Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 28 Number 13, December 2023 ISSN 1531-7714

A Tutorial on Cross Wave Measurement Invariance Testing with Item Factor Analysis

R. Noah Padgett, *Harvard University*

The consistency of psychometric properties across waves of data collection provides valuable evidence that scores can be interpreted consistently. Evidence supporting the consistency of psychometric properties can come from using a longitudinal extension of item factor analysis to account for the lack of independence of observation when evaluating the cross-wave properties. In this study, we demonstrate how to conduct and interpret a longitudinal examination of psychometric properties. This demonstration uses data on the Comprehensive Measure of Meaning across two waves of data collection. A simplified, structured set of **R** syntax for analyses is provided, and all remaining code is freely available in the accompanying Open Science Framework repository.

Keywords: measurement invariance, longitudinal, cohort studies, consistency

Introduction

The quality of any measurement tool in the social sciences relies on a set of scores that can be interpreted consistently. One aspect of scores having a consistent interpretation is whether the psychometric properties of the items are consistent over time without changes in the construct. The consistency of the interpretation of measure properties can be empirically evaluated to provide evidence that a researcher can measure the construct of interest even over relatively short time periods. The evaluation of repeated observations of a set of items to measure an underlying construct includes the added complexities of compromised independence of observation when testing measurement properties. This issue can be resolved by modeling the longitudinal observations; however, the issue is further complicated when the set of items uses categorical response options.

Categorical response scales such as *strongly disagree* to *strongly agree* are common in psycho-social measurement tools. A direct way of investigating test-retest reliability with such items is to compute the sum score, or total score, at each time point and estimate the correlation. However, sum scores make strong assumptions about all items being equally related to the domain (McNeish & Wolf, 2020; Widaman & Revelle, 2023), which may not hold or be tenable across time. The invariance of measurement properties is testable.

The tools needed to assess the measurement properties across time points are discussed in this paper. First, a brief introduction to factor analysis with categorical indicators is introduced. Then, the evaluation of measurement invariance of the factors across repeated measurements is discussed. Third, how the longitudinal factor model allows for evaluating the consistency of those measure properties. Then, a demonstration of invariance testing with data from a new measure of meaning-in-life is conducted. Lastly, a concise example write-up of the methods and results is given as part of the demonstration.

Item Factor Analysis

In educational and psychological research, a common approach to modeling the response to a

survey is factor analysis or item response theory (Brown, 2015; de Ayala, 2009; Wirth & Edwards, 2007). Factor analysis is frequently used in psycho-social scale development due to historical developments in psychology; however, the statistical foundations were built on the observed data being continuous. In psychological and educational measurement, truly continuous data are rare. Instead, data are commonly collected using discrete categories to capture information about a respondent, for example, using a Likert-type response format to assess attitudes towards a topic. A commonly used method for analyzing such data is to treat the observed data as representing a discretized underlying continuous response. The process by which this discretizing occurs is described in (Mislevy, 1986; Muthén, 1984).

Let an ordinal response variable (y) take on values $c = 1, 2, \cdots, C$, where C is the total number of response options. The responses to I such items are hypothesized to reflect m latent traits ($\boldsymbol{\eta}$). We want to relate $\boldsymbol{\eta}$ to y linearly; however, this is difficult due to the discrete nature of y. Instead, we presume that y is the observed manifestation of the categorization process

$$y_i = c, \text{ if } \tau_{c-1} < y_i^* \leq \tau_c, \quad (1)$$

where $\tau_0 = -\infty$, $\tau_c = \infty$, and y^* is the continuous latent response variable for item i. The threshold parameters (τ_c) may vary in magnitude and number across items/observed indicators. The linear relationship between y^* and η is now possible.

The relationship between the trait of interest, η, and the latent response variable, y^*, is modeled by the common factor model. The model for the vector of latent response $\boldsymbol{y}^*$ is

$$\boldsymbol{y}^* = \boldsymbol{\alpha} + \boldsymbol{\Lambda\eta} + \boldsymbol{\varepsilon} \quad (2)$$

$$\boldsymbol{\Sigma}(\boldsymbol{y}^*) = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta}, \quad (3)$$

where $\boldsymbol{\alpha}$ is the vector of latent response intercepts (typically assumed to be $\mathbf{0}$ within a factor analytic framework), $\boldsymbol{\Lambda}$ is the factor loading matrix, η is the latent trait (typically assumed that $\eta \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Phi})$), $\boldsymbol{\varepsilon}$ is the vector of residual (typically assumed $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Theta})$), and $\boldsymbol{\Sigma}(\boldsymbol{y}^*)$ is the model implied covariance matrix among the latent response variables. In applications, a common assumption is that the item responses are *locally* or *conditionally independent* given the latent trait $\boldsymbol{\eta}$. This results in a diagonal error-covariance matrix ($\boldsymbol{\Theta}$), where each item is statistically independent. Additionally, the latent response variable is commonly parameterized in terms of an intercept ($\boldsymbol{\alpha}$) and total variance or *scale* parameter (i.e., $\text{Var}(y_i^*) = \lambda_i^2\boldsymbol{\phi} + \theta_i$). Setting the latent response scale by fixing the scale parameter, or total variance of the latent response variable, is known as the DELTA parameterization. Alternatively, the THETA parameterization is specified in terms of the intercept and latent response residual variance ($\theta_i = \text{Var}(y^*) - \lambda^2\boldsymbol{\phi}$).

The above model is indeterminate with respect to location, scale, and orientation arising from $\boldsymbol{\eta}$ and $\boldsymbol{y}^*$ being unobserved. The indeterminacy can be resolved by restricting the parameter space of $\boldsymbol{\eta}$ and $\boldsymbol{y}^*$ according to the needs of the analysis (more on this in the measurement invariance section). Other restrictions can be made that allow for different interpretations of the model parameters. For example, the factor covariance matrix need not be diagonal or assume unit variances for the factors if the scale and orientation are set by restricting a factor loading to one for each factor. Kamata and Bauer (2008, Table 1, p. 139) described several approaches for resolving the indeterminacy in the item factor analysis model.

Once an approach to resolving the model indeterminacy has been decided, the latent factor and latent response variables can be used to compute the probability of the observed response using the categorization scheme described in Equation 1. The use of the threshold scheme implies the probability of an observed response for a single item is

$$Pr(y_i = c|\eta) = Pr(y_i^* \geq \tau_{c-1}|\eta) - Pr(y_i^* \geq \tau_c|\eta)$$
$$= F\left(\frac{\alpha_i + \lambda_i\eta - \tau_{c-1}}{\theta_i}\right) - F\left(\frac{\alpha_i + \lambda_i\eta - \tau_c}{\theta_i}\right) \quad (4)$$

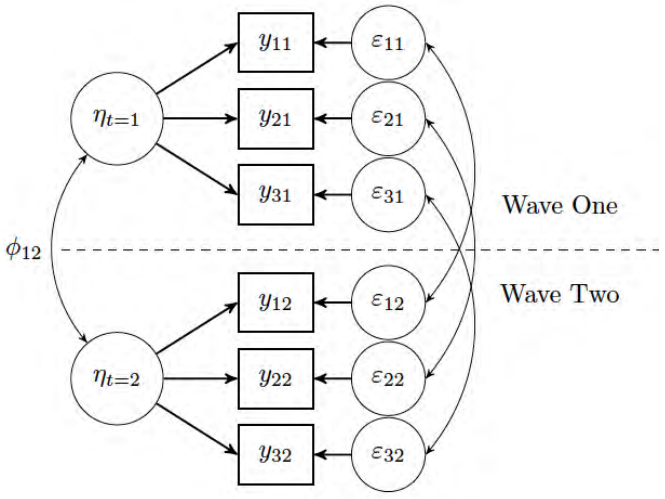
The link function (F) is commonly chosen to be the probit ($\Phi(\cdot)$) or logit ($\Psi(\cdot)$) link. The category probabilities can then be computed as the difference between the inequalities $Pr(y = 2) = Pr(y^* \geq \tau_2)$, $Pr(y = 1) = Pr(y^* \geq \tau_1) - Pr(y^* \geq \tau_2)$, and $Pr(y = 0) = 1 - Pr(y^* \geq \tau_1)$. This process is commonly used in item factor analysis (Wirth & Edwards, 2007).

Longitudinal Item Factor Model

Item factor analysis can be extended to model longitudinal relationships among domains. For instance, in a longitudinal model of a single domain measured by three items across two waves, the model

becomes a two-factor model with correlated residuals (see Figure 1). The unique component of this longitudinal model, relative to a multiple group factor model, is the correlation of residuals for each item pair is freely estimated between waves of data. Allowing for correlated residuals between item pairs across time helps account for the longitudinal aspect of the data. We expect the between-wave correlation of the same item to be high because the item content is the same. By specifying correlated residual variances, the known source of dependence among item responses is accounted for in our model (Brown, 2015; Liu et al., 2017).

Figure 1. Longitudinal item-factor model for cross-wave invariance



A major benefit of this longitudinal item factor model is examining how measurement properties change over time. Additionally, if the length of time is relatively short for the construct of interest, then the model may be useful for estimating test-retest reliability. The correlation of factors (ϕ_{12} in Figure 1) estimates test-retest reliability *if* length of time between waves is short, and the factors are comparable across waves. The comparability, or equality of measurement properties, across waves is necessary to ensure that test-retest reliability estimates are calibrated to the same factor across time. If factors are not comparable across waves, the correlation between factors will represent a correlation between some unclear common variance of a set of items and not the correlation between the same factors across time. The tenability of the equality of measurement properties can be evaluated by testing of *measurement invariance*. Testing measurement invariance for such models is the focus of this tutorial.

Longitudinal Measurement Invariance

Longitudinal measurement invariance investigations (Liu et al., 2017), with few time points, mirror closely with methods for multigroup invariance testing (Meredith & Horn, 2001; Millsap, 2011; Millsap & Cham, 2012). The approach for examining invariance follows closely with the recommendations of Wu and Estabrook (2016) for multigroup invariance testing with categorical indicators. The approach of invariance testing by Wu and Estabrook (2016) differs from the discussion of Liu et al. (2017) and Millsap (2011) in when item thresholds are evaluated. The difference has a practical implication for what models need to be specified in what order to evaluate the non-invariance of measurement properties across waves.

The issue is that evaluating measurement invariance with categorical indicators is less straightforward than with continuous outcomes. The difficulty arises due to the indeterminacy in the location and scale of the latent response variables. Wu and Estabrook (2016) identified the conditions in which latent responses (see the *Item Factor Analysis* section for more details) and factors in multigroup models are identified. The conditions Wu and Estabrook (2016) identified formed the basis for their recommendations for testing measurement invariance. The identification and invariance testing rules outlined in Wu and Estabrook (2016) are to test the 1) configural model, 2) threshold invariant model, 3) factor loading and threshold invariant model, (4) latent response intercept, factor loading, and threshold invariant model, and (5) latent response scale/residual variance, latent response intercept, factor loading, and threshold invariant model.

The important distinction of the order of models to test leads to the natural question of "why test threshold before factor loadings?" Wu and colleagues summarized their derivation of identification constraints in polytomous items as

> It is possible that adding one or multiple invariance conditions leads to an equivalent model. In this case, those invariance conditions cannot be tested because they can be satisfied trivially by a reparametrization. For example, not all invariance

> conditions impose restrictions on the baseline model. Models with invariant loadings, intercepts or unique variances alone are equivalent to the baseline model. Each of those invariance conditions cannot be tested alone. Once thresholds are assumed invariant, invariance of loadings, intercepts, and unique variances can be tested either separately or jointly. In light of this, we recommend investigating threshold invariance before considering other types of invariance. Once threshold invariance is established, the invariance of the remaining three types of parameters can be tested. (Wu & Estabrook, 2016, p. 1035)

A less technical reason that I believe summarizes their discussion is that thresholds for categorical items are a key feature that determines the location and scale of the latent response. The data do not fully identify the latent response variables, but, conditional on a level of the latent variable, the thresholds differentiate the probability of endorsing each response category. Evaluating whether the thresholds differentiate response propensity consistently across waves is a meaningful first step. Then, evaluating how other model parameters, such as factor loadings, differ across waves continues as normal. Starting with the average trend in response probability implied by thresholds allows for the more nuanced evaluation of loadings with more confidence. Additionally, starting with constraints on the thresholds makes the remaining constraints more easily explained to colleagues without a technical background but familiar with interpreting factor analysis. The summary explanation is not the most technical, but I believe having an easier explanation for colleagues is useful, especially considering the complexity of these models.

The technical issues of measurement invariance with ordered categorical variables above are joined with less statistical, but just as concerning, issues of the response process of individual respondents. For example, in the study of psychological constructs within a school setting, a limiting factor in the generalizability of results is that students commonly use environmental characteristics to make comparisons, also known as reference group bias (Duckworth & Yeager, 2015; Grutzmacher & Hartig, 2021). Responses to items will, therefore, be inherently connected to the specific sample used. Arguing that one's results generalize to other groups or that the local norm/reference is stable enough over time to make the case of the comparability of scores can be difficult. The potential difficulties of interpretation should not detract us from working to find evidence of the comparability of scores, and one such way of providing evidence in favor of the comparability of scores over time is to empirically assess the invariance of the statistical models used to make inferences.

A longitudinal investigation of measurement invariance adds to the complexity in that observations are necessarily correlated within the same person over time. The within-person correlation can be accounted for by allowing each item residual to freely correlate with itself over time (Liu et al., 2017). Adding a cross-wave covariance within an item adds a unique complexity to testing longitudinal measurement invariance compared to multiple group invariance testing. Adding the lagged covariance of each item has a practical implication for testing invariance by constraining how data need to be formatted to proceed with testing. In multigroup settings, the data are typically formatted vertically, with each row of one's dataset representing a unique person, and the data contain an identifier column for the group of each person. However, to more easily estimate the lagged covariance within each item, formatting the dataset horizontally where each row still represents a unique person, and the repeated measurements of the same items are represented by coded variable names (e.g., item1_t1 and item1_t2). This formatting difference is a subtle distinction between multigroup and longitudinal approaches to measurement invariance but is of practical importance for conducting longitudinal measurement invariance. The procedures for testing measurement invariance are as follows.

Invariance of dimensionality/structure

First, for the configural model, the observed items are set as indicators of a latent response, or "phantom", variable reflective of the construct (Rindskopf, 1984). The observed categorical items indirectly reflect the constructs through each item's respective latent response variable. The DELTA parameterization (see item factor analysis section) is used to specify the categorical indicators in these models so that the scale factors for the latent responses can be investigated in a more restrictive model. In the configural model, though, each item's scales (or total variance of the latent response) are fixed to 1 across waves along with the previously mentioned constraints. The factor

means and variances are fixed to 0 and 1, respectively. Parameters that are freely estimated across waves are the 1) item thresholds and 2) factor loadings between the latent factors and the phantom indicators. The fit of this configural model for equality across waves can be tested using the permutation test approach (Jorgensen et al., 2018). Under the permutation test, the index of time for each variable is permuted (shuffled), the mode is re-estimated, and the resulting fit indices are saved. The distribution of the fit indices based on many permutations of the indices is then compared to the observed value of the fit index. The resulting permutation p-value is the proportion of replications/permutations with a more "extreme" value than the observed index. For the RMSEA, the permutation p-value is the proportion of replications with values *greater* than the observed RMSEA for the configural model. For the CFI, the permutation p-value is the proportion of replications with values *less* than the observed CFI for the configural model. The permutation test better controls Type I error rates than the traditional χ^2 test or alternative fit indices specifically for testing configural invariance (Jorgensen et al., 2018). The permutation test is conducted using 10,000 permutations.

Invariance of Thresholds

Secondly, for the threshold invariance model, part of the identification is achieved by constraining the item thresholds to equality across waves with constraints on wave one parameters. In other words, restrictions on the latent response variables are relaxed once thresholds are equal across waves. For wave one, the latent response intercepts and scales are fixed at **0** and **1**, respectively. For wave two, the latent response intercept and scales are freely estimated. The restrictions on factor means and variances are still needed for both waves, though.

Invariance of Factor Loadings

Next, for factor loading invariance model is tested. In this model, the factor loadings are set to equality across waves. Models with factor loadings equal across waves/groups are sometimes called metric invariance or weak invariance models. The factor means are fixed to 0 across waves, but the factor variances are freely estimated in wave two, with the wave one factor variance fixed to 1 for identification. The location of the underlying factors cannot be uniquely determined across waves of data due to potential differences in the underlying latent response variables, which may differ across waves. However, the relative variability in the underlying factor can be determined once the factor loadings and thresholds are equal (Wu & Estabrook, 2016).

Invariance of Latest Response Intercepts

Next, the equality of the latent response intercepts (i.e., scalar invariance, strong invariance) is then tested across waves. The reader may notice that there is an almost back and forth in which parameters can be identified under which constraints. The back and forth creates a situation where we initially have to assume equality of some parts of the model (e.g., latent response locations) in order to test dimensionality but are able to test the initially assumed components once invariance is established for other parameters.

Invariance of Latest Response Scales

After the equality of the latent response variable intercepts has been supported (i.e., all equal to zero), the scales or variance of the latent responses can be tested for equality across waves of data collection (i.e., strict invariance). Under the DELTA parameterization, the total variance (scale) of the latent response distribution is fixed to unity (1), so what we are assessing in this step is essentially the equality of the residual variances of each item across waves. If invariance of the item-level characteristics (thresholds, intercepts, factor loadings, scales) is tenable, then evaluation of the between wave covariance of factors is possible. The evaluation of the item-level properties across waves is needed to ensure inferences about the factors (and then the domains of interest) are not distorted by differences in item-level measurement over time.

Models Summary and Model Fit

The models described above are summarized in Table 1. In all the models, the latent response variable variance-covariance matrix ($\boldsymbol{\Theta}$) contains the between-wave same-item covariance parameter, and these parameters are freely estimated. The steps described above follow the recommendations of Wu and Estabrook (2016) for the steps of sequential models and how to parameterize each successive model. A full description of the technical details for measurement invariance testing across groups is beyond the scope of this paper; interested readers are referred to Millsap (2011), Wu and Estabrook (2016), Liu et al. (2017), and

Svetina et al. (2020). The rules we followed for assessing the invariance of increasingly restrictive, sequential models were 1) the χ^2 difference test (Asparouhov et al., 2006; Liu et al., 2017); 2) residuals/modification indices (McDonald & Ho, 2002); 3) $\Delta CFI \geq -0.01$ (Cheung & Rensvold, 2002); 4) $\Delta RMSEA \leq 0.01$ (Rutkowski & Svetina, 2017); 5) $\Delta SRMR \leq .01$ (Chen, 2007) and 6) the plausibility of parameter estimates. The χ^2-difference test was conducted using the corrected Satorra and Bentler (2010) test statistic and is based on using the non-robust model χ^2 test statistics. The reported model χ^2 for each model is the robust variant and was not used in the χ^2 difference test.

Additionally, the use of the CFI, RMSEA, and SRMR is not fully endorsed for assessing the relative fit of the increasingly restrictive model set (Liu et al., 2017; Sass et al., 2014). Relying on any one of these metrics to decide if invariance is tenable is, therefore, dubious at best. To overcome this single-metric limitation, researchers should consider the breadth of information across indices and tests to make an informed decision about the tenability of invariance.

Brief Comparison to Traditional Approach

The approach Wu and Estabrook (2016) advocated is still relatively new. The more traditional approach to assessing measurement invariance with categorical indicators within a factor analytic tradition is to assess configural invariance, then metric (or factor loading) invariance, then scalar (or intercept) invariance (Liu et al., 2017; Millsap, 2011). The problem is that thresholds for categorical indicators have a unique role in identifying the statistical model. The thresholds are not simply a replacement of intercepts.

Traditionally, the item factor model was specified using the THETA parameterization when assessing measurement invariance (Millsap, 2011). This was intended to allow for the free estimation of the latent response intercepts, leading the item factor model to be conceptually identical to the factor model for continuous variables. However, Wu and Estabrook (2016) proved how fixing the latent response residual variance to unity does not guarantee the identification of the latent response intercepts relative to a reference group if the thresholds are not constrained. Under some circumstances, the latent response intercept can be empirically identified if the items have enough response categories, but using the THETA parameterization approach is not guaranteed to work when the number of response categories is only two or three.

Demonstration of Analysis

Data for analysis

Data for this tutorial were collected as part of an effort to study human flourishing. The measure itself

Table 1. Item-factor models for testing longitudinal invariance and estimating test-retest reliability

Model	Purpose	Specification
Configural	Test if dimensionality/specified structure is approx. equivalent across waves	$\boldsymbol{\tau}_1 \neq \boldsymbol{\tau}_2$, $\boldsymbol{\Lambda}_1 \neq \boldsymbol{\Lambda}_2$, $\boldsymbol{\alpha}_{1,2} = \mathbf{0}$, $\boldsymbol{\Theta}_{1,2} = \mathbf{1}$
Threshold	Test if thresholds are approx. equal across waves	$\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$, $\boldsymbol{\Lambda}_1 \neq \boldsymbol{\Lambda}_2$, $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$, $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_2$
Loading	Test if factor loadings are approx. equal across waves (metric/weak)	$\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$, $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$, $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$, $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_2$
Latent Response Intercept	Test if latent response intercepts are approx. equal across waves (scalar/strong)	$\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$, $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$, $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$, $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_2$
Latent Response Variance	Test if residual variances (uniqueness) are approx. equal across waves (strict)	$\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$, $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$, $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$, $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$

Note. Waves are denoted by the subscript number; e.g., $\boldsymbol{\tau}_1$ represents the thresholds for wave 1. $\boldsymbol{\tau}$ represents the item thresholds; $\boldsymbol{\Lambda}$ represents the factor loading matrix; $\boldsymbol{\alpha}$ represents the latent response intercepts; and $\boldsymbol{\Theta}$ represents the variance-covariance matrix of the latent responses. Models are specified using the DELTA parameterization.

was developed by Hanson and VanderWeele (2021) to bridge philosophical and psychological traditions of conceptualizations of meaning in life, and the measure is called the Comprehensive Measure of Meaning (CMM). The participants in the data collection were students at the University of British Columbia. The participants received course credit. Informed consent for the data collection was received and all procedures were approved by UBS's Ethics Board. The CMM consists of 21 items intending to broadly measure three domains of one's perceptions of meaning in life. Broadly defining meaning in life into three domains is considered the necessary minimum decomposition in order to appropriately cover the different nuances of one's views on meaning in life (George & Park, 2016; Hanson & VanderWeele, 2021; Heintzelman & King, 2014; King et al., 2006, 2016; Steger, 2012). The goal of the CMM which is to help measure individual differences in degrees of meaning and purpose so that potential ways of promoting meaning can be studied. The response scale for the CMM is seven categorical ordered responses ranging from *Strong disagree* to *Strong agree*. The items and all response labels are given in the appendix for reference. For simplicity of this demonstration, these data were subset to the cases with complete data across waves. The resulting subset of observation consists of 1235 individuals. A summary of the observed data for the demonstration is provided in the appendix.

Response scales with five or more response categories are commonly treated as continuous in scale evaluation. This decision is supported by the simulation work on the use of maximum likelihood with robust correction by Finney and DiStefano (2013) and Rhemtulla et al. (2012), and a tremendous amount of excellent work has followed from this approach. This study emphasizes the categorical nature of the observed data to demonstrate how not making the assumption of continuity is also sometimes useful to evaluate the properties of psycho-social measures.

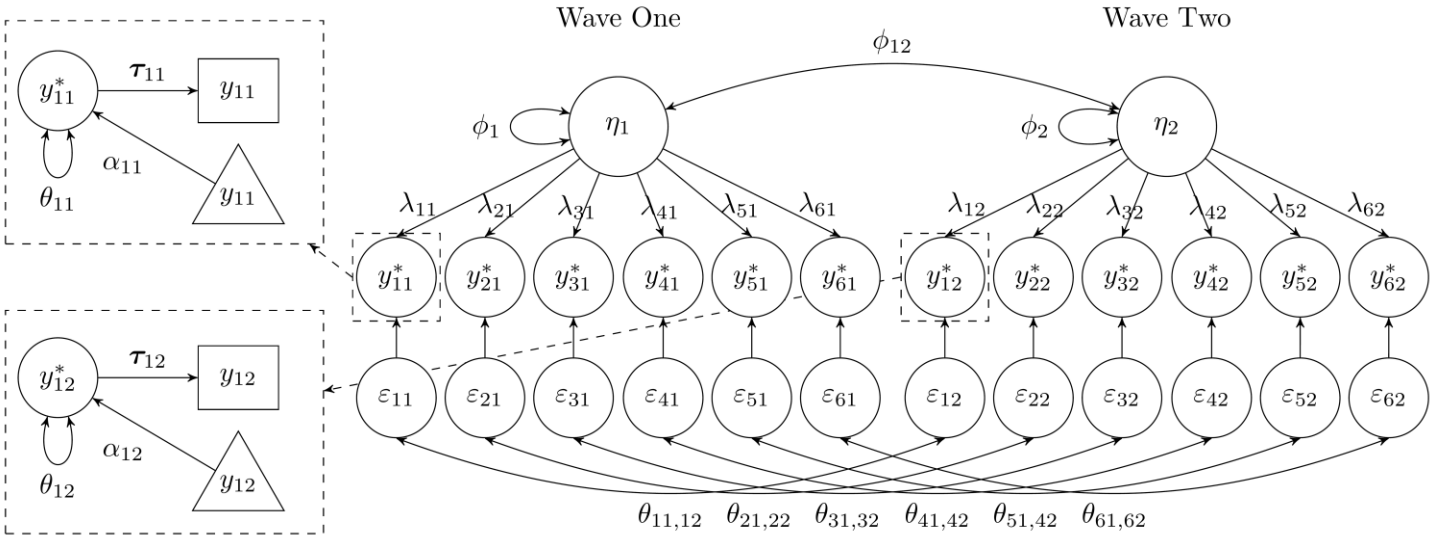
Invariance testing

The first step for longitudinal invariance testing is to specify the base or configural model where the structure is equivalent across waves. A simplified model of only one factor is shown in Figure 2. The path diagram shows the specified thresholds, factor loadings, latent response intercepts, latent response scales, and factor variances.

Configural invariance (i.e., Same dimensional structure and specification). The configural/structural invariance test relies on the permutation test developed by Jorgensen and colleagues (2018). The permutation test first requires ignoring the longitudinal component of these data, which may result in the test being less sensitive to variance (or non-invariance) due to the permutations containing the same individuals more than once in the same "group." The resulting distributions of the chi-square statistic, CFI, RMSEA, and SRMR are shown in Figure 3. In each panel of the figure, the observed value for each index is highlighted as a bolded-dashed (red) vertical line. The results of the permutation test provide evidence of dimensional invariance based on the χ^2 statistic (p=.055), the CFI (p=.989), and RMSEA (p=.055), but the permutation test based on the SRMR (p=.046) provides evidence against invariance.

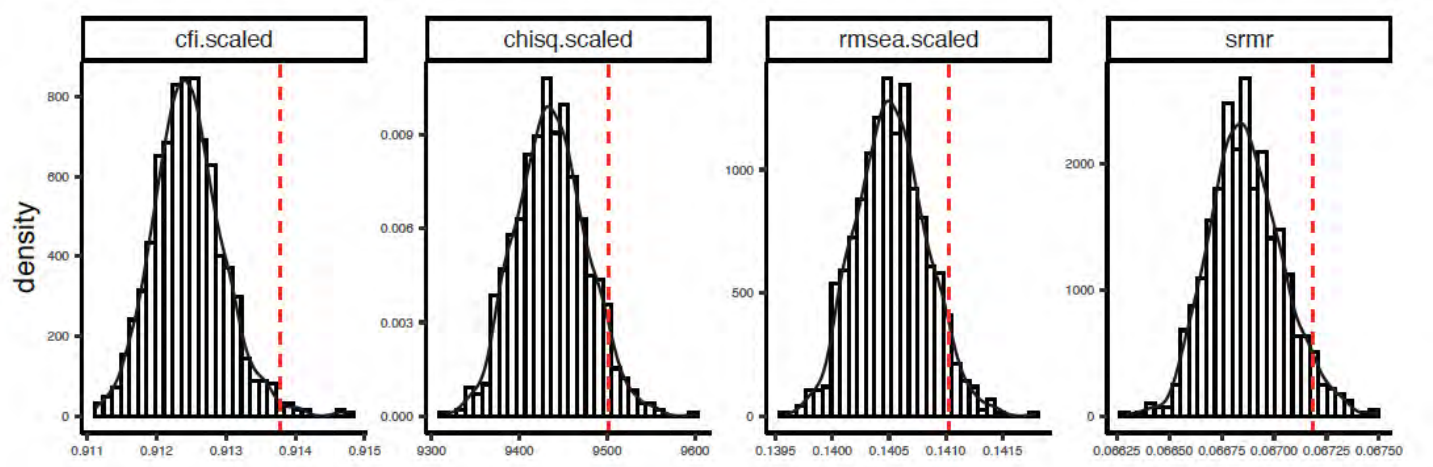
The lack of a consistent result across all permutation tests is somewhat concerning, but this could be due to a general misspecification of the measurement model. The model presented in this demonstration is a simplified version of the full model theorized by Hanson and VanderWeele (2021). In the full model, each domain is further decomposed into two or three subdomains, and the simplification in this demonstration is a major source of the large scaled-χ^2 test statistic. If this known source of model misspecification were not the case, the following steps could be taken to investigate the source of potential non-invariance (or variance) in the configural structure across waves. A common approach to investigate sources of variance (or non-invariance) in the dimensional structure or model specification is to look at modification indices (Kaplan, 1989; Sörbom, 1989). In this case, the top six most extreme values of the modification indices are shown in Table 2. Most of the remaining large modification indices also point to residual covariances within the same domain, which is expected given the theory that subdomains may exist. Another approach to investigate potential sources of variance in dimensional structure across waves is to split the data into two datasets and conduct exploratory factor analysis at each wave. This approach would help to identify any major deviations in the structure across waves, such as the number of factors.

Figure 2. Detailed path diagram (specification) of a general model for invariance testing



Note. The path with τ_{11} represents the categorization of the latent response into the observed ordered response. The boxes pulling out the specification of the latent response variables (e.g., y^*_{11}) represent the DELTA parameterization where the loop variance parameter θ_{11} is the total variance of the latent response.

Figure 3. Permutation test distributions resulting in mixed evidence of configural invariance



Note. The vertical dashed line represents the observed value for each fit index.

Threshold invariance. The investigation of threshold invariance resulted in evidence that the invariance of thresholds was tenable. The χ^2-difference test between the configural model and the model with thresholds constrained equal was nonsignificant ($\Delta\chi^2(84) = 53.76$, $p = .996$). The change in model fit statistics also gave evidence of the tenability of invariance of thresholds ($\Delta CFI = -0.002$, $\Delta RMSEA = -0.004$, $\Delta SRMR = 0.000$).

Loading invariance. The investigation of threshold invariance resulted in evidence that the invariance of thresholds was tenable. The χ^2-difference test between the configural model and the model with thresholds constrained equal was nonsignificant ($\Delta\chi^2(18) = 11.37$, $p = .878$). The change in model fit statistics also gave evidence of the tenability of invariance of thresholds ($\Delta CFI = 0.002$, $\Delta RMSEA = -0.002$, $\Delta SRMR = 0.000$).

Table 2. Modification indices point to residual covariance within domains

Parameter	Modification Index	Expected Parameter Change
$\theta_{D22,D32}$	489.8	0.190
$\theta_{C12,C22}$	460.3	0.206
$\theta_{C11,C21}$	395.7	0.221
$\theta_{D21,D31}$	394.2	0.201
$\theta_{D52,D62}$	312.3	0.249
$\theta_{S52,S62}$	290.1	0.234

Note. $\theta_{D22,D32}$ =residual covariance between item D2 and D3 at time 2 (D22=item D2 at time 2).

Latent response intercept invariance. The investigation of threshold invariance resulted in evidence that the invariance of thresholds was tenable. The χ^2-difference test between the configural model and the model with thresholds constrained equal was nonsignificant ($\Delta\chi^2(18) = 15.23$, $p = .646$). The change in model fit statistics also gave evidence of the tenability of invariance of thresholds ($\Delta CFI = 0.001$, $\Delta RMSEA = -0.001$, $\Delta SRMR = 0.000$).

Latest response scale invariance. The investigation of threshold invariance resulted in evidence that the invariance of thresholds was tenable. The χ^2-difference test between the configural model and the model with thresholds constrained equal was nonsignificant ($\Delta\chi^2(21) = 14.56$, $p = .845$). The change in model fit statistics also gave evidence of the tenability of invariance of thresholds ($\Delta CFI = 0.013$, $\Delta RMSEA = -0.008$, $\Delta SRMR = 0.001$).

Example Concise Reporting Write-up

In the write-up that follows, I wrote the methods and results section as if I were contributing to an article on the development of the Comprehensive Measure of Meaning. The purpose of such a study is not to spend great detail explaining the statistical methods but to make an argument that scores on the domains of the construct of interest can be interpreted consistently across waves of data collection. Providing evidence of the consistency of scores across waves provides greater evidence to the overall claim that the Comprehensive Measure of Meaning can be used to assess individuals' perceptions of their meaning in life. This is useful so that further comparisons of individuals on the dimension of the Comprehensive Measure of Meaning within and between time points are not influenced by an effect of time, thus ruling out one potential source that could explain the difference we observed between a focal group characteristic such as religious affiliation or changes in affiliation over time. The bulk of such an article on the development of the Comprehensive Measure of Meaning is to provide validity evidence in the form of the theoretical framework, test content, response process (usually through cognitive interviews), internal structure (e.g., dimensionality), relationship with other variables, and consequences of measurement. An evaluation of the longitudinal invariances can be considered one aspect of the evaluation of the internal structure of a measure. That being said, concisely writing the results of such analysis will necessarily lose information, especially in submissions to journals with tight word limits. The write-up below is one possible way to reduce the information while still providing the reader with enough information to evaluate your methods and results.

Methods. The three domains of the Comprehensive Measure of Meaning (CMM) are hypothesized as stable over time, at least within a one-year time gap. This study aims to provide evidence that the psychometric properties of the CMM are consistent over that one-year time period. The tenability measurement invariance across the two waves of data was investigated to provide that evidence. The steps to investigating invariance followed the recommendations of Wu and Estabrook (2016), where invariance was probed with respect to (1) dimensionality/structural specification, (2) thresholds, (3) factor loadings, (4) latent response intercepts, and (5) latent response scales. Configural invariance was assessed using the permutation approach (Jorgensen et al., 2018). The rules we followed for assessing the invariance of increasingly restrictive models were 1) the χ^2 difference test (Asparouhov et al., 2006; Liu et al., 2017); 2)

residuals/modification indices (McDonald & Ho, 2002); 3) $\Delta CFI \geq -0.01$ (Cheung & Rensvold, 2002); 4) $\Delta RMSEA \leq 0.01$ (Rutkowski & Svetina, 2017); 5) $\Delta SRMR \leq .01$ (Chen, 2007) and 6) the plausibility of parameter estimates. The longitudinal aspect of these data was accounted for in the models by specifying correlated residuals for the same item across waves. Finally, the consistency of the scores was summarised using the correlation of the same factors across waves (e.g., the correlation between the Coherence factor at wave one and the Coherence factor at wave two). Correlation estimates above 0.70 will be seen as evidence of sufficient consistency across time, and this means the scores account for at least 50% of the variability in scores over time, which we have determined is acceptable.

Results. Invariance testing resulted in evidence of comparability of factors across waves (results are summarized in Table 3). The permutation test for configural invariance gave some evidence for dimensionality invariance (χ^2 p-value = .055; CFI p-value= .989; RMSEA p-value = .055, and SRMR p-value = .046); however, investigation of modification indices pointed to residual covariance within domains in the same wave as expected given subdomains were written but not investigated in this study. In summary, invariance testing gives evidence that the domains measured by the CMM can be interpreted as the same factors across waves for comparisons.

The resulting cross-wave correlations are Coherence (.693, [.660, .727]), Significance (.730, [.695, .764]), and Direction (.691, [.659, .723]). All estimates are about 0.7 or higher, which gives evidence that scores from this measure provide a consistent measurement for these three domains. Our online supplemental material offers a more detailed summary of all fitted models (REFERENCE SUPPLEMENT MATERIAL).

Health Retirement Study Example (Concise Write-up Only)

The following example focuses on evaluating the Satisfaction with Life Scale (SWLS; Diener et al., 1985). The SWLS contains five items; "In most ways my life is close to my ideal"(Ideal); "The conditions of my life are excellent" (Excellent); "I am satisfied with my life" (Satisfied); "So far I have gotten the important things I want in life" (Important); and "If I could live my life over, I would change almost nothing" (Change). The following analysis demonstrates how three waves of data can be jointly used to assess the longitudinal measurement invariance.

Data. Data for this study are a three-wave (2008, 2012, and 2016) subset of the Health and Retirement Study (HRS; University of Michigan, 2016). The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. These data were limited to individuals who responded to at least one item per wave. The resulting

Table 3. Statistical evidence for longitudinal measurement invariance

Invariance Level	$\chi^2(df)$	S-B $\Delta\chi^2(df)$	CFI	RMSEA	SRMR
Configural	8515.93 (783)		.914	.089	.056
Permutation p-values	*.055*		*.989*	*.055*	*.046*
Threshold	8753.81 (867)	53.76 (84)†	.912	.086	.056
Loading (Metric)	8590.83 (885)	11.37 (18)†	.914	.084	.056
LR Intercept (Scalar)	8546.15 (903)	15.23 (18)†	.915	.083	**.056**
LR Scale (Strong)	**7368.28 (924)**	**14.55 (21)**†	**.928**	**.075**	.057

Note. $N = 1235$. Bolded values indicate the best approximating model of those tested. The $\chi^2(df)$ reported for each model is the robust (scaled) Satorra and Bentler (2001) test statistic. The S-B $\Delta\chi^2(df)$ is the robust difference test that uses the standard (not robust) test statistics which were not reported, where $^{\dagger}p > .05$ indicates evidence of invariance between models.

sample size was 6,576. The Health and Retirement Study data are publicly available online at https://hrsdata.isr.umich.edu/data-products/public-survey-data.

Methods. A single dimension of Life Satisfaction as measured by the SWLS is hypothesized to be stable in retirement. The aim of this study is to provide evidence that the psychometric properties of the SWLS are consistent over an eight-year period. The tenability measurement invariance across the three waves of data was investigated to provide that evidence. The steps to investigating invariance followed the recommendations of Wu and Estabrook (2016), where invariance was probed with respect to (1) dimensionality/structural specification, (2) thresholds, (3) factor loadings, (4) latent response intercepts, and (5) latent response scales. Configural invariance was assessed using the permutation approach (Jorgensen et al., 2018). The rules we followed for assessing the invariance of increasingly restrictive models were 1) the χ^2 difference test (Asparouhov et al., 2006; Liu et al., 2017); 2) residuals/modification indices (McDonald & Ho, 2002); 3) $\Delta CFI \geq -0.01$ (Cheung & Rensvold, 2002); 4) $\Delta RMSEA \leq 0.01$ (Rutkowski & Svetina, 2017); 5) $\Delta SRMR \leq .01$ (Chen, 2007) and 6) the plausibility of parameter estimates. The longitudinal aspect of these data was accounted for in the models by specifying correlated residuals for the same item across waves. Finally, the consistency of the scores was summarised using the correlation of the factor across waves. Correlation estimates above 0.70 will be seen as evidence of sufficient consistency across time, and this means the scores account for at least 50% of the variability in scores over time, which we have determined is acceptable.

These data were collected over three waves and ranged over eight years (2008, 2012, and 2016). When more than two waves of data are available, an investigation of how the cross-wave residual covariances of items over time can be informative of which items are more related over time than expected by a single factor model. In this evaluation of longitudinal invariance, the cross-wave residual covariances were compared to identify how lag effects may change. Code for these analyses are available in our Online Supplement.

Results. Invariance testing results in evidence of comparability of factors across waves (results are summarized in Table 4). The permutation test for configural invariance gave some evidence of non-invariance (χ^2 p-value = .012; CFI p-value= .001; RMSEA p-value = .012, and SRMR p-value = .001). Additionally, the residuals and modification indices pointed to residual covariance among items within and between waves. In summary, invariance testing gives evidence that Life Satisfaction measured by the SWLS does not provide a consistent measurement of Life Satisfaction over the eight-year period in this HRS sample. The lack of configural invariance over time provides evidence that individuals in this sample respond to the five items sufficiently differently enough that unique statistical models should be used for each wave to evaluate Life Satisfaction.

If one assumes invariance despite the evidence of variance in the measurement model parameters over time, the resulting correlation between waves 2008 and 2012 was .638, ([.612, .663]95% CI), between waves 2012 and 2016 was .615, ([.588, .641]95% CI), and between waves 2008 and 2016 was .559, ([.530, .588]95% CI). Similar estimates of correlations were found even in the configural model only (see Table 5). All estimates are below the target of 0.70, which gives evidence that scores on the SWLS are not necessarily consistent over four to eight-year ranges. The cross-wave residual covariances of the five SWLS items are shown in Table 5. The cross-wave correlations provide evidence that changes in responses to item *Excellent* over time are captured well by a single factor but that changes in responses to item *Change* are not captured well by a single factor (residual correlations of 0.42-0.48). Our online supplemental material offers a more detailed summary of all fitted models (REFERENCE SUPPLEMENT MATERIAL).

Concluding Remarks

Measurement invariance is a combination of the validity of interpretations we can give to a set of scores on a measure and a technical component in the statistical models used to evaluate differences among individuals or groups on a measure. The former, validity, component of measurement invariances emphasizes how comparable scores and responses to items are among groups or over time (Horn & Mcardle, 1992; Meredith, 1993). The latter, statistical, component of measurement invariance emphasizes the

Table 4. Statistical evidence for longitudinal measurement invariance of the SWLS

Invariance Level	$\chi^2(df)$	S-B $\Delta\chi^2(df)$	CFI	RMSEA	SRMR
Configural	1991.2 (72)		.987	.064	.031
Permutation p-values	*.012*		*.001*	*.012*	*.001*
Threshold	3416.8 (86)	1621.5 (14)	.977	.077	.031
Loading (Metric)	4096.6 (94)	575.3 (8)	.972	.080	.032
LR Intercept (Scalar)	5433.8 (102)	1242.1 (8)	.963	.089	.033
LR Scale (Strong)	2864.6 (112)	−1001.8 (10)†	.964	.094	.032

Note. $N = 6576$. The $\chi^2(df)$ reported for each model is the robust (scaled) Satorra and Bentler (2001) test statistic. The S-B $\Delta\chi^2(df)$ is the robust difference test that uses the standard (not robust) test statistics which were not reported, where $^{\dagger}p > .05$ indicates evidence of invariance between models.

Table 5. SWLS cross-wave correlations of SWL factor and items

Item/Factor	cor(2008w,2012w)	cor(2012w,2016w)	cor(2008w,2016w)
Configural Model			
SWL Factor	0.60	0.60	0.56
Ideal	0.18	0.17	0.03
Excellent	−0.03	0.06	0.01
Satisfied	0.21	0.18	0.02
Important	0.34	0.38	0.31
Change	0.43	0.42	0.46
Full Invariance Model			
SWL Factor	0.64	0.62	0.56
Ideal	0.20	0.12	0.02
Excellent	−0.01	0.04	0.00
Satisfied	0.20	0.22	0.03
Important	0.34	0.40	0.32
Change	0.47	0.47	0.48

Note. $N = 6576$. cor(2008w,2012w) = correlation between 2008 wave and 2012 wave. Item correlations represent the estimated residual correlations over time.

assumptions and particular statistical models used to make group comparisons (Van de Vijver et al., 2019, p.92). This tutorial focuses on comparing scores across time, and the comparability of scores can be empirically assessed as demonstrated. As the number of time points increases, the method described in this tutorial can grow difficult to implement, and researchers may find helpful solutions in how comparisons across many groups are often made in the context of large-scale cross-cultural studies. Interested readers are referred to the chapters in Van de Vijver et al. (2019) for an excellent discussion of some recent developments in large-scale measurement invariance assessment.

The evaluation of scale properties is a necessary component of research with psycho-social constructs, especially for self-report questionnaires. Being able to demonstrate that scores can be interpreted consistently over time is one piece of evidence that researchers can provide to make this argument. As I hoped to demonstrate in this tutorial, researchers can use more methodological rigor and estimate the degree of consistency of properties and scores within an item factor analysis framework. The results of such an

expanded analysis will provide more insight into the longitudinal measurement properties of their data. The deeper insight can help to provide stronger evidence of the consistency of psychometric properties on the measure under investigation.

Data Availability Statement

The data used in this study are openly available, along with the analysis scripts to reproduce the results on the Open Science Framework at (https://osf.io/2vs3w/).

References

Asparouhov, Muthén, & Muthén. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics (mplus web notes no. 10)* (tech. rep.). http://www.statmodel.com/download/webnotes/webnote10.pdf

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second). The Guilford Press.

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. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9* (2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Publications.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49* (1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Duckworth, A., & Yeager, D. (2015). Measurement matters. *Educational Researcher*, *44* (4), 237–251. https://doi.org/10.3102/0013189x15584327

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd, pp. 439–492). Information Age Publishing.

George, L. S., & Park, C. L. (2016). Meaning in life as comprehension, purpose, and mattering: Toward integration and new research questions. *Review of General Psychology*, *20* (3), 205–220.

Grutzmacher, S. V., L., & Hartig, J. (2021). Are questionnaire scales which measure non-cognitive constructs suitable as school effectiveness criteria? a measurement invariance analysis. *School Effectiveness and School Improvement*, *0* (0), 1–18. https://doi.org/10.1080/09243453.2021.1903511

Hanson, J. A., & VanderWeele, T. J. (2021). The comprehensive measure of meaning: Psychological and philosophical foundations. Oxford University Press. https://doi.org/10.1093/oso/9780197512531.003.0013

Heintzelman, S. J., & King, L. A. (2014). Life is pretty meaningful. *American Psychologist*, *69* (6), 561–574.

Horn, J., & Mcardle, J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18* (3), 117–144. https://doi.org/10.1080/03610739208253916

Jorgensen, T. D., Kite, B. A., Chen, P.-Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*, *23* (4), 708–728. https://doi.org/10.1037/met0000152

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15* (1), 136–153. https://doi.org/10.1080/10705510701758406

Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate Behavioral Research*, *24* (3), 285–305. https://doi.org/10.1207/s15327906mbr2403_2

King, L. A., Heintzelman, S. J., & Ward, S. J. (2016). Beyond the search for meaning: A contemporary science of the experience of meaning in life. *Current Directions in Psychological Science*, *25* (4), 211–216.

King, L. A., Hicks, J. A., Krull, J. L., & Del Gaiso, A. K. (2006). Positive affect and the experience of meaning in life. *Journal of Personality and Social Psychology*, *90* (1), 179–196.

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22* (3), 486–506. https://doi.org/10.1037/met0000075

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7* (1), 64–82. https://doi.org/10.1037/1082-989x.7.1.64

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52* (6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58* (4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In *New methods for the analysis of change* (pp. 203–240, Vol. 442). American Psychological Association. https://doi.org/10.1037/10409-007

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. Millsap, R. E., & Cham, H. (2012). Handbook of developmental research methods. In B. Laursen, T. Little, & N. Card (Eds.), *Investigating factorial invariance in longitudinal data* (pp. 109–127). Guilford.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11* (1), 3–31. https://doi.org/10.3102/10769986011001003

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49* (1), 115–132. https://doi.org/10.1007/BF02294210

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17* (3), 354–373. https://doi.org/10.1037/a0029315

Rindskopf, D. (1984). Using phantom and imaginary latent variables to parameterize constraints in linear structural models. *Psychometrika*, *49* (1), 37–47. https://doi.org/10.1007/BF02294204

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*, *30* (1), 39–51. https://doi.org/10.1080/08957347.2016.1243540

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21* (2), 167–180. https://doi.org/10.1080/10705511.2014.882658

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75* (2), 243–248. https://doi.org/10.1007/s11336-009-9135-y

Sörbom, D. (1989). Model modification. *Psychometrika*, *54* (3), 371–384. https://doi.org/10.1007/BF02294623

Steger, M. F. (2012). *Experiencing meaning in life: Optimal functioning at the nexus of well-being, psychopathology, and spirituality* (P. Wong, Ed.; 2nd). Routledge.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using mplus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27* (1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

University of Michigan. (2016). Health and Retirement Study, cross-wave tracker file, public use dataset. https://hrsdata.isr.umich.edu/data-products/public-survey-data

Van de Vijver, F. J. R., Avvisati, F., Davidov, E., Eid, M., Fox, J.-P., Le Donne, N., Lek, K., Meuleman, B., Paccagnella, M., & van de Schoot, R. (2019). *Invariance analyses in large-scale studies*. OECD Publishing. https://doi.org/10.1787/254738dd-en

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior research methods*, *55* (2), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12* (1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81* (4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Citation:
Padgett, R. N. (2023). A tutorial on cross wave measurement invariance testing with item factor analysis. *Practical Assessment, Research, & Evaluation*, 28(13). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/13/

Corresponding Author:
R. Noah Padgett
Harvard T.H. Chan School of Public Health, Department of Epidemiology, Harvard University
677 Huntington Avenue, Boston, MA 02115, USA
Email: npadgett@hsph.harvard.edu

Appendix A Comprehensive Measure of Meaning

The 21 items of the CMM are given below.

- Coherence
  C1. I have a clear understanding of the ultimate meaning of life.
  C2. The meaning of life in the world around us is evident to me.
  C3. I have a framework that allows me to understand or make sense of human life.
  C4. I understand my life's meaning.
  C5. I can make sense of the things that happen in my life.
  C6. I have a philosophy of life that helps me understand who I am.

- Significance
  S1. I am living the kind of meaningful life I want to live.
  S2. Living is deeply fulfilling.
  S3. I feel like I have found a really significant meaning in my life.
  S4. The things I do are important to other people.
  S5. I have accomplished much in life as a whole.
  S6. I make a significant contribution to society.

- Direction
  D1. I have been aware of an all-encompassing and consuming purpose towards which my life has been directed.
  D2. I have a sense of mission or calling.
  D3. I have a mission in life that gives me a sense of direction.
  D4. I have a sense of direction and purpose in life.
  D5. I can describe my life's purposes.
  D6. My current aims match with my future aspirations.
  D7. In my life I have very clear goals and aims.
  D8. I have goals in life that are very important to me.
  D9. I have definite ideas of things I want to do.

- Response Scale: *Strong disagree, Disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Agree, Strongly agree*

Table A1. Summary statistics of observed items

				Category Proportions						
Item	Wave	Mean	SD	1	2	3	4	5	6	7
Coherence										
C1	1	3.78	1.63	0.09	0.19	0.10	0.27	0.19	0.11	0.04
	2	3.67	1.64	0.07	0.19	0.12	0.25	0.21	0.11	0.04
C2	1	3.81	1.58	0.03	0.10	0.10	0.22	0.31	0.19	0.06
	2	3.66	1.60	0.07	0.14	0.13	0.23	0.23	0.16	0.05
C3	1	4.47	1.45	0.01	0.05	0.09	0.16	0.36	0.28	0.06
	2	4.33	1.53	0.02	0.09	0.10	0.21	0.28	0.21	0.09
C4	1	4.08	1.61	0.04	0.10	0.14	0.22	0.26	0.18	0.05
	2	3.97	1.62	0.03	0.05	0.09	0.19	0.23	0.28	0.12
C5	1	4.85	1.29	0.04	0.14	0.16	0.23	0.22	0.15	0.06
	2	4.82	1.30	0.02	0.06	0.09	0.24	0.28	0.25	0.05
C6	1	4.61	1.48	0.06	0.13	0.19	0.21	0.25	0.13	0.03
	2	4.52	1.50	0.06	0.15	0.17	0.27	0.22	0.11	0.02
Significance										
S1	1	4.32	1.50	0.07	0.16	0.13	0.33	0.20	0.09	0.03
	2	4.12	1.54	0.07	0.16	0.13	0.18	0.24	0.15	0.06
S2	1	4.85	1.52	0.05	0.13	0.13	0.16	0.27	0.18	0.07
	2	4.71	1.48	0.03	0.08	0.10	0.14	0.34	0.23	0.08
S3	1	4.11	1.56	0.05	0.13	0.13	0.19	0.26	0.18	0.05
	2	3.95	1.55	0.01	0.04	0.06	0.18	0.29	0.31	0.09
S4	1	4.65	1.36	0.02	0.09	0.14	0.15	0.31	0.20	0.08
	2	4.60	1.39	0.01	0.03	0.05	0.10	0.28	0.35	0.19
S5	1	3.97	1.50	0.02	0.05	0.09	0.11	0.31	0.27	0.15
	2	3.91	1.55	0.10	0.22	0.14	0.19	0.22	0.11	0.03
S6	1	3.83	1.45	0.09	0.20	0.14	0.22	0.21	0.10	0.03
	2	3.75	1.51	0.04	0.12	0.11	0.20	0.31	0.17	0.05
Direction										
D1	1	3.80	1.48	0.07	0.15	0.16	0.19	0.23	0.15	0.04
	2	3.64	1.54	0.02	0.05	0.08	0.15	0.38	0.27	0.05
D2	1	4.07	1.69	0.03	0.09	0.11	0.20	0.29	0.21	0.07
	2	3.87	1.67	0.05	0.13	0.18	0.20	0.25	0.15	0.04
D3	1	4.30	1.63	0.03	0.06	0.12	0.21	0.25	0.24	0.09
	2	4.11	1.67	0.05	0.16	0.18	0.23	0.21	0.12	0.05
D4	1	4.66	1.50	0.02	0.08	0.10	0.23	0.28	0.24	0.05
	2	4.49	1.56	0.07	0.15	0.19	0.19	0.24	0.13	0.03
D5	1	4.24	1.59	0.07	0.17	0.18	0.24	0.21	0.11	0.02
	2	4.07	1.62	0.10	0.17	0.15	0.29	0.18	0.08	0.03
D6	1	5.00	1.31	0.09	0.17	0.16	0.18	0.22	0.13	0.05
	2	4.88	1.36	0.07	0.16	0.12	0.16	0.27	0.17	0.05
D7	1	4.58	1.50	0.05	0.10	0.10	0.16	0.31	0.22	0.06
	2	4.53	1.56	0.07	0.14	0.16	0.19	0.24	0.17	0.04
D8	1	5.39	1.28	0.02	0.05	0.07	0.18	0.33	0.27	0.08
	2	5.26	1.39	0.04	0.10	0.13	0.14	0.30	0.21	0.08
D9	1	5.07	1.44	0.02	0.04	0.06	0.10	0.27	0.35	0.16
	2	5.03	1.46	0.03	0.05	0.09	0.10	0.30	0.30	0.13

Note. $N = 1235$.

Appendix B Satisfaction with Life Scale

The SWLS contains five items; "In most ways my life is close to my ideal"(Ideal); "The conditions of my life are excellent" (Excellent); "I am satisfied with my life" (Satisfied); "So far I have gotten the important things I want in life" (Important); and "If I could live my life over, I would change almost nothing" (Change).

Table B1. Summary statistics of observed items

Item	% Miss	Mean	SD	P1	P2	P3	P4
		Wave – 2008					
Ideal	19.33	2.51	0.94	0.15	0.35	0.34	0.15
Excellent	19.45	2.52	0.96	0.16	0.34	0.32	0.16
Satisfied	19.05	2.87	0.98	0.10	0.24	0.34	0.10
Important	19.08	2.87	0.97	0.10	0.25	0.34	0.10
Change	19.05	2.36	1.02	0.24	0.33	0.26	0.24
		Wave – 2012					
Ideal	4.46	2.37	0.94	0.19	0.39	0.29	0.19
Excellent	4.43	2.39	0.96	0.19	0.38	0.29	0.19
Satisfied	4.23	2.76	1.00	0.13	0.27	0.32	0.13
Important	4.08	2.79	0.98	0.11	0.28	0.32	0.11
Change	3.94	2.32	1.02	0.25	0.35	0.23	0.25
		Wave – 2016					
Ideal	21.86	2.57	0.93	0.13	0.35	0.34	0.13
Excellent	21.85	2.55	0.94	0.14	0.35	0.33	0.14
Satisfied	21.62	2.87	0.97	0.10	0.24	0.34	0.10
Important	21.58	2.90	0.95	0.08	0.26	0.34	0.08
Change	21.48	2.40	1.01	0.22	0.34	0.26	0.22

Note. $N = 6576$.

Appendix C

R Code for Longitudinal Invariance Testing

```
# Longitudinal model
configural.model.wide <- "
# Wave 1
# factor structure
f1 =~ NA*item1_1 + item2_1 + item3_1
# factor variance
f1 ~~ 1*f1
# Wave 2
f2 =~ NA*item1_2 + item2_2 + item3_2
f2 ~~ 1*f2
# cross wave factor covariances
f1 ~~ f2
# cross wave item residual covariances
item1_1 ~~ item1_2
item2 _ 1 ~~ item2 _ 2
item3 _ 1 ~~ item3 _ 2 "
# multigroup equivalence
configural.model.multigroup <- "
# factor structure
f1 =~ NA*item1_1 + item2_1 + item3_1
# factor variance
f1 ~~ 1*f1 "

# ==================================== #
# Test configural invariance
# Jorgenson et al. (2018) permutation test

# obtain multigroup syntax
syntax.config <- measEq.syntax(
    configural.model = configural.model.multigroup,
    data = analysis.dat,
    ordered=T,
    parameterization = "delta",
    ID.cat = "Wu.Estabrook.2016",
    ID.fac = "std.lv",
    group.equal = "configural",
    group = "wave"
  )
  mod.config <- as.character(syntax.config)
  cat(mod.config) # to see how base model is changed

fit.config <- cfa(mod.config, data=analysis.dat, group = "wave")
# note , the permuteMeasEq takes a LONG time to run,
# recommend to go get coffee
```

```
out.config <- permuteMeasEq(nPermute = 10000, con = fit.config, show
   Progress = T)
summary(out.config)

# estimate longitudinal model
fit.config.wide <- cfa(
  model=configural.model.wide,
  data = analysis.dat.wide, ordered = T
)
summary(fit.config.wide, standardized=T, fit.measures=T)

# ================================== #
# Test threshold invariance

threshold.model.model.wide <- "
# Wave 1
# factor structure
f1 =~ NA*item1_1 + item2_1 + item3_1
# factor variance
f1 ~~ 1*f1
# factor mean
f1 ~ 0*f1
# thresholds (assuming 3 response categories)
item1_1 | item1.thr1*t1 + item1.thr2*t2
item2_1 | item2.thr1*t1 + item2.thr2*t2
item3_1 | item3.thr1*t1 + item3.thr2*t2
# latent response intercepts
item1_1 ~ 0*1
item2_1 ~ 0*1
item3_1 ~ 0*1
# latent response scales
item1_1 ~*~ 1*item1_1
item2_1 ~*~ 1*item1_1
item3_1 ~*~ 1*item1_1
# Wave 2
# factor loadings
f2 =~ NA*item1_2 + item2_2 + item3_2
# factor variance
f2 ~~ 1*f2
# factor mean
f2 ~ 0*f2
# thresholds (assuming 3 response categories)
item1_2 | item1.thr1*t1 + item1.thr2*t2
item2_2 | item2.thr1*t1 + item2.thr2*t2
item3_2 | item3.thr1*t1 + item3.thr2*t2
# latent response intercepts - free - relative to wave 1
item1_2 ~ NA*1
item2_2 ~ NA*1
```

```
item3_2 ~ NA*1
# latent response scales - free - relative to wave 1
item1_2 ~*~ NA*item1_2
item2_2 ~*~ NA*item2_2
item3_2 ~*~ NA*item3_2
# cross wave factor covariances
f1 ~~ f2
# cross wave item residual covariances
item1_1 ~~ item1_2
item2 _ 1 ~~ item2 _ 2
item3 _ 1 ~~ item3 _ 2 "

#       fit       model
fit.threshold.wide <- cfa(
  model=threshold.model.wide,
  data = analysis.dat.wide, ordered = T
)
summary(fit.threshold.wide, standardized=T, fit.measures=T)

## test equivalence of thresholds, given equivalence of configural model
lavTestLRT(fit.config.wide, fit.threshold.wide, method = "satorra. bentler.
   2010")


# ================================== #
# Test loading invariance

loading.model.wide <- "
# Wave 1
# factor structure
# NOTE: you need to enter the first item of each factor twice in order to set
   the label (X.lambda.1) so that lavaan estimates the loading instead of
   fixing it to 1
f1 =~ NA*item1_1 + lambda1*item1_1 + lambda2*item2_1 + lambda3*item3_1
# factor variance
f1 ~~ 1*f1
# factor mean
f1 ~ 0*f1
# thresholds (assuming 3 response categories)
item1_1 | item1.thr1*t1 + item1.thr2*t2
item2_1 | item2.thr1*t1 + item2.thr2*t2
item3_1 | item3.thr1*t1 + item3.thr2*t2
# latent response intercepts
item1_1 ~ 0*1
item2_1 ~ 0*1
item3_1 ~ 0*1
# latent response scales
item1_1 ~*~ 1*item1_1
```

```
item2_1 ~*~ 1*item1_1
item3_1 ~*~ 1*item1_1
# Wave 2
# factor loadings
f2 =~ NA*item1_2 + lambda1*item1_2 + lambda2*item2_2 + lambda3*item3_2
# factor variance  - free - relative to wave 1
f2 ~~ NA*f2
# factor mean
f2 ~ 0*f2
# thresholds (assuming 3 response categories)
item1_2 | item1.thr1*t1 + item1.thr2*t2
item2_2 | item2.thr1*t1 + item2.thr2*t2
item3_2 | item3.thr1*t1 + item3.thr2*t2
# latent response intercepts - free - relative to wave 1
item1_2 ~ NA*1
item2_2 ~ NA*1
item3_2 ~ NA*1
# latent response scales - free - relative to wave 1
item1_2 ~*~ NA*item1_2
item2_2 ~*~ NA*item2_2
item3_2 ~*~ NA*item3_2
# cross wave factor covariances
f1 ~~ f2
# cross wave item residual covariances
item1_1 ~~ item1_2
item2 _ 1 ~~ item2 _ 2
item3 _ 1 ~~ item3 _ 2 "

#      fit      model
fit.loading.wide <- cfa(
  model = loading . model . wide ,
  data = analysis . dat . wide , ordered=T
)
summary(loading.model.wide, standardized=T, fit.measures=T)

## test equivalence of loadings, given equivalence of configuration and
   threshold
lavTestLRT(fit.threshold.wide, loading.model.wide, method = "satorra
   .bentler.2010")


# ================================== #
# Test latent response intercept invariance

lrintercept.model.wide <- "
# Wave 1
# factor structure
```

```
# NOTE: you need to enter the first item of each factor twice in order to set
   the label (X.lambda.1) so that lavaan estimates the loading instead of
   fixing it to 1
f1 =~ NA*item1_1 + lambda1*item1_1 + lambda2*item2_1 + lambda3*item3_1
# factor variance
f1 ~~ 1*f1
# factor mean
f1 ~ 0*f1
# thresholds (assuming 3 response categories)
item1_1 | item1.thr1*t1 + item1.thr2*t2
item2_1 | item2.thr1*t1 + item2.thr2*t2
item3_1 | item3.thr1*t1 + item3.thr2*t2
# latent response intercepts
item1_1 ~ 0*1
item2_1 ~ 0*1
item3_1 ~ 0*1
# latent response scales
item1_1 ~*~ 1*item1_1
item2_1 ~*~ 1*item1_1
item3_1 ~*~ 1*item1_1
# Wave 2
# factor loadings
f2 =~ NA*item1_2 + lambda1*item1_2 + lambda2*item2_2 + lambda3*item3_2
# factor variance  - free - relative to wave 1
f2 ~~ NA*f2
# factor mean
f2 ~ 0*f2
# thresholds (assuming 3 response categories)
item1_2 | item1.thr1*t1 + item1.thr2*t2
item2_2 | item2.thr1*t1 + item2.thr2*t2
item3_2 | item3.thr1*t1 + item3.thr2*t2
# latent response intercepts
item1_2 ~ 0*1
item2_2 ~ 0*1
item3_2 ~ 0*1
# latent response scales - free - relative to wave 1
item1_2 ~*~ NA*item1_2
item2_2 ~*~ NA*item2_2
item3_2 ~*~ NA*item3_2
# cross wave factor covariances
f1 ~~ f2
# cross wave item residual covariances
item1_1 ~~ item1_2
item2 _ 1 ~~ item2 _ 2
item3 _ 1 ~~ item3 _ 2 "

#       fit        model
fit.lrintercept.wide <- cfa(
```

```
  model=lrintercept.model.wide,
  data = analysis.dat.wide, ordered = T
)
summary(fit.lrintercept.wide, standardized=T, fit.measures=T)

## test equivalence of latent response intercepts, given equivalence of
    configuration, thresholds, and loadings
lavTestLRT(fit.loading.wide, fit.lrintercept.wide, method = "satorra
   .bentler.2010")

# ================================== #
# Test latent response scale invariance

lrscale.model.wide <- "
# Wave 1
# factor structure
# NOTE: you need to enter the first item of each factor twice in order to set
   the label (X.lambda.1) so that lavaan estimates the loading instead of
   fixing it to 1
f1 =~ NA*item1_1 + lambda1*item1_1 + lambda2*item2_1 + lambda3*item3_1
# factor variance
f1 ~~ 1*f1
# factor mean
f1 ~ 0*f1
# thresholds (assuming 3 response categories)
item1_1 | item1.thr1*t1 + item1.thr2*t2
item2_1 | item2.thr1*t1 + item2.thr2*t2
item3_1 | item3.thr1*t1 + item3.thr2*t2
# latent response intercepts
item1_1 ~ 0*1
item2_1 ~ 0*1
item3_1 ~ 0*1
# latent response scales
item1_1 ~*~ 1*item1_1
item2_1 ~*~ 1*item1_1
item3_1 ~*~ 1*item1_1
# Wave 2
# factor loadings
f2 =~ NA*item1_2 + lambda1*item1_2 + lambda2*item2_2 + lambda3*item3_2

# factor variance  - free - relative to wave 1
f2 ~~ NA*f2
# factor mean
f2 ~ 0*f2
# thresholds (assuming 3 response categories)
item1_2 | item1.thr1*t1 + item1.thr2*t2
item2_2 | item2.thr1*t1 + item2.thr2*t2
item3_2 | item3.thr1*t1 + item3.thr2*t2
```

```
 # latent response intercepts
 item1_2 ~ 0*1
 item2_2 ~ 0*1
 item3_2 ~ 0*1
 # latent response scales - fixed to same value as wave 1
 item1_2 ~*~ 1*item1_2
 item2_2 ~*~ 1*item2_2
 item3_2 ~*~ 1*item3_2
 # cross wave factor covariances
 f1 ~~ f2
 # cross wave item residual covariances
 item1_1 ~~ item1_2
 item2 _ 1 ~~ item2 _ 2
 item3 _ 1 ~~ item3 _ 2 "

 #      fit      model
 fit.lrscale.wide <- cfa(
   model = lrscale . model . wide , data =
   analysis . dat . wide , ordered = T
 )
 summary(fit.lrscale.wide, standardized=T, fit.measures=T)

 ## test equivalence of latent response scales, given equivalence of
    configuration, thresholds, loadings, and latent response intercepts
 lavTestLRT(fit.lrintercept.wide,fit.lrscale.wide, method = "satorra. bentler.
    2010")


 # ==================================== #
 # Obtain a summary of all models at once
 fit.comp.wide <- compareFit(fit.config.wide, fit.threshold.wide, fit
    .loading.wide, fit.lrintercept.wide, fit.lrvariance.wide, argsLRT
     = list(method="satorra.bentler.2010"))
summary(fit.comp.wide)


 # ==================================== #
 # Use final model to extract estimates of cross-wave correlations with
    confidence intervals
 summary(fit.lrscale.wide, standardized=T, ci=T)
```