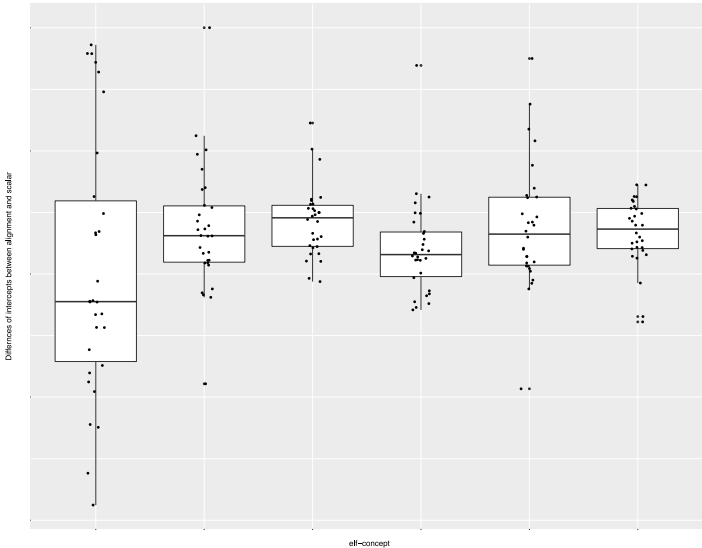| modV1	operator	modV2	MI	EPC	Std_EPC	StdYX_EPC	Group
[ST17Q04]			2928	0.312	0.312	0.356	484
[ST37Q01]			2887	0.244	0.244	0.33	484
JOYSCIE	BY	ST17Q04	1850	0.283	0.176	0.2	484
SCIFUT	BY	ST29Q04	1774	0.198	0.157	0.192	484
SCIACT	BY	ST19Q04	1740	0.137	0.173	0.194	484
[ST16Q01]			1674	-0.209	-0.209	-0.266	380
JOYSCIE	BY	ST37Q01	1659	0.214	0.132	0.179	484
PERSCIE	BY	ST17Q04	1640	0.282	0.176	0.2	484
[ST29Q02]			1609	-0.171	-0.171	-0.208	484
[ST16Q03]			1560	-0.242	-0.242	-0.277	724
SCIACT	BY	ST37Q01	1531	0.108	0.136	0.184	484
SCIACT	BY	ST17Q04	1505	0.134	0.169	0.193	484
PERSCIE	BY	ST37Q01	1475	0.212	0.133	0.18	484
JOYSCIE	BY	ST29Q04	1475	0.198	0.122	0.15	484
[ST16Q03]			1429	-0.177	-0.177	-0.248	484
[ST29Q04]			1401	0.169	0.169	0.207	484
[ST16Q05]			1396	0.193	0.193	0.222	724
PERSCIE	BY	ST29Q04	1388	0.201	0.126	0.154	484
[ST19Q06]			1383	0.207	0.207	0.336	724
SCIFUT	BY	ST37Q01	1383	0.185	0.147	0.199	484
[ST18Q01]			1364	0.193	0.193	0.311	724
SCIACT	BY	ST29Q04	1341	0.099	0.124	0.152	484
[ST19Q05]			1295	-0.255	-0.255	-0.342	826
JOYSCIE	BY	ST29Q02	1266	-0.17	-0.106	-0.128	484
INSTSCIE	BY	ST29Q04	1249	0.211	0.141	0.173	484
JOYSCIE	BY	ST19Q04	1232	0.246	0.152	0.171	484
SCIACT	BY	ST29Q02	1228	-0.088	-0.111	-0.134	484
PERSCIE	BY	ST29Q02	1222	-0.176	-0.11	-0.133	484
[ST19Q03]			1183	0.249	0.249	0.339	826
[ST19Q04]			1183	0.218	0.218	0.245	484
INSTSCIE	BY	ST37Q01	1175	0.208	0.139	0.188	484
[ST29Q02]			1151	0.145	0.145	0.145	124
INSTSCIE	BY	ST17Q04	1137	0.259	0.173	0.197	484
[ST18Q07]			1130	0.202	0.202	0.261	380
[ST18Q06]			1108	-0.196	-0.196	-0.283	724
SCIFUT	BY	ST17Q04	1105	0.209	0.166	0.189	484
SCIFUT	BY	ST19Q04	1079	0.219	0.174	0.195	484
SCIFUT	BY	ST29Q02	1054	-0.147	-0.117	-0.142	484
[ST17Q05]			1043	0.201	0.201	0.224	124
SCSCIE	BY	ST29Q04	1027	0.165	0.124	0.151	484
PERSCIE	BY	ST19Q04	992	0.232	0.145	0.163	484
INSTSCIE	BY	ST29Q02	970	-0.173	-0.116	-0.14	484
[ST18Q04]			964	0.349	0.349	0.519	410
[ST16Q03]			946	0.19	0.19	0.248	826
SCSCIE	BY	ST17Q04	932	0.203	0.153	0.174	484
[ST18Q02]			915	-0.385	-0.385	-0.513	410
[ST18Q03]			914	-0.186	-0.186	-0.249	380
[ST19Q05]			900	0.202	0.202	0.227	380
[ST16Q01]			860	0.109	0.109	0.175	484
SCSCIE	BY	ST29Q02	853	-0.14	-0.105	-0.127	484
[ST37Q02]			850	0.145	0.145	0.202	380
[ST19Q05]			810	-0.129	-0.129	-0.173	36
INSTSCIE	BY	ST19Q04	808	0.229	0.153	0.172	484
[ST16Q03]			791	0.144	0.144	0.191	380
[ST17Q01]			787	-0.195	-0.195	-0.225	724
GENSCIE	BY	ST17Q04	750	0.206	0.159	0.181	484
[ST17Q02]			743	-0.15	-0.15	-0.177	484
[ST19Q06]			740	0.487	0.487	0.504	300
[ST29Q02]			739	0.17	0.17	0.181	826
SCIACT	BY	ST16Q01	737	-0.123	-0.134	-0.17	380
[ST29Q03]			735	0.227	0.227	0.298	703
[ST16Q01]			728	0.179	0.179	0.204	756
[ST16Q02]			724	-0.158	-0.158	-0.202	826
GENSCIE	BY	ST29Q04	710	0.153	0.118	0.145	484
[ST18Q02]			684	0.188	0.188	0.28	756
[ST29Q03]			675	0.267	0.267	0.291	392
SCSCIE	BY	ST19Q04	674	0.182	0.137	0.153	484
JOYSCIE	BY	ST19Q06	673	-0.08	-0.079	-0.215	124
[ST16Q01]			667	-0.15	-0.15	-0.19	826
[ST37Q01]			665	-0.162	-0.162	-0.196	36
[ST19Q03]			656	0.135	0.135	0.188	36
[ST19Q01]			648	-0.15	-0.15	-0.199	724
SCIACT	BY	ST17Q02	645	-0.085	-0.107	-0.127	484
[ST29Q02]			639	0.13	0.13	0.129	724
GENSCIE	BY	ST29Q02	636	-0.135	-0.104	-0.126	484
[ST16Q05]			621	-0.106	-0.106	-0.153	484
[ST16Q03]			615	0.116	0.116	0.144	36
[ST37Q01]			610	0.142	0.142	0.166	724
[ST18Q09]			610	0.264	0.264	0.401	410
JOYSCIE	BY	ST17Q02	593	-0.155	-0.096	-0.113	484
[ST29Q04]			590	-0.121	-0.121	-0.124	124
SCIACT	BY	ST19Q06	579	0.313	0.413	0.428	300
SCIFUT	BY	ST17Q02	579	-0.146	-0.116	-0.137	484
PERSCIE	BY	ST19Q06	576	-0.078	-0.075	-0.205	124
JOYSCIE	BY	ST16Q03	555	-0.117	-0.073	-0.102	484
[ST17Q05]			554	-0.397	-0.397	-0.414	703
[ST18Q05]			553	0.274	0.274	0.366	410
[ST29Q03]			551	-0.1	-0.1	-0.113	36
[ST35Q02]			543	-0.149	-0.149	-0.167	826
SCIFUT	BY	ST29Q03	542	0.113	0.09	0.108	484
[ST19Q05]			536	0.402	0.402	0.417	352
[ST17Q04]			533	0.292	0.292	0.335	392
[ST18Q06]			528	0.108	0.108	0.165	124
[ST18Q05]			528	-0.112	-0.112	-0.159	484
[ST17Q01]			526	-0.296	-0.296	-0.376	578
[ST37Q05]			524	0.244	0.244	0.28	392
SCIACT	BY	ST19Q06	522	-0.062	-0.068	-0.186	124
[ST29Q03]			522	0.19	0.19	0.249	203
JOYSCIE	BY	ST18Q05	509	-0.125	-0.077	-0.11	484
GENSCIE	BY	ST37Q01	508	0.133	0.103	0.139	484
PERSCIE	BY	ST17Q02	507	-0.151	-0.094	-0.112	484
[ST29Q03]			507	-0.1	-0.1	-0.108	124
SCIFUT	BY	ST29Q01	505	-0.103	-0.082	-0.098	484
GENSCIE	BY	ST19Q06	502	-0.076	-0.073	-0.197	124
[ST29Q02]			501	0.075	0.075	0.079	36
[ST18Q07]			500	0.305	0.305	0.328	392
[ST37Q02]			495	-0.088	-0.088	-0.133	484
[ST17Q02]			491	0.196	0.196	0.226	756
[ST17Q08]			490	0.169	0.169	0.171	724
[ST29Q03]			485	0.103	0.103	0.124	484
[ST19Q02]			476	0.23	0.23	0.336	410
GENSCIE	BY	ST16Q03	464	-0.161	-0.139	-0.159	724
PERSCIE	BY	ST16Q03	457	-0.12	-0.075	-0.105	484
[ST17Q03]			454	0.361	0.361	0.368	250
SCIACT	BY	ST18Q05	454	-0.062	-0.078	-0.11	484
INSTSCIE	BY	ST29Q03	454	-0.126	-0.126	-0.143	36
SCIACT	BY	ST29Q03	452	0.058	0.074	0.089	484
[ST18Q06]			452	-0.326	-0.326	-0.369	40
[ST18Q02]			448	-0.234	-0.234	-0.311	392
[ST17Q07]			442	0.169	0.169	0.165	724
PERSCIE	BY	ST18Q03	441	0.115	0.072	0.104	484
[ST17Q08]			440	-0.138	-0.138	-0.152	36
[ST17Q04]			436	0.341	0.341	0.379	348
[ST18Q04]			434	-0.114	-0.114	-0.149	484
INSTSCIE	BY	ST29Q04	434	-0.119	-0.114	-0.118	124
SCSCIE	BY	ST18Q02	433	0.238	0.226	0.301	410
PERSCIE	BY	ST29Q03	431	0.115	0.072	0.087	484
[ST37Q01]			429	-0.172	-0.172	-0.228	392
[ST16Q01]			428	-0.23	-0.23	-0.256	372
[ST18Q09]			426	-0.288	-0.288	-0.382	208
JOYSCIE	BY	ST19Q06	424	-0.074	-0.074	-0.22	36
[ST17Q06]			421	-0.117	-0.117	-0.134	484
[ST37Q04]			421	0.094	0.094	0.114	724
SCSCIE	BY	ST19Q06	421	-0.059	-0.064	-0.175	124
SCIACT	BY	ST19Q06	419	0.186	0.208	0.266	616
PERSCIE	BY	ST16Q01	419	-0.161	-0.111	-0.141	380
SCIACT	BY	ST19Q06	417	0.121	0.124	0.202	724
[ST18Q09]			415	-0.337	-0.337	-0.381	352
INSTSCIE	BY	ST29Q02	415	0.1	0.096	0.096	124
[ST19Q03]			414	0.102	0.102	0.138	124
[ST17Q06]			413	0.256	0.256	0.311	616
SCIACT	BY	ST19Q06	411	0.247	0.267	0.359	348
SCIFUT	BY	ST18Q05	409	-0.106	-0.084	-0.119	484
[ST29Q03]			408	-0.11	-0.11	-0.134	826
INSTSCIE	BY	ST17Q02	408	-0.149	-0.1	-0.118	484
[ST18Q02]			408	-0.227	-0.227	-0.342	528
[ST17Q08]			405	0.272	0.272	0.311	578
SCIACT	BY	ST19Q06	404	-0.093	-0.099	-0.267	56
[ST16Q04]			401	0.239	0.239	0.25	410
PERSCIE	BY	ST18Q05	400	-0.109	-0.068	-0.097	484

Note. Modification Indices (MI) and Expected Parameter Change (EPC) are not available in the alignment approach, but can be obtained for the basic alignment-within-CFA (AwC) model. The 63 (out of 44 items x 30 groups = 1,320 factor loadings and 1,320 intercepts) StdYX_EPC values greater than .2 (in absolute value) are highlighted in yellow. ModV1 and modV2 = specific items associated with MI. Parameters refer to factor loadings (indicated by "BY") as an operator or as item intercepts (indicated by variable names in brackets). Group refers to the 30 different OECD countries.