Meta-Psychology, 2022, vol 6, MP.2020.2595,
https://doi.org/10.15626/MP.2020.2595
Article type: Tutorial
Published under the CC-BY4.0 license

Open data: Not Applicable
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Rickard Carlsson
Reviewed by: Sacha Epskamp, Ulrich Schimmack
Analysis reproduced by: Marco Lauriola
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/H9DV5

A Tutorial in Longitudinal Measurement Invariance and Cross-lagged Panel Models Using Lavaan

Sean P. Mackinnon
Dalhousie University

Robin Curtis
Dalhousie University

Roisin M. O'Connor
Concordia University

In longitudinal studies involving multiple latent variables, researchers often seek to predict how iterations of latent variables measured at early time points predict iterations measured at later time points. Cross-lagged panel modeling, a form of structural equation modeling, is a useful way to conceptualize and test these relationships. However, prior to making causal claims, researchers must first ensure that the measured constructs are equivalent between time points. To do this, they test for measurement invariance, constructing and comparing a series of increasingly strict and parsimonious models, each making more constraints across time than the last. This comparison process, though challenging, is an important prerequisite to interpretation of results. Fortunately, testing for measurement invariance in cross-lagged panel models has become easier, thanks to the wide availability of R and its packages. This paper serves as a tutorial in testing for measurement invariance and cross-lagged panel models using the lavaan package. Using real data from an openly available study on perfectionism and drinking problems, we provide a step-by-step guide of how to test for longitudinal measurement invariance, conduct cross-lagged panel models, and interpret the results. Original data source with materials: https://osf.io/gduy4/. Project website with data/syntax for the tutorial: https://osf.io/hwkem/.

Keywords: cross-lagged panel; lavaan; measurement invariance; R; tutorial; perfectionism; social anxiety

The proliferation of R as a free and versatile programming language and analytic tool, coupled with the increasing power of modern computers, has made possible a great range of new statistical tests for students and professionals across varied disciplines. However, the learning curve for R is steep, and many statistical topics are so specialized that they lack coherent step-by-step guides with accompanying syntax. This paper and its accompanying OSF page (https://osf.io/hwkem/) demonstrate the process of selecting, running, and evaluating cross-lagged panel models with the lavaan package in R, using real data from an open-access source. While our target audience for this paper is graduate students with a basic understanding of R, confirmatory factor analysis, and structural equation modelling, we seek to present the material in such a way that parts of it may be useful to researchers at a range of levels. Readers unfamiliar with structural equation modelling might start with Ullman (2006), which is a relatively accessible introduction.

Measurement Invariance

Much of the time in psychology, we do not measure the construct of interest directly, but rather infer it through a series of items associated with it.

Such constructs are called latent variables, such as drinking motives (Mackinnon et al., 2017) and perfectionism (Rice, Loscalzo, Giannini, & Rice, 2020).[1] Confirmatory factor analysis (CFA) is a statistical technique that allows us to test whether clusters of items in our measure are indeed reflective of the latent construct to which we have assigned them.

When studying constructs over time, we administer the same measurement instruments repeatedly. To make logical claims about how latent variables change across time, we must first establish that our instruments are measuring the construct consistently over time. Measurement invariance (MI) is upheld in a study when "participants across all [time periods] interpret the individual questions, as well as the underlying latent factor, in the same way" (Van de Schoot et al, pp. 1 - 2). If MI is not upheld, then the nature of the latent construct changes from over time, making comparisons across measurement occasions difficult. Since the proposal of MI as a concept thirty years ago (Byrne, Shavelson, and Muthen, 1989), researchers have considered MI an important quality to check for in longitudinal studies incorporating latent variables. In simpler terms, a fundamental problem in longitudinal measurement is that the mere passage of time (or the act of observing one's own thoughts through repeated measurement) can sometimes change how people interpret questionnaire items. To make comparisons over time, we want MI to ensure that the nature of the construct has not changed substantially over time.

To use CFA to test for MI, researchers first set up a set of nested model comparisons; essentially, a series of CFA models with increasingly strict constraints on equality over time. More strict and parsimonious models allow fewer parameters to vary over time for the same latent construct. Broadly speaking, the parameters we are concerned with are: 1) factor loadings, which show how representative each item is of its latent factor; 2) intercepts, which relate to the mean levels of each item; and 3) residual variances, which represent the other unexplained influences predicting item responses besides latent variables. The most parsimonious model that maintains adequate CFA fit indices determines the level of invariance. We can further subdivide MI according to four levels (Widaman & Reise, 1997; Widaman et al., 2010). The least stringent level of invariance, referred to as configural invariance, allows factor loadings, item intercepts, and residual variances to vary across waves. This establishes that the same factor structure applies across waves (i.e., the same number of latent variables with the same items loading on each factor). The next level is metric – sometimes called weak[2] – invariance and constrains factor loadings to equality across waves. This establishes that items do not become more (or less) representative of the latent construct at different measurement occasions. That is, as factor loadings get larger, they are stronger indicators of the latent variable; metric invariance proposes that items do not vary in how representative they are of the construct over time. The following level is scalar invariance, and constrains not only factor loadings, but also item intercepts to equality across waves. Constraining item intercepts to equality establishes that the mean levels of the underlying items themselves do not vary significantly between time periods. That is, if scalar invariance is violated it means that the interpretation of the absolute value of a score changes as time goes on. It is analogous to how $100 today does not have the same value as it did 100 years ago. Using a psychological example, a lack of scalar invariance might make it appear that means are increasing over time when it is really that participants are changing how they interpret the response scale over time. The final level is residual invariance, and constrains factor loadings, item intercepts, and residual variances to equality across waves. Residual variance represents the degree to which the model deviated from the actual data due to external factors. Fixing residual item variation to equality across waves therefore establishes that any external factors (e.g., unmeasured variables that are changing over time and predicting variation in the measured latent construct) also display minimal change over time (Van de Schoot et al., 2012).

Why does this matter? Without accounting for MI, researchers may misinterpret the causes behind observed effects. For instance, in an investigation of student narcissism levels between the 1990s to

[1] These two papers also serve as good templates for reporting measurement invariance results for a beginning learner.

[2] Terminology for model strictness can vary. In this paper, we follow Van de Schoot et al. (2012) in using the terms "metric," "scalar," and "residual." On the other hand, Widaman et al. (2012) and Wu et al. (2007) opt for the terms "weak," "strong," and "strict" in referring to the same concepts.

2010s, Wetzel et al. (2017) examined MI to check the validity of prior findings suggesting that narcissism is increasing among today's youth (Twenge & Campbell, 2009). In a three-wave model incorporating responses from more than 50,000 students, Wetzel et al. found nonequivalence in several aspects of the measurement of narcissism on the Narcissistic Personality Inventory. Specifically, facets of leadership and vanity were not invariant, suggesting that students' interpretations of questions pertaining to these aspects were changing over time, a feature of the data not previously acknowledged in Twenge & Campbell's (2009) study. When accounting for this partial nonequivalence, their model actually suggested a decrease in narcissism over time. Checking for MI is thus an important practice prior to interpreting longitudinal results.

Cross-lagged Panel Models

Another important aspect of longitudinal studies are directional effects between variables over time. Researchers can use cross-lagged panel models (CLPM) to investigate how well different variables predict future iterations of each other, helping to make stronger causal claims by establishing temporal precedence (Cole and Maxwell, 2003). Results can sometimes help clarify the direction of relationships in a way cross-sectional correlations cannot. For example, Mackinnon (2012) found a small positive correlation between perceived social support and school grades cross-sectionally; however, a cross-lagged panel model suggested that higher grades led to more social support, rather than the reverse, contrary to common belief. This study used a three-wave design over several years, but diary studies using more waves over shorter time periods are also common (e.g., Sherry & Hall, 2009).

Cross-lagged panel models are one attempt to make stronger causal claims with longitudinal data. However, it is important to note that cross-lagged panel models have been criticized quite early on (e.g., Rogosa, 1980) and more recently by Hamaker, Kuiper, & Grasman (2015). The crux of the criticism is that the traditional cross-lagged panel model does not properly disentangle within-person processes (e.g., state-like, day-to-day change) from between-person processes (e.g., trait-like, stability from day-to-day). As a result, traditional cross-lagged panel models can produce incorrect results for statistical significance, which relationship is larger, and even the sign/direction of the relationship (Hamaker et al., 2015)! As a potential solution to these issues, Hamaker et al. (2015; see also Mulder & Hamaker, 2020 for the extension to multiple indicators, as used in this paper) introduces the random intercepts cross-lagged panel model, which properly accounts for the stable, trait-like nature of many constructs. Thus, though the bulk of the paper will focus on the traditional cross-lagged panel model, we also present code and interpretation for a random intercepts cross-lagged panel model.[3]

MI in Cross-lagged Panel Models

When testing for MI in CLPMs, model complexity and number of participants can also have an impact on interpretations. The greater the number of waves and items, the more complex the model will be, and the greater the number of participants needed to facilitate reliable testing. This is generally true of structural equation models in general (Kyriazos, 2018). The configural model estimates the greatest number of parameters and is thus the least parsimonious model. To simplify the problem of longitudinal studies slightly in each wave, actual scores of participants will differ somewhat from the model's prediction, but the same participants are responding to the measure each time, creating non-independence of observations across waves. We model this non-independence as a covariance between the residuals of the same items among waves (see the annotated syntax file that accompanies this paper). It is a reasonable a-priori assumption to expect the magnitude of covariances between residuals across waves to be similar. It is common for researchers to fix residual covariance to equality across waves.[4] Fixing these values to equality is thus often theoretically

[3] Readers should also note that the random intercepts cross-lagged panel model requires a minimum of three measurement occasions, unlike the traditional cross-lagged panel model, which is identifiable with only two.

[4] Another common set of constraints used in longitudinal data is an first-order autoregressive or AR(1) correlated error structure. That is, a set of constraints that predict the covariances will get smaller as the time lags increase (i.e., constructs measured more closely in time should be more strongly related).

justified and helps reduce the total number of parameters estimated by the configural model (Cole and Maxwell, 2003)[5]. Prior to setting up and testing the final structural equation model, it is useful to map out the predicted relationship between variables as figures.

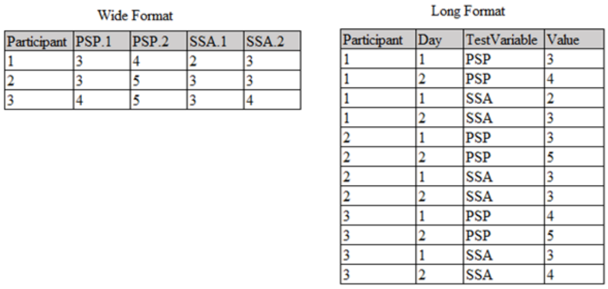
The dataset used in this study derives from a 21-day diary study investigating perfectionism, motivations for drinking, and alcohol-related problems. It is open-access, and all data are free to download at https://osf.io/hwkem/. For the purpose of this paper, our key latent variables of interest are: 1) perfectionistic self-presentation (PSP), which (as operationalized in these data) measures an individual's desire to hide their imperfections; and 2) and state social anxiety (SSA), which measures transitory feelings of anxiety associated social situations. PSP was first proposed an aspect of perfectionism by Hewitt et al. (2003), whereas SSA was proposed as measure of social anxiety by Kashdan and Steger (2006). For a recent study using these data discussing the relation between PSP and SSA in more detail, please see Kehayes and Mackinnon (2019). To make the example easier to follow we focus on only 5 of the 20 days perfectionistic self-presentation and social anxiety were measured (arbitrarily, days 7-11). Thus, in the present example we first (a) establish longitudinal measurement invariance over 5 days and then (b) test a cross-lagged panel model. In general, think of the measurement invariance portion as a necessary first step to proceed with hypothesis testing in the cross-lagged panel model. Though theory suggests that perfectionism would cause social anxiety rather than the reverse, we do not concern ourselves with formally testing confirmatory hypotheses in this paper – even though the cross-lagged panel model would be the spot where hypotheses about directionality of relationships are formally tested in a traditional paper. Instead, we focus on the technical, analytical aspects with the goal of teaching readers how to conduct the analysis.

Method

Dataset

Our study uses a simplified version of the dataset published by Mackinnon et al. (2021). We first trimmed and reformatted the dataset to contain only the variables of interest (see Appendix A). The abridged dataset contains responses given by 251 participants for two latent variables (PSP, composed of three items; and SSA, composed of seven items) across five days (days seven through eleven of the study). PSP items were measured using a 7-point scale from 1 to 7. SSA items were measured using a 5-point scale from 0 to 4.

Beyond trimming the dataset to only the variables of interest, we also converted the data to wide format (in which every participant receives a unique row, and each variable receives a unique column) from long format (in which every data point receives a unique row, with participants and categorical variables recurring across rows) For a simple illustration of this pertaining to SSA and PSP values across days, see Figure 1. This format conversion aided in setting up the code for our later models.



Wide Format

Participant	PSP.1	PSP.2	SSA.1	SSA.2
1	3	4	2	3
2	3	5	3	3
3	4	5	3	4

Long Format

Participant	Day	TestVariable	Value
1	1	PSP	3
1	2	PSP	4
1	1	SSA	2
1	2	SSA	3
2	1	PSP	3
2	2	PSP	5
2	1	SSA	3
2	2	SSA	3
3	1	PSP	4
3	2	PSP	5
3	1	SSA	3
3	2	SSA	4

Figure 1. Examples of Wide Vs. Long Format in PSP and SSA Values Across Two Days

Data Analysis Strategy

This section describes our strategy for comparing and selecting models. Our goal was to compare nested versions of our CLPM, using CFA, to determine the most appropriate parameters for our final structural model. We sought to use the simplest model (i.e., model estimating the fewest parameters) that maintained a good fit for our data, while also

[5] Note that measurement invariance between independent groups using multi-group modelling (e.g., comparing men, women, and nonbinary groups) is comparatively simpler than measurement invariance in the longitudinal context because of this correlated error structure. Readers interested in measurement invariance for independent groups can take advantage of the measurementInvariance() function in lavaan for convenience functions that are much shorter than the code in this tutorial (https://lavaan.ugent.be/tutorial/groups.html), even though the core principles are the same.

making good theoretical sense. Because lavaan allows undefined parameters to vary freely by default, simpler models are those with more constraints defined; thus, somewhat counterintuitively, more parsimonious models appear more complicated in the code. In constructing our models through lavaan, we relied upon four key operators: 1) =~ , which is used for factor loadings, and can be thought of as “is measured by;” 2) ~ , which is used for regression formulas, and can be thought of as “is regressed on;” 3) ~~ , which is used for defining variance and residual covariance, and can be thought of as “varies with;” and 4) ~ 1, which is a special notation for defining intercepts. Within any given formula, labels may be assigned to terms using the asterisk (*). Any items to which the same label is applied are fixed to equality in the model’s calculations. To better understand the full definitions of each model below, we recommend referring to Figures 2-5, which show simplified versions of each model with their respective constraints. It may also be useful to simultaneously follow along with our annotated code on our OSF page.

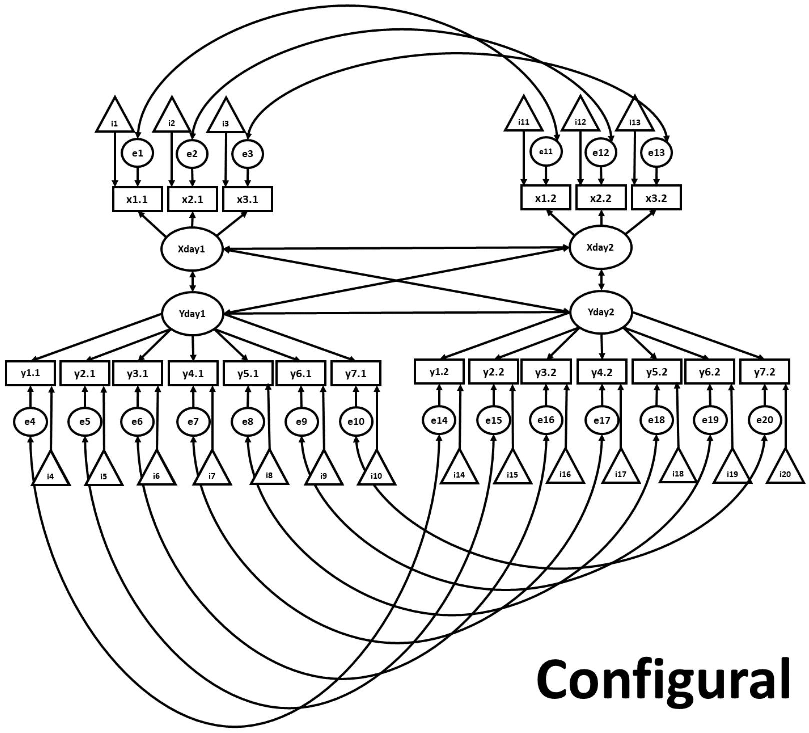
Our approach involves five steps total: (a) configural model; (b) metric model; (c) scalar model; (d) residual model; and (e) structural model. Notably, in the first four steps we use covariances (~~) rather than regressions (~) for relationships between variables. In the fifth step, we do a true cross-lagged panel model where temporal directionality of relationships is assumed (e.g., day 7 predicting day 8). This choice has no impact on the overall fit indices, but it is worth noting that the relationships between variables in the first four steps are more akin to bivariate correlations (albeit, corrected for measurement unreliability), while the last step is the true test of hypotheses with paths allowing for stronger causal inferences than correlations by adjusting for past-day levels of each variable.

Configural Model. We began by defining our configural model. Excluding the correlated error structure, the full configural model is as follows:

```
configural.v1 <-
'
# PSP factor loadings defined

PSP.7 =~ NA*psp1.7 + psp2.7 + psp3.7
PSP.8 =~ NA*psp1.8 + psp2.8 + psp3.8
PSP.9 =~ NA*psp1.9 + psp2.9 + psp3.9
PSP.10 =~ NA*psp1.10 + psp2.10 + psp3.10
PSP.11 =~ NA*psp1.11 + psp2.11 + psp3.11

# PSP variance constrained to 1

PSP.7 ~~ 1*PSP.7
PSP.8 ~~ 1*PSP.8
PSP.9 ~~ 1*PSP.9
PSP.10 ~~ 1*PSP.10
PSP.11 ~~ 1*PSP.11

# SSA factor loadings defined

SSA.7 =~ NA*ssa1.7 + ssa2.7 + ssa3.7 + ssa4.7 + ssa5.7 +
ssa6.7 + ssa7.7
SSA.8 =~ NA*ssa1.8 + ssa2.8 + ssa3.8 + ssa4.8 + ssa5.8
+ ssa6.8 + ssa7.8
SSA.9 =~ NA*ssa1.9 + ssa2.9 + ssa3.9 + ssa4.9 + ssa5.9
+ ssa6.9 + ssa7.9
SSA.10 =~ NA*ssa1.10 + ssa2.10 + ssa3.10 + ssa4.10 +
ssa5.10 + ssa6.10 + ssa7.10
SSA.11 =~ NA*ssa1.11 + ssa2.11 + ssa3.11 + ssa4.11 +
ssa5.11 + ssa6.11 + ssa7.11

# SSA variance constrained to 1

SSA.7 ~~ 1*SSA.7
SSA.8 ~~ 1*SSA.8
SSA.9 ~~ 1*SSA.9
SSA.10 ~~ 1*SSA.10
SSA.11 ~~ 1*SSA.11

'

configural.model <- paste(configural.v1, errorstruc-
ture, sep = ' ', collapse = NULL)
```

Figure2

Configural Model (2 days only).



Note. Ovals represent the latent variables. Rectangles labeled represent the measured items, with their corresponding factor loadings shown as arrows to the latent variables. Triangles represent the item intercepts, and circles represent the residuals. Double-headed arrows represent covariances. In the configural model, we allowed factor loadings, item intercepts, and residual variances to vary freely across waves. The model tested in the tutorial uses 5 days of data; to reduce visual clutter, this diagram shows only the first two days of data.

For this and all subsequent models, we first defined the factor loadings of our latent variables. By default in lavaan, the first factor loading for each latent variable would normally be constrained to 1, whereas the variance of each latent variable would be unconstrained. However, as a matter of preference, we invert these settings, allowing each latent variable's first factor loading to vary freely, achieved through the "NA*" in the code above, while constraining the variance of each latent variable to 1, achieved through the addition of the code below. By overriding this default setting, we can more easily constrain the factor loadings to equality in later models.

Note that in the code above, we use the ~~ operator to correlate PSP.7 with itself, which effectively defines its variance[6], which we set to 1 with the operator 1*. Though not shown in the code snippet above, we also constrained our residual covariances to equality across waves, as justified in the previous section. We achieved this by assigning unique labels (e.g., "psp1cov") to the defined covariance parameters of each residual across all days in the study, for example:

```
# Here "psp1cov" is a label, while "psp1.7" and
"psp1.8" are variables.
psp1.7 ~~ psp1cov*psp1.8
```

[6] This seems unintuitive but makes sense when you consider that lavaan works with variance-covariance matrices, which places variances on the diagonal. Thus, "covarying something with itself" refers to the cell on the diagonal, or the variance for that variable.

The final part of the code using “paste” merely combines two vectors of characters together (i.e., the error structure and the configural model). Because the correlated error structure is 108 lines of code, it would be tedious to re-write this every time. Thus, we instead save the error structure as an object named “errorstructure” and append it to each model to make the code shorter.

Having defined our model, we used it to run our first CFA using the code below:

```
configural.fit <- cfa(configural.model,
                      data = model.test.dat,
                      estimator = "MLR",
                      se = "robust",
                      missing = "ML",
                      std.lv = TRUE)
```

Note that for this and all future analyses, we added the line “std.lv = TRUE.” This automatically fixes the variance of factors to 1, rather than their factor loadings, which is the default setting in lavaan. Because we will be predicting good model fits after imposing equality of factor loadings in our testing of measurement invariance, it is better to fix factor variance to 1 in this manner. As for the other code statements: “data = model.test.dat” calls our dataset for use in the model; “estimator = ‘MLR’” sets our model’s estimation method to maximum likelihood estimation with robust standard errors; “se = ‘robust’” similarly implements robust standard errors in the estimation; “missing = ‘ML’” implements full information maximum likelihood estimation for missing data. Standard practice for comparing nested models dictates that investigators establish increasingly strict levels of invariance, stopping once a model fails to display adequate fit criteria. In our Results section, we discuss our model fit indices in more detail. Within this section, however, we will simply acknowledge whether fit criteria were met and move on to test all 4 levels of invariance for pedagogical purposes.

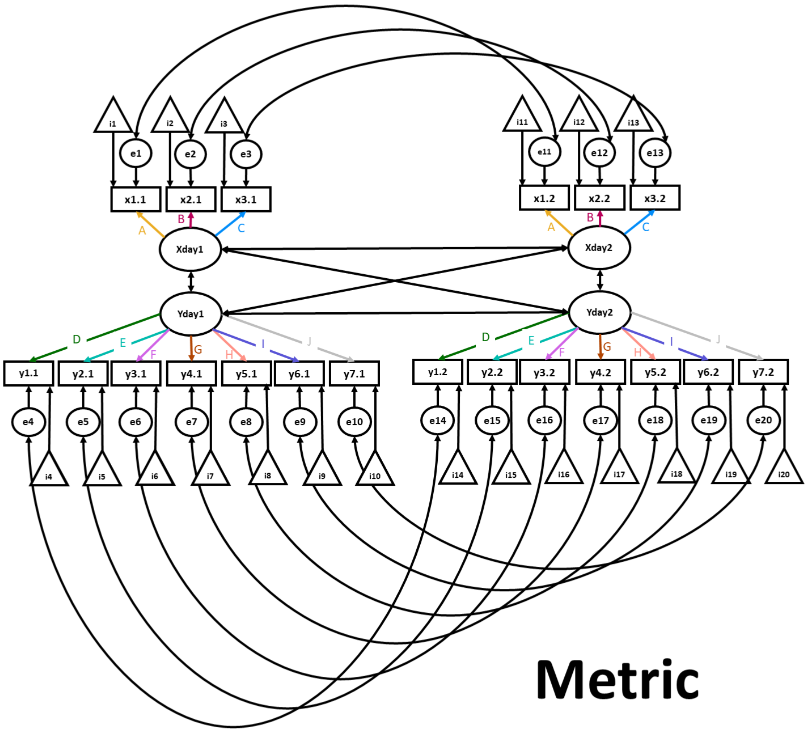
Metric Model. Having found evidence for configural invariance in our latent variables, we moved on to check for metric invariance. In this model, in addition to the constraints applied to our configural model, we constrained factor loadings to equality across waves. We achieved this by assigning unique labels (e.g., “psp1f*”) to all five iterations of each factor (e.g., “psp1.7”, “psp1.8”, etc.) within our factor loading definition section, for example:

```
# Here “psp1f”, “psp2f”, and “psp3f” are labels, while
the terms to the right of these labels are variables.
PSP.7 =~ psp1f*psp1.7 + psp2f*psp2.7 + psp3f*psp3.7
PSP.8 =~ psp1f*psp1.8 + psp2f*psp2.8 + psp3f*psp3.8
```

In the code snippet above, you might read it as: “PSP items psp1.7, psp2.7 and psp3.7 (i.e., items from day 7) all load on the latent factor called PSP.7. PSP items psp1.8, psp2.8 and psp3.8 (i.e., items from day 8) all load on the latent factor called PSP.8. The factor loadings for psp item 1 at both day 7 and day8 is given the label ‘psp1f’. Since they have the same label, they are constrained to be equal to each other.”

Figure 3

Metric Model (2 days only).

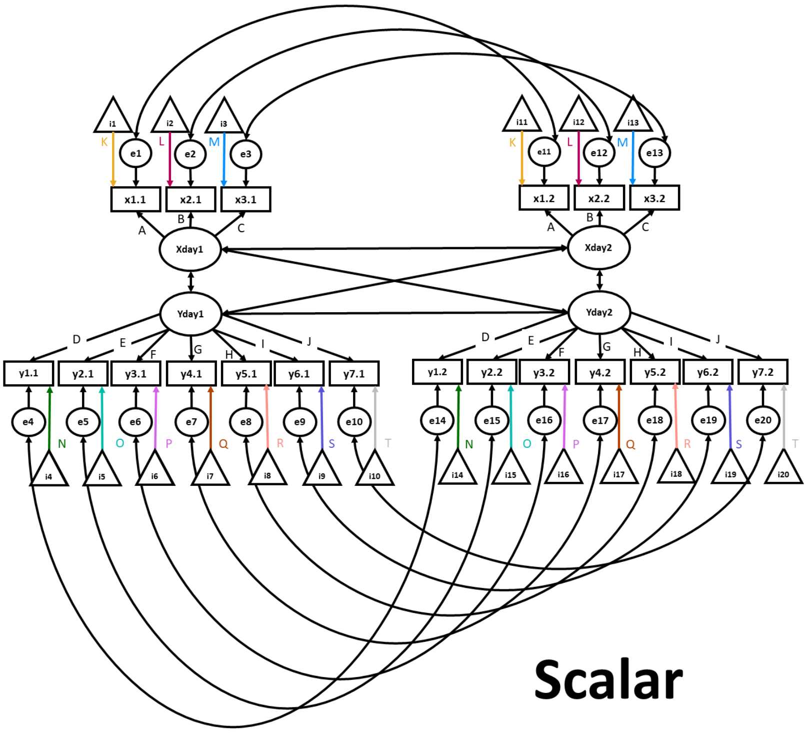


Note. In our metric model, we fixed factor loadings to equality across waves, as shown by letters A – J (and matching colors) in this simplified diagram. Factor loadings that share the same letter and color are constrained to equality. The model tested in the tutorial uses 5 days of data; to reduce visual clutter, this diagram shows only the first two days of data.

Scalar Model. We then moved on to check for scalar invariance. In this model, in addition to the constraints applied to our metric model, we fixed item intercepts to equivalence across waves. We achieved this by applying a single label (e.g., psp1i) to all five iterations of an item's intercept. For example:

```
# Here psp1i is a label, while the 1 refers to the intercept of the variable on the left side of the ~.
psp1.7 ~ psp1i*1
psp1.8 ~ psp1i*1
```

Figure 4
Scalar Model (2 days only).



Note. In our scalar model, we additionally fixed item intercepts to equality across waves, as shown by letters K-T (and matching colors) in this diagram. Intercepts that share a letter and color are constrained to equality. The model tested in the tutorial uses 5 days of data; to reduce visual clutter, this diagram shows only the first two days of data.

Residual Model. Finally, we applied our most rigorous set of constraints in our residual model. In this model, in addition to the constraints applied to our scalar model, we constrained the residual error of each factor to equality across waves. Once again, we did this by applying the same label (e.g., “psp1u”) to each iteration of a given residual across waves. For example:

```
# Here “psp1u” is a label, while “psp1.8” is a variable.
psp1.7 ~~ psp1u*psp1.7
psp1.8 ~~ psp1u*psp1.8
```

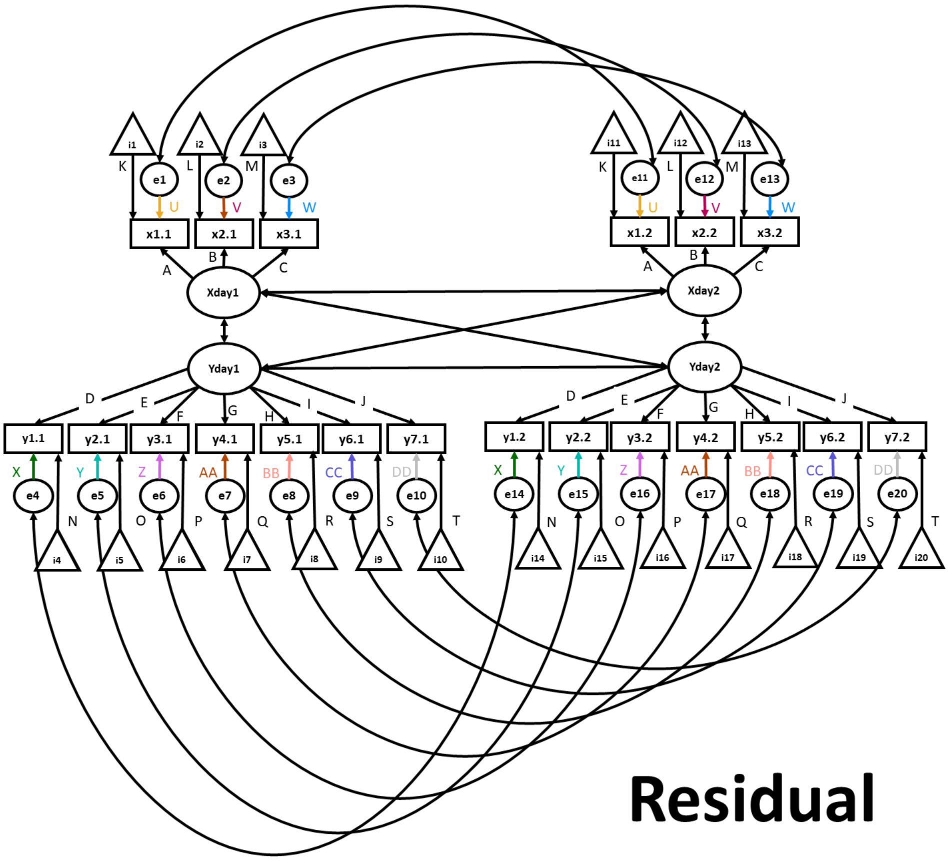
Model Selection. Having thus defined our nested models, we next compare all of their fit indices at once and select the best candidate for our subsequent structural equation model (SEM). We grouped the models and compared their fit indices using the code below:

```
round(cbind(configural.error=inspect(configural.fit,
'fit.measures'),
           metric=inspect(metric.fit, 'fit.measures'),
          scalar=inspect(scalar.fit, 'fit.measures'),
           residual=inspect(residual.fit,
'fit.measures')),3)

anova(configural.fit, metric.fit)
anova(metric.fit, scalar.fit)
anova(scalar.fit, residual.fit)
```

Here, the inspect() function extracts the fit indices from the specified models, the cbind() function places each extracted item into a dataframe, and the round(...,3) function rounds all values to three decimal places. The anova() functions are a shorter version to get key statistics: AIC, BIC, and a chi-squared difference test. We discuss the fit indices and selection process in the next section. After selecting our residual model as our preferred choice, we set up and ran our SEM. In addition to the constraints of our residual model, our SEM defined regression paths between our latent variables, using simple labels to fix them across days (see Figure 6).

```
SSA.8 ~ A*SSA.7 + C*PSP.7
PSP.8 ~ D*SSA.7 + B*PSP.7
```

Figure 5
Residual Model (2 days only).



Note. In our residual model, we additionally fixed residual variances to equality across waves, as shown by letters U-DD (and matching colors) in this diagram. Residual variances that share a letter and color are constrained to equality. The model tested in the tutorial uses 5 days of data; to reduce visual clutter, this diagram shows only the first two days of data.

Results

Fit Indices

The fit indices[7] for each of our four models (configural, metric, scalar, and residual) are shown in Table 1. Number of estimated parameters is self-explanatory. For the sake of model parsimony, we sought to estimate the fewest possible number of parameters, while still using a model that maintains a good fit for our data.

The Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) compare the given proposed model to the null model. Both prioritize the average correlation of variables within the proposed model, and their results are highly correlated. However, TLI places an additional emphasis on model parsimony, implementing a modest penalty for each additional estimated parameter. These fit indices are negatively biased, meaning that the resultant value for poor-fitting models tends to be lower. For both CFI and TLI, values between .90 and .95 are considered marginally acceptable, whereas values above .95 are considered good. Cheung & Rensvold (2002) suggest that a ΔCFI of -.01 or more suggests that the model with the largest CFI should be chosen; otherwise, prefer the more parsimonious model. Using this criterion, the residual model would be preferred.

On the other hand, two popular positively biased (meaning the lower the value, the better the fit) measures of fit are the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR). The RMSEA applies a penalty for model complexity according to the ratio between the chi-square and degrees of freedom. Models with fewer degrees of freedom are penalized more strongly, and values of .06 to .08 or lower generally indicate an acceptable level of fit. SRMR, meanwhile, calculates the difference between the observed correlation and predicted correlation, with no penalty for model complexity. As with RMSEA, SRMR values should be no greater than .08. To our knowledge, these are not commonly used for nested model comparisons, though it is important than any final selected model still fit well by this criterion. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are comparative fit measures. They provide an independent value for assessing each model's fit, which makes most sense when compared between models. The lower the value, the better the fit, with differences of 6 or greater typically constituting strong evidence of model difference (Raftery, 1995). Both use a likelihood function in conjunction with the number of estimated parameters and the sample size to assess model fit, with BIC implementing a slightly higher penalty for more complex models. Due to this difference, AIC values preferred our scalar model, whereas BIC values preferred our residual model in Table 1.

Log-likelihood ratio tests (sometimes called "chi-squared difference tests") are also a popular way to compare models, though they tend to be overly sensitive in the same ways the chi-squared test for model fit is. The log-likelihood ratio test compares the loglikelihood of two models and produces a test statistic that has a χ^2 distribution when the null hypothesis is true. A statistically significant p-value indicates that the less parsimonious model should be chosen. A non-significant p-value suggests that we should default to the most parsimonious model. In Table 1, the loglikelihood ratio test prefers the metric model.

Model Comparison

As shown in Table 1, the fit indices for each of our models were generally favorable, but due to the differences in index calculations just described, different indices preferred different models. In case of such potential conflicts, we therefore recommend the researcher decide which model aspect they deem most important (e.g., goodness of fit, number of estimated parameters, sample size), and decide a priori which fit index to use to determine their final model. For a more in-depth (but challenging) review of these issues, see Lin, Huang, & Weng (2017). We tend to prefer ΔBIC as an a-priori criterion, but there is a great deal of variability among different analysts on this point. Failing this, you might choose the model recommended most often out of the 4 choices (loglikelihood ratio, AIC, ΔBIC, ΔCFI). While we are not formally hypothesis testing in this paper, we chose to continue with the model with the best

[7] This paper does not seek to fully explain the mathematics behind each fit index, but rather to give a brief overview of the features prioritized by each. For a more substantive description of fit indices, and for a list of further reading, please see Kenny (2015).

BIC score (the residual model). Given that it was not significantly worse than our scalar model by most common metrics, our residual represented the best trade-off between model fit and model complexity. Factor loadings, unstandardized intercepts, and unstandardized residual variances for each model are provided in the Appendix, within Tables A1, A2, and A3, respectively.

Table 1
Nested Model Fit Indices

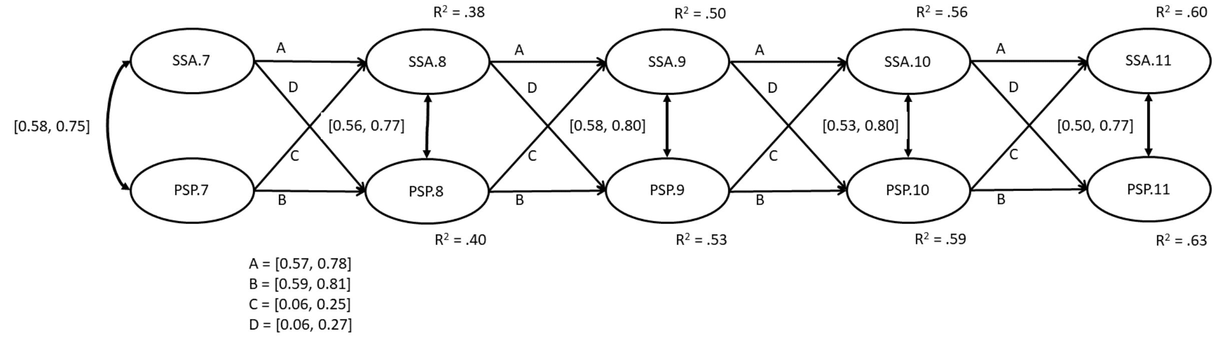
	Model			
	Configural	Metric	Scalar	Residual
No. of Estimated Parameters	205	165	125	85
Raw Loglikelihood	-12632	-12651	-12685	-12741
Δ χ2, p-value	N/A	**54.829, .059**	67.585, .004	67.151, .005
Robust CFI	0.953	0.952	0.950	**0.946**
Raw AIC	25674	25633	25619	25652
Δ AIC	N/A	-41	**-14**	+33
Raw BIC	26397	26215	26060	25951
Δ BIC	N/A	-182	-155	**-109**
Robust RMSEA	0.048	0.048	0.048	0.049
SRMR	0.054	0.058	0.059	0.059

Note. Bold font indicates preferred model/s according to index. Chi-square difference test (i.e., log likelihood ratio tests) values are calculated between the current column model and the preceding column model – e.g., the value listed for the metric model column compares the metric model to the configural model. AIC and BIC are calculated as the difference from the preceding model – e.g., the residual model AIC is 33 greater than the scalar model AIC, while the scalar model is 14 less than metric AIC.

Structural Model for Traditional CLPM

Using the constraints defined in our residual model, we conclude by testing the structural model. In this model, we estimate regressions instead of covariances. We also include only relationships from time t to time t+1. That is, relationships from time t to t+2, t+3 and so on are omitted. One may wish to include these as well, but for pedagogical purposes here we will focus only on these pathways and will constrain relationships between variables at time lags greater than +1 to be zero. This model fit the data ok, but the model fit is worse relative to the CFA model, χ2(1276) = 2139, p < .001, Robust CFI = 0.93, Robust TLI = 0.93, RMSEA = 0.05, SRMR = 0.23. Notably, the SRMR index suggests considerably worse fit than the other measures here. This probably means we should investigate whether more paths should be added (e.g., the cross-lagged paths from time t to time t+1). For brevity's sake in this tutorial, we do not explore this further (as this is an idiosyncratic feature of this sample dataset, rather than the general approach most people would be taking), but it is worth mentioning – it probably means that the relationships of interest are not confined to a 1-day time lag, and that they can influence relationships on a longer time frame. This model had R^2 values ranging from .38 to .63. Figure 6 shows our latent variables across the five days, and 95% confidence intervals for the path values. As shown in the figure, previous-day perfectionistic self-presentation strongly predicted next-day perfectionistic self-presentation. Similarly, previous-day social anxiety strongly predicted next-day social anxiety. Meanwhile, each latent variable's cross-lagged path predicting the opposite variable was weaker but still non-zero and statistically significant. This kind of pattern (large autoregressive paths, small cross-lagged paths) is typical of most cross-lagged panel models with adequate statistical power.

Figure 6
Structural Model for the Traditional Cross-lagged Panel Model.



Note. Day-to-day paths were fixed to equality. All paths and covariances are shown as unstandardized 95% confidence intervals. Much is omitted from this diagram, including the factor loadings and correlated error structure that is depicted in Figures 2-5. All pathways are statistically significant at p < .05.

Random Intercepts Cross-Lagged Panel Model

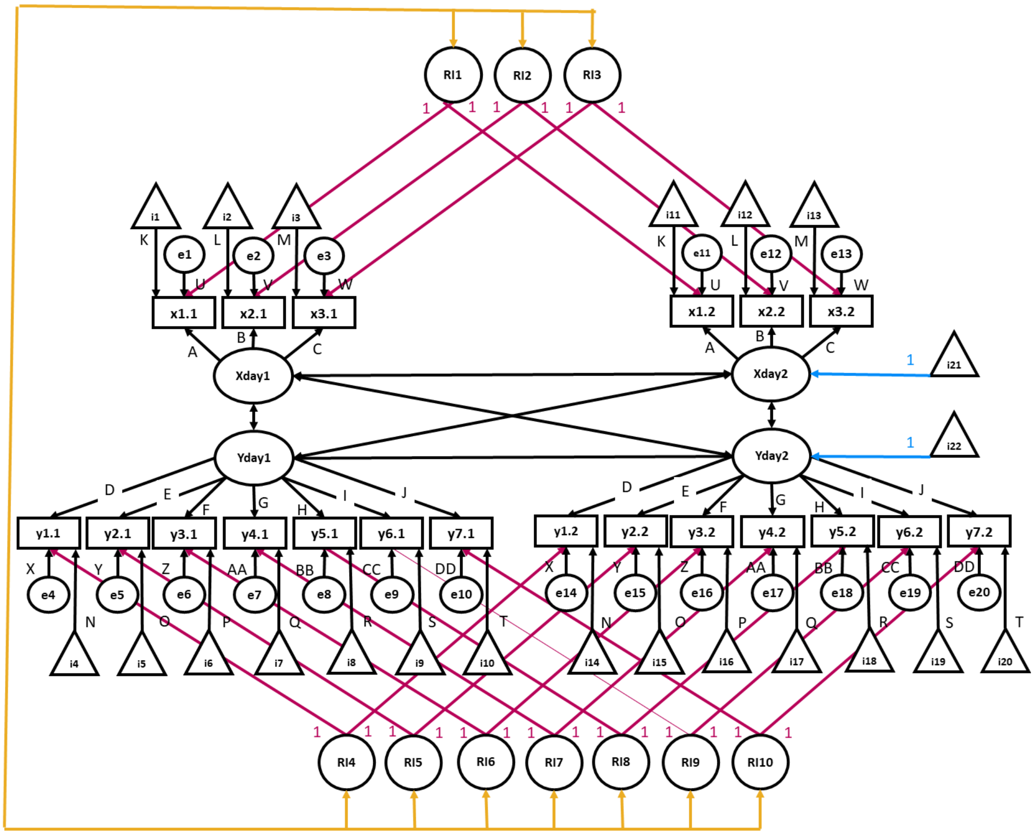
Though the above summarizes the traditional cross-lagged panel model approach, we will also briefly review the multiple indicator random intercepts cross-lagged panel model described by Mulder & Hamaker (2020). This model makes four primary changes: (a) the correlated residual structure is omitted; (b) random intercepts are specified for each indicator; (c) the random intercepts are allowed to covary together; (d) the latent means for all latent variables from the second time point onwards are freed[8]. See Figure 7 for a visual depiction of this model. Please note that the RI-CLPM can only be estimated with 3 or more time points; the diagram in Figure 7 is for pedagogical purposes and presents only two time points due to the practical requirements of creating an image with a font size large enough to read! See the online supplementary materials for the statistical syntax of the full model.

Excepting the chi-squared statistic and SRMR, this model fit the data reasonably well, χ2(1225) = 1884, p < .001, Robust CFI = 0.94, Robust TLI = 0.94, RMSEA = 0.05, SRMR = 0.16. Figure 8 shows the 95% CIs for the path coefficients of interest. Here, we have a different pattern of results. The auto-regressive paths now represent within-person carry-over effects (Mulder & Hamaker, 2020). For perfectionistic self-presentation, when people experience elevated perfectionism scores relative to their own expected score, they are more likely to experience elevated perfectionism scores relative to their own expected score at the next occasion as well. In comparison, within-person carry-over effects were inconclusive (i.e., non-significant) for social anxiety. Moreover, the cross-lagged paths now suggest that perfectionistic self-presentation predicts increases in social anxiety over time, but not the reverse. Interestingly, this result is more in line with theoretical predictions in this research area.

[8] Mulder & Hamaker (2020) note "…if we would not freely estimate the latent means, we would not only specify strong factorial invariance, but also specify a model in which there cannot be mean changes over time" (pg. 8).

Figure 7

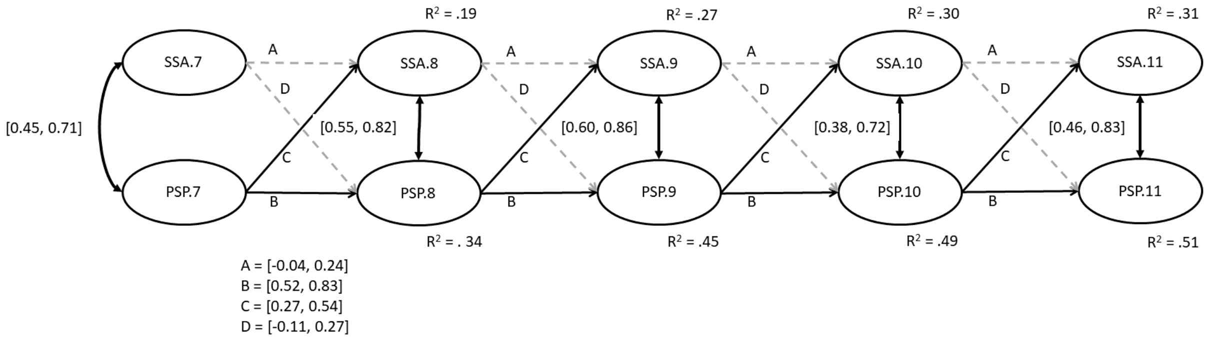
The Random-Intercepts Cross-Lagged Panel Model (2 days only).



Random Intercepts Residual

Note. This model adds random intercepts (RI1-RI10; red), covariances between the random intercepts (orange), and frees the latent means from the second time point onwards (i21 and i22; blue). Please note that the RI-CLPM can only be estimated with 3 or more time points; this diagram is for pedagogical purposes and presents only two time points to ensure the font size is still legible.

Figure 8
Coefficients of Interest for the Random Intercepts Cross-Lagged Panel Model.



Note. Day-to-day paths were fixed to equality. All paths and covariances are shown as unstandardized 95% confidence intervals. Much is omitted from this diagram, including the factor loadings and random intercepts that are depicted in Figure 7. Black solid lines are statistically significant at p < .05. Grey dotted lines are non-significant p > .05.

Discussion

To re-cap, we used confirmatory factor analysis to test two related variables for measurement invariance across five waves, using four increasingly restrictive nested models: configural, metric, scalar, and residual. After selecting our residual model as the simplest that maintained good fit, we applied its constraints to a structural equation model, allowing us to quantify the cross-lagged relationships of our two variables across days.

Interestingly, not all our fit indices preferred the same model. The loglikelihood ratio tests preferred the metric model, as this test tends to prefer less parsimonious models. Meanwhile, our CFI preferred the residual model based on Cheung & Rensvold's (2002) criteria. AIC preferred the scalar model, due to its higher parsimony coupled with good overall fit. On the other hand, BIC, our final deciding criterion, placed a higher emphasis on parsimony, and therefore preferred our residual model. There are two main take-away points here. First, the relatively good fits of all of our models across indices suggest that our theoretical reasons for investigating this relationship between variables were most likely sound. Second, while researchers may wish to report multiple fit indices in their papers, it is important that they decide beforehand which index they will use when determining their final model. It is worth noting that the structural models had poor fit based on the SRMR index, which is probably due to constraining some paths to zero that still evince a positive relationship. If this were an a-priori criterion used for assessing model fit in a research paper, you would need to investigate the source of this misfit further. Remember, SRMR differs from RMSEA insomuch as it has no penalty for model complexity; to the extent that you value model parsimony in model selection, you might prefer to use RMSEA as your selection tool instead.

In our case, the final model conformed to the strictest level of measurement invariance. However, in some studies, this will not be the case. Scalar invariance is typically sufficient for general data analysis, as it indicates that participants do not vary greatly across waves in the ways they interpret and answer questions. However, should only metric invariance be upheld, researchers must qualify any subsequent results by acknowledging that, although the latent factors are loaded similarly across waves, individual interpretations of the items may change over time. For more examples of this, see Steenkamp & Baumgartner (1998). It is also worth noting that this method for testing MI in CLPMs works best on models with a low to moderate number of waves. For a 20-day diary study, it may be more pragmatic for researchers to instead implement multi-level modelling techniques or multilevel structural equation modelling.

Though the present tutorial dataset is ill-suited for examining differences in latent means, it is worth noting that scalar invariance is often a preliminary step towards examining differences between means. That is, most substantive research questions on longitudinal data are not about the measurement invariance per se, but rather about regression/covariance and mean differences over time. Though longitudinal latent mean differences are beyond the scope of this tutorial, readers interested in learning more might read Bishop, Christian & Cole (2015) for three approaches for modelling latent growth curves with multiple indicators. Moreover, Breitsohl (2019) is an excellent tutorial for converting common experimental designs (e.g., ANOVA) to SEM frameworks.

Author Contact

Correspondence concerning this article should be addressed to Sean P. Mackinnon. Email: mackinnon.sean@dal.ca.

http://orcid.org/0000-0003-0921-9589

Conflict of Interest and Funding

The authors have no conflict of interest to declare. The data used for this tutorial was collected with the help of a Social Sciences and Humanities Research Council Insight Development Grant [#430-2016-00805].

Author Contributions

Authors are in order of most to least contribution.

Sean Mackinnon took the lead role in conceptualizing the tutorial, supervised Robin's work as part of his Ph.D. comprehensives, edited the first draft manuscript and code, wrote original sections, created the figures, analyzed the data with the Random Intercepts CLPM.

Robin Curtis created the majority of the R Syntax and online appendices for our OSF page, excluding the Random Intercepts model. He also took a lead role in writing the first draft of the manuscript.

Roisin O'Connor assisted with editing the writing in the manuscript and editing/reviewing the tutorial materials.

Open Science Practices

This article earned the Open Materials badge for making the materials openly available. It is a tutorial that used data from a published study, and as such has no (new) collected data. It was not pre-registered. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Bishop, J., Geiser, C., & Cole, D. A. (2015). Modeling latent growth with multiple indicators: A comparison of three approaches. *Psychological Methods*, *20*(1), 43-62. https://doi.org/10.1037/met0000018

Breitsohl, H. (2019). Beyond ANOVA: An introduction to structural equation models for experimental designs. *Organizational Research Methods*, *22*(3), 649-677. https://doi.org/10.1177/1094428118754988

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

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

Cole, D., & Maxwell, S. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 558-577.

https://doi.org/10.1037/0021-843X.112.4.558

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20(1),* 102-116. http://dx.doi.org/10.1037/a0038889

Hewitt, P. L., Flett, G. L., Sherry, S. B., Habke, M., Parkin, M., Lam, R. W., ... Stein, M. B. (2003). The interpersonal expression of perfection: Perfectionistic self-presentation and psychological distress. *Journal of Personality and Social Psychology*, *84(6)*, 1303–1325. 10.1037/0022-3514.84.6.1303.

Kashdan, T., & Steger, M. (2006). Expanding the topography of social anxiety: An experience-sampling assessment of positive emotions, positive events, and emotion suppression. *Psychological Science*, *17*(2), 120-128. https://doi.org/10.1111/j.1467-9280.2006.01674.x

Kehayes, I.-L. L., & Mackinnon, S. P. (2019). Investigating the relationship between perfectionistic self-presentation and social anxiety using daily diary methods: A replication. *Collabra: Psychology*, *5*(1), 33. https://doi.org/10.1525/collabra.257

Kenny, D. A. (2015). Measuring Model Fit. *http://www.davidakenny.net/cm/fit.htm*

Kyriazos, T. (2018) Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *9,* 2207-2230. https://doi.org/10.4236/psych.2018.98126.

Lin, L. C., Huang, P. H., & Weng, L. J. (2017). Selecting path models in SEM: A comparison of model selection criteria. *Structural Equation Modeling*, *24*, 855-869. https://doi.org/10.1080/10705511.2017.1363652

Mackinnon, S. (2012). Perceived social support and academic achievement: Cross-lagged panel and bivariate growth curve analyses. *Journal of Youth and Adolescence*, *41*(4), 474-485. https://doi.org/10.1007/s10964-011-9691-1

Mackinnon, S. P., Couture, M-E., Cooper, M. L., Kuntsche, E., O'Connor, R. M., Stewart, S. H., & the DRINC Team. (2017). Cross-cultural comparisons of drinking motives in 10 countries: Data from the DRINC Project. *Drug and Alcohol Review, 36, 721-730.* https://doi.org/10.1111/dar.12464.

Mackinnon, S. P., Ray, C. M., Firth, S. M., & O'Connor, R. M. (2021). Data from "Perfectionism, Negative Motives for Drinking, and Alcohol-Related Problems: A 21-day Diary Study". Journal of Open Psychology Data, 9: 1, pp. 1–6. doi: https://doi.org/10.5334/jopd.44

Mulder, J. D., & Hamaker, E. L. (2020). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling*, 1-11. https://doi.org/10.1080/10705511.2020.1784738

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163. https://doi.org/10.2307/271063

Rice, S. P., Loscalzo, Y., Giannini, M., & Rice, K. G. (2020). Perfectionism in Italy and the USA: Measurement invariance and implications for cross-cultural assessment. *European Journal of Psychological Assessment*, 36, 207-211. https://doi.org/10.1027/1015-5759/a000476

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88(2),* 245–258. https://doi.org/10.1037/0033-2909.88.2.245

Sherry, S., & Hall, P. (2009). The perfectionism model of binge eating: Tests of an integrative model. *Journal of Personality and Social Psychology, 96(3),* 690-709. https://doi.org/1010.1037/a0014528

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national Consumer Research. *Journal of Consumer Research*, *25(1),* 78-107. https://doi.org/10.1086/209528

Twenge, J. M., & Campbell, W. K. (2009). The narcissism epidemic: Living in the age of entitlement. New York, NY: Atria.

Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*(1), 35-50.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9(4),* 486-492. https://doi.org/10.1080/17405629.2012.686740

Wetzel, E., Brown, A., Hill, P., Chung, J., Robins, R., & Roberts, B. (2017). The narcissism epidemic is dead; long live the narcissism epidemic. *Psychological Science*, *28*(12), 1833-1847. https://doi.org/10.1177/0956797617724208

Widaman, K., Ferrer, E., & Conger, R. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

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.

Appendix A: Supplementary Tables

Table A1

Factor Loadings

	Configural	Metric	Scalar	Residual
psp1	0.84 - 0.88 (1.62 - 1.75)	0.86 (1.68)	0.63 (1.68)	0.86 (1.68)
psp2	0.93 - 0.96 (1.89 - 2.08)	0.95 (1.99)	0.90 (1.99)	0.95 (1.98)
psp3	0.89 - 0.92 (1.73 - 1.96)	0.91 (1.87)	0.89 (1.87)	0.91 (1.86)
ssa1	0.85 - 0.90 (1.01 - 1.06)	0.87 (1.04)	0.87 (1.04)	0.87 (1.04)
ssa2	0.87 - 0.92 (1.11 - 1.18)	0.90 (1.14)	0.90 (1.14)	0.90 (1.14)
ssa3	0.89 - 0.93 (1.11 - 1.21)	0.91 (1.16)	0.90 (1.15)	0.90 (1.15)
ssa4	0.85 - 0.94 (1.03 - 1.12)	0.90 (1.13)	0.90 (1.13)	0.90 (1.13)
ssa5	0.85 - 0.91 (1.06 - 1.15)	0.88 (1.12)	0.88 (1.12)	0.88 (1.11)
ssa6	0.72 - 0.81 (0.87 - 0.99)	0.75 (0.93)	0.75 (0.93)	0.75 (0.92)
ssa7	0.60 - 0.67 (0.64 - 0.75)	0.62 (0.68)	0.62 (0.68)	0.63 (0.68)

Note. Values are formatted as “standardized (unstandardized).” Value ranges (min-max) are provided for the configural model because factor loadings varied by day. For the latter three models, factor loadings were constrained to equality across days. However, for the metric and scalar models, standardized factor loading scores still fluctuated very slightly across days, because variances differed across days. This will be typical in most real data. Therefore, arithmetic mean values are provided for standardized scores in the metric and scalar models, just for ease of presentation in the table. In the residual model, residual error was also constrained to equality across days, which resulted in no fluctuation of standardized scores.

Table A2

Unstandardized Intercepts

	Configural	Metric	Scalar	Residual
psp1	3.63 - 3.94	3.63 - 3.94	3.75	3.76
psp2	3.49 - 3.74	3.49 - 3.75	3.64	3.64
psp3	3.28 - 3.66	3.28 - 3.66	3.47	3.48
ssa1	1.53 - 1.64	1.53 - 1.64	1.60	1.61
ssa2	1.39 - 1.56	1.39 - 1.56	1.52	1.52
ssa3	1.31 - 1.53	1.31 - 1.53	1.44	1.44
ssa4	1.40 - 1.64	1.40 - 1.64	1.54	1.54
ssa5	1.31 - 1.47	1.32 - 1.47	1.41	1.41
ssa6	1.09 - 1.23	1.09 - 1.23	1.16	1.16
ssa7	0.97 - 1.10	0.97 - 1.10	1.01	1.02

Note. Ranges are provided for the configural and metric models, for which intercepts varied by day. For the latter two models, intercepts were fixed across days.

Table A3
Unstandardized Residual Variances.

	Model			
	Configural	Metric	Scalar	Residual
psp1	0.89 - 1.11	0.87 - 1.11	0.89 - 1.12	1.00
psp2	0.34 - 0.57	0.33 - 0.57	0.34 - 0.57	0.44
psp3	0.67 - 0.80	0.68 - 0.82	0.67 - 0.80	0.74
ssa1	0.26 - 0.38	0.26 - 0.38	0.26 - 0.38	0.33
ssa2	0.23 - 0.40	0.23 - 0.40	0.23 - 0.40	0.31
ssa3	0.22 - 0.35	0.22 - 0.35	0.22 - 0.36	0.30
ssa4	0.18 - 0.41	0.18 - 0.41	0.18 - 0.41	0.30
ssa5	0.27 - 0.44	0.26 - 0.44	0.27 - 0.44	0.37
ssa6	0.54 - 0.76	0.53 - 0.76	0.54 - 0.76	0.66
ssa7	0.68 - 0.77	0.68 - 0.76	0.68 - 0.77	0.73

Note. In the residual model, residual error was constrained to equality across days. For all other models, the range of values across the five days is provided.