Abstract

Latent variable models (LVMs) are incredibly flexible tools that allow users to address research questions they might otherwise never be able to answer (McDonald, 2013). However, one major limitation of LVMs is evaluating model fit. There is no universal consensus about how to evaluate model fit, either globally or locally. Part of the reason evaluating these models is difficult is because fit is typically reduced to a handful of statistics that may or may not reflect the model's adequacy and/or assumptions. In this paper we argue that proper evaluation of model fit *must* include visualizing both the raw data and the model-implied fit. Visuals reveal, at a glance, the fit of the model and whether the model's assumptions have been met. Unfortunately, tools for visualizing LVMs have historically been limited. In this paper, we introduce new plots and reframe existing plots that provide necessary resources for evaluating LVMs. These plots are available in a new open-source R package called `flexplavaan`, which combines the model plotting capabilities of `flexplot` with the latent variable modeling capabilities of `lavaan`.

Seeing the Impossible: Visualizing Latent Variable Models with `flexplavaan`

Introduction

It is currently an unprecedented time in the social sciences; multiple scientific disciplines are reeling from a "replication crisis" (Camerer et al., 2018; Ioannidis, 2005; Pashler & Wagenmakers, 2012), new norms for credibility are becoming more prevalent (Nelson, Simmons, & Simonsohn, 2018; Nosek, Ebersole, DeHaven, & Mellor, 2018), and the push for open science is accelerating at a rapid pace (Nosek, Ebersole, DeHaven, & Mellor, 2018). Amidst this push for open science practices, some have called for greater use of visualization techniques (Fife, 2020b; Fife, Longo, Correll, & Tremoulet, 2021; Fife & Rodgers, 2019; Tay, Parrigon, Huang, & LeBreton, 2016). As noted by Tay, et al. (2016), "[visualizations]. . . can strengthen the quality of research by further increasing the transparency of data. . . " (p. 694). In other words, one of the best, and most efficient ways of making data analysis open and transparent is to display each and every data point through visualization techniques. This is particularly important in research applications where participant-level data cannot be shared.

Not only do visualizations adhere to the principles of openness and transparency, but they offer several additional advantages; they vastly improve encoding of information (Correll, 2015), they highlight model misfit (Healy & Moody, 2014), and they are an essential component in evaluating model assumptions (Levine, 2018; Tay et al., 2016). As such, we (as well as others, e.g., Fife, 2019, 2020b; Fife et al., 2021; Wilkinson & Task Force on Statistical Inference, 1999) recommend every statistical model ought to be accompanied by a graphic.

Unfortunately, this suggestion is easier said than done. While visualizing some statistical models is trivial (e.g., regressions, t-tests, ANOVAs, multiple regression), visualizing others is not. One particularly troublesome class of models to visualize is latent variable models (LVMs). While researchers routinely visualize LVMs with conceptual models (e.g., via path diagrams), visualizing the fits of statistical models is not so easy. The former

visualizations are common, while the latter are not (Hallgren, McCabe, King, & Atkins, 2019). The reason statistical visualizations of LVM are not intuitive is because they rely on unobserved variables (Bollen, 1989). If the variables of interest are unobserved, how can we possibly visualize them?

Though it is not, at first glance, intuitive how to visualize unobserved variables, that does not mean visualizing them is any less important. On the contrary, visualizing latent variables is more important because their presence is unobserved. In the following section, we elaborate on how LVMs are traditionally evaluated and why visualizations are particularly crucial. We then review previous approaches others have used for visualizing LVMs, and note their strengths and weaknesses. We then introduce our approach to visualizing LVMs. Our strategy approaches visualization differently from existing approaches and is able to detect misfit in ways undetected by traditional fit indices and visualization strategies. Furthermore, our approach allows users to detect misfit in just/under-identified models, as well as differences in fit for equivalent models. This approach and the corresponding R package `flexplavaan` allows users to visualize both `lavaan` (Rosseel, 2012) and `blavaan` (Merkle & Rosseel, 2018) objects in R. We conclude with examples that highlight how visualizations assist in identifying problematic statistical models.

Evaluating Model Fit in LVMs

The validity of LVM-based inferences assume structural models closely approximate real-world causal processes (Bollen, 2019; Hayduk, 2014). Unfortunately, evaluating model adequacy in LVMs is rife with obstacles. For one, misspecifications in any one part of the model can lead to biases that spread throughout the full systems of equations (Bollen, 2019). To address this issue, LVM practitioners generally rely on global fit tests and approximate fit indices to evaluate the model adequacy (Jackson, Gillaspy, & Purc-Stephenson, 2009).

Yet global fit indices themselves represent an obstacle to intelligent model evaluation.

Models can yield desirable values (e.g., a non-significant χ^2 test), even when specific aspects of the model are substantially misspecified (Goodboy & Kline, 2017; Hayduk, 2014; Tomarken & Waller, 2003). In other words, a nonsensical model can still yield estimates that lead one to believe in their own statistical models. Conversely, fit indices might penalize even well-specified models for fit-irrelevant characteristics of the model (Browne, MacCallum, Kim, Andersen, & Glaser, 2002). Moreover, in our experience, applied users often lack an intuitive understanding of what global fit statistics tell them about their models. Does a Tucker-Lewis Index (TLI) value of 0.95 mean we have established a strong theoretical foundation for a model? If the root mean square error of approximation (RMSEA) value dips below 0.05, should we assumed that the model estimates closely approximate real-world relationships? This lack of understanding is evident when applied users express confusion about why global fit indices sometimes provide seemingly conflicting assessments of model fit (Lai & Green, 2016). While global fit measures may be useful in comparing models and/or identifying problematic model features, users instead rely on conventional cutoffs (e.g, Hu & Bentler, 1998) to determine whether a model is "adequate," a practice that has received resounding criticism (Barrett, 2007; Chen, Curran, Bollen, Kirby, & Paxton, 2008; Hayduk, 2014; McIntosh, 2007).

A number of scholars have emphasized the importance of supplementing LVM global fit indices with local fit assessment, investigating the tenability of all specific model implications individually (Bollen, 2019; Goodboy & Kline, 2017; Hayduk, 2014; Thoemmes, Rosseel, & Textor, 2018; Tomarken & Waller, 2003, 2005). Local fit evaluation procedures such as inspection of residual correlation matrices (e.g., Bollen, 1989), confirmatory tetrad analysis (Hipp & Bollen, 2003), and equation-based overidentification tests (Bollen, 2019) can help identify individual model specifications that are inconsistent with the data and may give clues about effective remedial strategies (Bollen, 1989; Goodboy & Kline, 2017; Thoemmes et al., 2018; Tomarken & Waller, 2003, 2005).

While local fit indices are an important part of model evaluation (Kline, 2004), they too have limitations. First, although local fit evaluation can be helpful in identifying which implied bivariate associations are most discrepant with the data, they do not necessarily pinpoint the source of the problems (Hayduk, 2018). For instance, an error in the specification of the measurement of an exogenous latent variable could lead to a large residual correlation between measured indicators for the exogenous and endogenous latent variables even if the structural portion of the model is properly specified.[1] Second, it is unclear when a non-zero residual correlation should be interpreted as meaningful. Imposing a cutoff value is one approach (Goodboy & Kline, 2017), though the magnitude of discrepancies is an imperfect metric for determining its practical import (Hayduk, 2014). Standardized residual correlations allow for significance tests for each element of the residual correlation matrix, though residual correlation matrices often include many elements, requiring an adjustment to the p-value to avoid an inflated Type I error rate (Bollen & Arminger, 1991). Additionally, residual correlations may not alert modelers to strong nonlinear associations among observed variables (Flora, LaBrish, & Chalmers, 2012). Finally, local fit indices often produce an overwhelming amount of statistical information that must be filtered. This can be a daunting process, particularly for those not accustomed to LVMs.

Another problem shared by both global and local fit indices is that they are a highly *compressed* representations of both the data and the model. Many readers may be familiar with Anscombe's quartet (reproduced in Figure 1). Anscombe demonstrated there are *many* different raw data patterns that can yield identical correlations/slopes/intercepts. For this reason, simply reporting statistics without visualizing them can be misleading. Likewise, with LVMs, there is a many-to-one relationship between data and indices; very different types of data may yield identical statistics. Some of these data patterns may be very poorly

[1] Notably, the model-implied instrumental variable (MIIV) approach can contain problems caused by misfit in one aspect of a model and prevent ill effects from spreading (Bollen, 2019).

represented by the model. This problem is only exacerbated with LVMs simply because traditional algorithms compress the data at multiple levels (e.g., raw data are compressed into means/covariances, which are then compressed into model parameter estimates and/or indices that evaluate model fit). Put differently (and perhaps quite cynically), LVM is the process of compressing hundreds or thousands of data points into a handful (or less) of estimates that may or may not represent the data-generating process. When considered from this perspective, the most common practices of evaluating model fit seem primitive, at best.

The best defense against this compression is to supplement statistics with evaluation strategies that evaluate *un*compressed data. Perhaps the best way to evaluate uncompressed data is through visualizations, particularly visuals that display raw data (Fife, 2020b).

Previous Approaches to Visualizing LVMs

There is sparse literature describing visual strategies for evaluating the adequacy of LVMs, at least when compared to the extensive literature discussing the merits of global fit tests and indices (e.g., Barrett, 2007; Chen et al., 2008; Hayduk, 2014; Hu & Bentler, 1998; McIntosh, 2007; Shi & Maydeu-Olivares, 2020; Smith & McMillan, 2001; Steiger, 2007). Additionally, while reporting of numeric fit indices is ubiquitous in LVM applications (Jackson et al., 2009), plots of data underlying LVMs are rare (Hallgren et al., 2019). There are, however, some strategies for using visuals to evaluate the tenability of model assumptions, diagnose causal misspecifications, and select the best model from a group of competitors.

A common visual approach for identifying model-data discrepancies is to plot the distribution of the residuals of the covariances/correlation matrix (e.g., using stem-and-leaf plots or histograms; Bollen, 1989). This can aid in identifying specific components of the data that the model struggles to capture (Bollen, 1989; Bollen & Arminger, 1991), which might not be detected with global fit indices (Goodboy & Kline, 2017; Tomarken & Waller,

2003, 2005). Unfortunately, these plots suffer from two major problems. First, the residuals in this case are themselves compressed estimates. As such, we might have a model that is poorly represented by linear correlations (e.g., if the data contain nonlinear relationships), but that problem would never be uncovered by studying residual plots.

A second problem with plots of correlation/covariance residuals is they are extremely limited in the amounts of misspecification they might reveal. For example, suppose we posit a latent variable with three indicators: X_1, X_2, and X_3 (see left image in Figure 2). However, the data-generating model actually has X_3 associated with X_2/X_1 (right diagram in Figure 2).[2] These two models will have the same implied variance/covariance matrix.[3] In other words, a stem and leaf plot will not signal any problems, despite problems existing.

Rather than visualizing aggregates in LVMs, the raw data themselves ought to be visualized. Bollen and Arminger (1991) developed methods for calculating raw and standardized individual case residuals (ICRs), which represent the difference between observed and model-estimated case values for outcome variables. These ICRs are then plotted to help locate outlying and influential observations. Pek and MacCallum (2011)

---

[2] This problem is similar, but not identical to equivalent models (Lee & Hershberger, 1990; MacCallum, Wegener, Uchino, & Fabrigar, 1993). When two models are equivalent, they have identical variance/covariance matrices *and* identical degrees of freedom (DF). In this case, the two models do not have identical DF. However, the visuals we present do offer an intuitive approach for evaluating equivalent models. Two equivalent models with different latent models will have different visuals. Additionally, these visuals will allow one to detect problems with equivalent models, which cannot be detected with traditional model-evaluation statistics.

[3] The model on the left is the user-specified model. The model on the right is the data-generating model. In the left model, the standardized relationship between X_1/X_3 is $a_1 \times c_1$, while in the right model it is r_2. Likewise, the X_1/X_2 relationship is $a_1 \times b_1$ in the left model, while it is r_1 in the right model. The LVM machinery will attempt to make $a_1 \times c_1 = r(X_1, X_3)$, which is the same as setting $a_1 \times c_1 = r_2$ (and it will set $a_1 \times b_1$ to r_1). In other words, the implied variance/covariance matrix will not reveal any problems, despite having a misspecified model.

demonstrated how diagnostic procedures commonly used in generalized linear models (e.g., Mahalanobis distance, generalized Cook's D, and DFBETAs) can be applied to LVMs to detect influential cases with index plots. Flora et al. (2012) applied these diagnostic procedures and others specifically to factor analysis models, and Yuan and Hayashi (2010) used visualizations of Mahalanobis distance metrics to identify high-leverage cases and outliers. Open-source R packages, including `faoutlier` (Chalmers & Flora, 2015) and `influence.SEM` (Pastore & Altoe', 2018), have used these visualization procedures to show case influence on model fit (e.g., likelihood differences) and parameter estimations (e.g., generalized Cook's D). Lee and MacCallum (2015) also demonstrated how visualization can be used to identify specific parameter estimates with strong influence on global model fit.

Aside from outlier detection, other visualization approaches have been proposed to detect structural problems. Asparouhov and Muthén (2014), for example, proposed a method for extending the diagnosticity of ICRs to detect specific structural misspecifications. They demonstrated that plots of estimated factor scores against observed predictor variables can be used to detect unspecified nonlinear effects of the predictor on the latent outcome. Furthermore, they used ICR scatterplots to detect violations of local independence in a congeneric latent factor model. Finally, they demonstrated in a latent factor model how plotting predicted values for a reflective indicator against the observed indicator values could aid in uncovering unmodeled heterogeneity that could be better captured using a mixture model.

Others have introduced visual techniques for growth curve models. Raykov and Penev (2014), for example, compared linear and quadratic growth curve models for the same data. They showed that a scatterplot of the ICRs for the quadratic model vs. ICRs for a linear model can help identify which model best minimizes model-data discrepancies. In the context of growth mixture modeling, Wang, Hendricks Brown, and Bandeen-Roche (2005) showed how visualization of empirical Bayes residuals (e.g., Q-Q and trajectory plots) can

aid in determining the appropriate number of classes, an adequate shape of within-class growth trajectories, and missing confounders.

In short, ICRs have been used to identify high influence/leverage data points, nonlinear effects, heterogeneity, and to compare models. While these are certainly a step in the right direction, existing approaches suffer from a few weaknesses. First, ICRs rely on factor score estimates. Individual latent factor scores cannot be uniquely determined (Grice, 2001; Rigdon, Becker, & Sarstedt, 2019; Steiger, 1996). In cases where factors are highly indeterminate – e.g., factors with few indicators only weakly predicted by the latent factor – different factor score estimation methods can yield highly discrepant values, potentially even estimates that are negatively correlated (Grice, 2001). Also, ICRs are computed under the assumption the model is correct. When the model is misspecified, it is unclear how these visual diagnostics will behave. It is possible, of course, that for misspecified models, ICRs will reveal that misfit. (In fact, we show later that, at least under some circumstances, this is indeed the case). A final limitation of some existing approaches is that many of their visuals were achieved using coding without a dedicated package (e.g., Raykov & Penev, 2014).

In this paper, we introduce `flexplavaan`, which is an amalgamation of the model-plotting capabilities of `flexplot` (Fife, 2020a) and the latent variable modeling capabilities of `lavaan` (Rosseel, 2012). Before we introduce `flexplavaan` and its core functions, however, we explain the types of graphics used and the rationale behind them.

Our Approach (Linear LVMs)

Our approach will introduce five different plots: hopper plots for visualizing residual correlations, trail plots and disturbance-dependence plots for visualizing model-implied fit, and measurement plots/structural plots for visualizing relationships with/between latent variables. These plots are adept at identifying various types of misfit and model assumption violations, such as nonlinearity, correlated errors, missing paths, outliers, nonnormality, etc.

They are also capable of evaluating models that existing numeric approaches cannot evaluate (e.g., equivalent models and just/under-identified models). Before we introduce these approaches, we introduce our simulated example dataset.

Example Data

To motivate our discussion/explanation of visualizing LVMs, we begin with a simulated dataset. Suppose the Jedi Council is attempting to identify Padawans who will make good Jedi Knights. To do so, they develop seven indicators (light saber score, fitness score, midichlorian levels, and a Jedi history exam, as well as three written exams completed at the end of Jedi training). Unbeknownst to the Jedi, these indicators measure two latent factors (force and Jedi), according to the relationships specified in Figure 3. Notice that one variable (history) has cross loadings on both factors. Also notice the path from force to Jedi is represented by a curved *one*-headed arrow. This is to indicate there is a curvilinear relationship between the two variables. (Specifically, the relationship between Jedi and force was simulated to be a quadratic).

Unfortunately, the Jedi Council posits the model shown in Figure 4. (Jedi are notoriously poor at psychometrics.) While most of the important elements are there, the Jedi's (incorrect) model specifies that history loads only onto force, and they posit a linear (rather than nonlinear) path from force to Jedi. According to conventional measures of fit, the χ^2 for the misspecified model is not statistically significant. Also, the CFI (1.00), TLI (0.99), RMSEA (0.03), and SRMR (0.03) indicate respectable fit. By most standards, this is a well-fitting model. However, as we show, these statistics are misleading. In the examples that follow, we will use this example both to demonstrate how the visualization algorithms function and how they are able to identify misspecification that traditional fit statistics fail to capture.

Hopper Plots

Traditionally, conscientious researchers wanting to engage in model evaluation will often produce stem-and-leaf diagrams of the residual variance/covariance matrix. However, stem-and-leaf diagrams are somewhat less intuitive, especially for those without experience interpreting them. As an alternative, we suggest using what we call "hopper plots," which plot the size of the residuals against the rank-ordered correlations (see Figure 5). In other words, as we go from top to bottom, the residual correlations will get smaller and smaller. The dots represent the residual size, while the lines show the absolute value of the residual (on the right) and $-1\times$ the absolute value of the residual (on the left). These plots end up looking like a funnel, but the name "funnel plot" was already snagged by meta-analytic researchers (Egger, Smith, Schneider, & Minder, 1997). Instead, we call them hopper plots. (A hopper is a type of funnel, frequently used to dispense grains.) By default, hopper plots only show those variables with residuals larger than 0.01 (in absolute value). Ideally, the width of the top of the hopper will have a small residual value. In the plot shown in Figure 5, we see that there are problems with fitting correlations with the variables saber, fitness, and midichlorian.

Trail Plots

While hopper plots easily convey misfit at the correlation (or covariance level), they are still quite limited in that they only show compressed data. Ideally, we would have plots that reveal misfit at the raw data level. In order to conceptualize our approach to visualizing LVMs, let us first consider how typical linear models are visualized. In a standard regression, each dot in a scatterplot represents scores on the observed variables. Often, analysts overlay additional symbols to represent the fit of the model (e.g., a line to represent the fitted regression model, or large dots to represent the mean). Sometimes additional symbols are overlaid to represent uncertainty (e.g., confidence bands for a regression line or standard error bars). In either case, the dots represent observed information, while the fitted

information is conveyed using other symbols.

Likewise, visualizing LVMs might follow similar conventions; the dots should represent the observed information, as in Bauer (2005). In his visuals, pairwise relationships between observed variables are represented in a scatterplot. However, Bauer's approach did not overlay a model-implied fit, as we seek to do. When the line represents the model-implied fit, it denotes the "trail" left behind by the unobserved latent variable. As such, we call these plots "trail plots." Importantly, which variable falls on the X versus Y axis is arbitrary.

How then does one identify the slope/intercept of the LVM's model-implied fit? It is quite easy to do so when standard linear LVMs are used. Recall how our force factor has four indicators (e.g., saber, fitness, midichlorians, and history). To visualize the bivariate relationship between saber and fitness, for example, we can simply utilize the model-implied variance/covariance matrix. Recall the relationship between a covariance and a slope:

$$\beta_{y|x} = \frac{\sigma(x, y)}{\sigma_x^2}$$

For our example,

$$b_{\mathrm{S|F}} = \frac{\hat{\sigma}_{\mathrm{S,F}}}{\hat{\sigma}_F^2}$$

where S and F represent "saber" and "fitness," respectively, and $\hat{\sigma}_F^2$ represents the *residual* variance of fitness. (Put differently, this is the expected slope between saber and fitness, unadjusted for unreliability). With the slope, one can then estimate the intercept using basic algebra:

$$b_0 = \overline{S} - \beta_{\mathrm{S|F}} \times \overline{\mathrm{F}}$$

Figure 6 shows the LVM model-implied fit in red with a regression line in blue. Because the regression line minimizes the sum of squared errors, we would hope the LVM fitted line (red) closely approximates the regression line (blue). In this case, the two are very similar, though it may suggest the model (red line) possibly underestimates the relationship between the two variables.

Of course, Figure 6 only shows one pairwise relationship between variables. If we wished to visualize all the variables in our model, we would have to utilize a scatterplot matrix, as in Figure 7. The diagonal elements show histograms of ICRs, enabling researchers to (somewhat) evaluate the assumption of normality.[4] Naturally, this becomes quite cumbersome when users have more than seven or eight variables. In this case, it is best to visualize only a subset of variables. By default, `flexplavaan` sorts the scatterplot matrix using a blockmodeling algorithm, then presents the block with the largest residuals first. For our figure (Figure 7), we have examined only the variables associated with the latent force variable, while Figure 8 shows the variables associated with the latent Jedi variable. We have also asked `flexplavaan` to show a loess line instead of a regression line, which will allow us to detect nonlinear patterns. These figures reveals some potential problems, including some nonlinearity (e.g., between the exam variables), some underestimation (e.g., between saber and midichlorian), and some overestimation (e.g., between saber and history).

The primary advantage of trail plots is that they easily show many types of misspecification in LVMs. Another advantage is they visually (and often times strikingly) show how little information a model might capture. Returning to Figure 7, we see that many of the bivariate plots reveal quite weak relationships; many of the slopes are quite near zero. Recall that global fit indices suggested a well-fitting model. The trail plots, however, suggest

[4] Technically, maximum likelihood-based LVMs assume multivariate normality, while these plots show univariate normality. However, the univariate plots at least suggest when multivariate normality might be violated.

there's little information to fit from the beginning for at least some of these relationships.

Disturbance-Dependence Plots

One common technique for visualizing the adequacy of statistical models in classic regression is residual-dependence plots. With these graphics, one simply plots the residuals of the model (Y axis) against the predicted values (X axis). The rationale behind this is simple: the model should have extracted any association between the prediction and the outcome. The residuals represent the remaining information after extracting the signal from the model. If there is a clear trend remaining in the data (e.g., a nonlinear pattern or a "megaphone" shape in the residuals), this indicates the model failed to capture important information.

Likewise, in LVMs, we can apply this same idea to determine whether the fit implied by the LVM has successfully extracted any association between any pair of predictors. However, in LVMs, residuals refer to the discrepancy between the model-implied and the actual variance/covariance matrix (or correlation matrix). As such, naming these plots "residual-dependence plots" would be a misnomer. Rather, misfit at the raw data level is typically called either a disturbance or an individual case residual (ICR), as mentioned previously. In this paper, we call these plots disturbance-dependence plots.

Like trail plots, we visualize disturbance-dependence plots for each pair of observed variables. To do so, `flexplavaan` subtracts the fit implied by the model from each individuals' observed scores. For example, a disturbance dependence plot for an X_2/X_1 relationship (assuming X_2 is the outcome) would subtract the fitted (model-implied) X_2 values from the actual X_2 values, then plot those against the raw X_1 values. If the trail-plot fit actually extracts all association between the pair of observed variables, we would expect to see a scatterplot that shows no remaining association between the two. If there is a pattern in the scatterplot remaining, we know the fit of the model misses information about that specific relationship.

To aid in interpreting these plots, we can overlay the plot with a flat line (with a slope of zero), as well as a regression (or loess) line. The first line indicates what signal should remain after fitting the model, while the second line shows what actually remains.

Figure 9 shows an example of trail plots in the upper triangle and disturbance-dependence plots in the lower triangle of a scatterplot matrix. These plots are for the same data shown in the right image of Figure 8. Notice how many of the plots have nonlinear patterns showing up in both the DDP and the trail plots.

Together, these plots (hopper, trail plots and DDPs) serve as a critical diagnostic check. All these plots will signal certain types of misfit both in the measurement and structural components of the model. However, these plots suffer from a major weakness. Recall how earlier we referenced Figure 2 and noted that sometimes severe misspecification will go undetected simply because an incorrect model will often yield a model-implied covariance matrix that well approximates the actual covariance matrix. However, this sort of misspecification may show up in measurement plots, which we address next.

Measurement Plots

Earlier we mentioned how Asparouhov and Muthén (2014) utilized factor score estimates to visualize ICRs as a diagnostic tool. One of their plots showed the latent variable on the Y axis and the observed (indicator) variable on the X axis. While these may be capable of revealing nonlinearities, they cannot reveal other sorts of misspecification (e.g., many types of cross-loadings and residual correlations) without a simple modification. The modification we propose is similar to the trailplots: overlay the model-implied slope and a regression (or loess) line.

Fortunately, similar to the trail plots, we can use the model-implied variance/covariance matrix of the observed/latent variables to determine the model-implied slope. The advantage of these plots is they are far more sensitive to many types of

misspecification than trail plots. This is because trail plots are unable to pick up misspecification unless that misspecification introduces bias in estimating the observed variance/correlation matrix. However, as shown in Figure 2, not all misspecification manifests itself as bias in estimating correlations between observed variables. While the two models in Figure 2 will not yield different observed covariances, they will yield different latent variable estimates (and thus, different covariances between latent/observed).

To create these plots, `flexplavaan` does the following:

1. Obtain empirical Bayes estimates of the factor scores (using the `lavPredict` function in the `semTools` package).
2. Convert each of the observed variables to z-scores (using the means/SDs of the empirical Bayes factor scores). This allows one to plot multiple observed variables in the same plot.
3. Use the estimated latent/observed correlation (i.e., standardized) matrix to compute the slope between each observed/latent variable.[5]

Figure 10 plots several graphs of the relationship between the observed variables and the latent variables. `flexplavaan` defaults to displaying only four observed variables at a time. Which four are chosen is determined by the degree of discrepancy between the observed (blue) and model-implied (red) slope, such that the four observed variables with the largest discrepancy are chosen. From Figure 10, we see a consistent pattern of nonlinearity between the observed and latent variables that is not captured by the model (shown in red). It is also interesting to note that the history indicator is nearly synonymous with the force latent variable.

---

[5] Because we use the standardized (correlation) matrix between latent/observed, these estimates are unaffected by scaling decisions (e.g., setting a particular indicator's weight to one).

Structural (Cross-Hair) Plots

When modeling latent variables, often the visuals of interest are not the observed variables, but the latent variables. In other words, the measurement model is ancillary to the substantive model. Naturally, we might wish to visualize the relationship between the latent variables.

However, latent factor scores computed by LVM software are merely predictions. As such, we ought to have visuals that reflect uncertainty in our predictions for the latent scores. In `flexplavaan`, this uncertainty is represented as crosshairs. The widths of each line of the crosshair (for both the X axis and the Y axis) are obtained from prediction intervals for `lavaan` objects (using the `plausibleValues` function of the `semTools` package; Jorgensen, Pornprasertmanit, Schoemann, and Rosseel, (2020)). Importantly, the value of factor score estimates is dependent on the degree of factor indeterminacy (Grice, 2001; Rigdon et al., 2019; Steiger, 1996). Figure 11 shows these plots, which we call "structural plots," or "cross-hair plots" with a loess line (in blue). We also added a regression line (in red). Interestingly, the relationship between the two, though simulated to be significantly nonlinear, shows up as fairly linear. (The blue loess line and red regression line are quite similar). This seems to suggest that, by the time a model's factor scores are estimated, any nonlinearity that exists in the data has been discarded. For this reason, we disagree with the approach recommended by Hallgren et al. (2019), who suggest researchers plot a structural plot as evidence for model fit; simply plotting structural plots without the previous graphics will yield a very misleading picture of data/model fit. This final plot is little more than a summary of the model, provided the model diagnostics check out. As such, showing these is like plotting only a fitted line for a regression model without the underlying raw data.

When there are multiple latent variables, there is a great deal of flexibility in how one visualizes the structural model. `flexplavaan` makes a best guess at how to visualize this relationship using the model specified by the user. However, the user can always specify how

to plot the structural model using a `flexplot` equation (Fife, 2020a). In our example, we only had two variables to visualize, so a simple bivariate plot was most natural. When more variables are included, we might utilize paneling, added variable plots, beeswarm plots, etc. For a review of the types of plots possible, see Fife (2020a).

Model Comparisons

It is quite common to compare two different models when using LVMs. Indeed, model comparisons are a recommended strategy for building LVMs (Rodgers, 2010). Not only can we use statistics to do model comparisons, but we can also compare them visually. All plots previously mentioned (hopper plots, trail plots, DDPs, measurement, and structural plots) can visualize two models at the same time. While visualizing two models side-by-side makes differences more visually detectable, we also recommend visualizing each model individually since certain aspects of misfit in a model might be masked by the comparison.

For illustration purposes, we're going to fit a second model that attempts to address both the nonlinearity and the missing cross-loading. The second model is shown in Figure 12. To model the nonlinearity, we have created three nonlinear indicators (saber^2, fitness^2, and midichlorian^2), using the "product indicator" approach (Little, Bovaird, & Widaman, 2006). We have also added the cross-loading from Jedi to history.

Figure 13 shows the residual plots, for all pairwise relationships. The new model seems to have reduced the largest residuals (though it has, in some instances, slightly increased the size of some residuals).

Figure 14 shows the trail/DDPs for the two models. As before, the red line is the original model, while the blue line shows the nonlinear model for the variables saber, fitness, midichlorian, and history. There are some minor discrepancies (e.g., the saber/midichlorian relationship), with the nonlinear model suggesting stronger relationships between the variables. However, as before, these plots are not terribly sensitive to misspecification. (In

fact, the trail/DDPs for the other factor showed hardly any discrepancies.)

On the other hand, the measurement plots (Figure 15) do show differences between the two models, at least for the force latent variable. The nonlinear model shows consistently weaker relationships between the force variable and the indicators shown (force history, exam one, exam two, and exam three).

Finally, the structural plot, shown in Figure 16 shows the relationship between the force/Jedi variable. The data from the left plot are for the original model, while the right plot are for the nonlinear model. Not surprisingly, the right model captures the nonlinearity between the two much better than does the original model.

Recall that the original estimates of fit (e.g., TLI, RMSEA, RMSR) indicated the original model fit quite well. However, the visuals have revealed these estimates were misleading. Rather the original model failed to capture an important nonlinear relationship, overestimated some relationships, and underestimated others. The nonlinear model, however, has smaller residuals, seems to better approximate the actual data, and suggests the latent variables are associated nonlinearly. Table 1 shows fit statistics for each model. We display these *not* to validate the efficacy of the visuals (we trust the visuals more than we trust the fit statistics). Rather, we show these to highlight the convergent validity of the statistics and the graphics.

Comparing Under/Just-Identified/Equivalent Models. As we mentioned previously, simply viewing model-implied variance/covariance matrices may be misleading since two very different models will yield identical variance/covariance matrices (as in Figure 2). Notably, if one were to fit these models, one would be just-identified, while the other is under-identified. Additionally, the two models produce equivalent variance/covariance matrices. (Granted, all just-identified models are also equivalent models). However, when we use `flexplavaan` to visualize these models, we see very clear problems. The top plot in Figure 17 shows the measurement plot for the proposed model in the left image of Figure 2.

Notice how the implied fit (red line) is consistently lower than the regression-based fit (blue line), even though the model is just-identified. This should signal problems that something is wrong with the model. If we were to then fit the right model in Figure 2 and compare the two, we would see very different model-implied relationships between our variables and the latent factor (particularly for $x3$).

The reason `flexplavaan` is able to evaluate models that standard LVM machinery cannot is because the degrees of freedom for fit statistics are based on the variance/covariance matrix. The visual degrees of freedom from `flexplavaan`, on the other hand, are based on the *raw data*. While the LVM degrees of freedom only allow six total sources of misfit (a.k.a., degrees of freedom), the raw data allow a total of 1,000 sources of misfit (because $N = 1000$). To be clear, we don't necessarily recommend routinely fitting just-identified (or under-identified) models; plots are no substitute for statistics (and vice versa). However, in situations where a just/under-identified model is the appropriate model, `flexplavaan` allows one to evaluate these sorts of models.

While we do not necessarily recommend routinely fitting just/under-identified models, we do recommend users visualize equivalent models using `flexplavaan`. While the fit statistics cannot inform users about which model is more appropriate, `flexplavaan` can, as shown in Figure 17.

Applied Example: Social Distress & Wellbeing

To demonstrate the utility of `flexplavaan` with real-world data, we use publicly-accessible data from the NIH Toolbox Norming Study (Gershon, 2016). Our primary aim is to estimate the effect of social distress on emotional well-being. We use scale scores to represent overall levels of constructs measured using multi-item questionnaires from the *Emotion Domain* (Salsman et al., 2013). For the purposes of this demonstration, we limit our sample to adults, defined as age 18 or older (N = 1,629). Our hypothesized model

is depicted in Figure 18:

The model freely estimates 15 parameters labeled with red Greek letters. The primary response variable, η_2, measures emotional well-being as a latent factor with two reflective indicators: *life satisfaction* (y_3) and *meaning* (y_4). The primary exposure variable is social distress (η_1), measured with two reflective indicators: *loneliness* (y_1) and *sadness* (y_2). Social distress also has two causal indicators, *perceived hostility* and *rejection*. Rather than assuming these two causal indicators are perfectly measured, we use reliability estimates reported in Cyranowski et al. (2013) to form single-indicator latent variables (ξ_1 and ξ_2) with error variances equal to $(1 - \alpha_i) \times \sigma(\xi_i)$, where α_i is the estimate of reliability for measured variable i (Hayduk & Littvay, 2012). We allow the *sadness* reflective indicator to have a direct causal effect on well-being, indicating the hypothesis that the association between sadness and well-being is not merely due to the having social distress as a common cause.

It is sometimes difficult to distinguish between causal and effect indicators for latent factors (Bollen & Bauldry, 2011), and modelers may differ in how they conceptualize the same indicators. Figure 19 presents an alternative model in which *perceived hostility* and *perceived rejection* are treated as reflective indicators of the *social distress* factor, rather than causal indicators. In our hypothesized model, changes in perceived hostility and rejection lead to corresponding changes in levels of the social distress latent factor and its effect indicators; however, changes in social distress do not lead to changes in perceived hostility and rejection. In the alternative model, changes in the social distress factor lead to changes in perceived rejection and hostility. This difference between models has important clinical implications. In the hypothesized model, interventions reducing perceived rejection and hostility would lead to improvements in both social distress and well-being. The same is not true in the alternative model where rejection and hostility exert no causal influence on social distress or well-being. The alternative model estimates one fewer parameter than the hypothesized model (14 vs. 15) and, consequently, has one additional df.

We use `flexplavaan` to evaluate the tenability of these two models. All measured variables are treated as continuous. Preliminary graphical analyses showed evidence of floor and ceiling effects causing considerable skewness and kurtosis for several variables, so we used the MLR estimator in `lavaan` which provides robust standard errors and a scaled χ^2 test statistic.

We evaluate the global fit of the competing models using the scaled χ^2 fit test and three approximate fit indices: the RMSEA with a 90% confidence interval, the TLI, and the square root mean residual (SRMR). (See Table 2). The scaled χ^2 tests for both models indicate that there are model-data discrepancies that exceed the conventional threshold (p=.05) used to distinguish chance vs. non-chance expectation, though the χ^2 test statistic is considerably smaller in the hypothesized model. The approximate fit indices favor the hypothesized model. However, as we soon show, these fit statistics paint an overly simplified picture of the model.

We begin by generating hopper plots to visualize which bivariate correlations are most poorly approximated by the model. (See Figure 20). The hypothesized model yields smaller residual correlations for nearly all variable pairs, with the absolute value of all residual correlations $< .03$. The alternate model does a relatively poor job reproducing the bivariate associations among the indicators of *social distress*. For example, when including *hostility* and *rejection* as reflective indicators, the alternative model underestimates the *lonely-sad* and *reject-hostility* correlations, whereas the hypothesized model reproduces these relationships nearly perfectly. The hypothesized model also generally does better reproducing correlations between the *social distress* indicators and the *well-being* indicators. The alternative model most notably overestimates the *loneliness-meaning* correlation.

At this point, we expect that many SEM users might end their critique of the hypothesized model. Most reviewers in applied journals would accept that the model provides a "close fit" to the data as all approximate fit indices are within conventional thresholds. With $N > 1000$, the significant χ^2 test might be dismissed as overly sensitive,

likely detecting negligible discrepancies between the model and data. However, we will pursue visual inspection of the model using `flexplavaan` and see what other insights we might glean from visual analysis.

First, we generate trail and disturbance dependence plots for our hypothesized model. To make the plot matrix manageable, we initially include only the indicators for the social distress latent factor. These are shown in Figure 21. The trail plots show floor effects with large clusters of points in the bottom-left corners. Additionally, the loess lines summarizing the observed patterns of association show some evidence of curvature for several variable pairs at high levels of the X-axis variables where the data are sparse. There appears to be clusters of outlying points with high leverage pulling the loess lines downward away from the implied linear slope. For example, there appears to be a number of participants reporting high levels of hostility and rejection but low levels of sadness. This suggests these two indicator variables might actually not be indicators of a common latent variable. Furthermore, several disturbance dependence plots also show clusters of aberrant points contributing to deviations from the model-implied zero slopes. The floor effects also appear to be contributing to heteroscedasticity.

Similar patterns can be seen when visualizing relationships between the social distress causal indicators (*rejection* and *hostility*) and the well-being reflective indicators (*meaning* and *satisfaction*) in Figure 22. There are departures from the model-implied linear effects, seemingly driven by clusters of aberrant cases (e.g., participants reporting high levels of rejection *and* high levels of meaning). Once again, we have evidence our indicators may not share the same common factor. We can also view how the hypothesized and alternative models differ in their implications by including both models in the same trail and disturbance dependence plots, as in Figure 23.

The models produce generally similar trail and disturbance dependence plots. However, a close evaluation of the trail plots reveal how the models differ in their implications. The

implied correlations between the hypothesized causal indicators and reflective indicators of the *social distress* factor are consistently stronger in the alternative model, where all four indicators are treated as reflective indicators. In contrast, the implied correlations between the two hypothesized causal indicators (*hostility* and *rejection*) and between the two hypothesized reflective indicators (*sadness* and *loneliness*) are stronger in the hypothesized model.

Figure 24 plots measurement plots for the hypothesized model for the distress latent variable. These figures generally show that the model-implied relationships between latent/observed variables are consistent with the observed relationships between factor score estimates and the observed variables. However, there is some evidence of curvilinearity. This is particularly notable for the hostility, meaning, and life-satisfaction indicators. This could be for multiple reasons: it may be that the distress to well-being latent variable relationship is nonlinear. Or it may be that relationships from the indicators to the latent variables are curvilinear. Or, perhaps, this may be at least partly an artifact of the floor effect for the hostility variable, resulting in an underestimate of the effect of hostility on the social distress factor. There also appear to be a cluster of outliers – individuals reporting low levels of meaning *and* low levels of distress – pulling the loess line for the plot of distress against meaning away from the implied linear relationship.

As many of the `flexplavaan` visualizations showed outlying points that were inconsistent with the model-implied relationships, we use the `fitinfluence()` function in the `influence.SEM` package (Pastore & Altoe', 2018) to identify cases with the greatest influence on the hypothesized model's χ^2 test. The 10 cases with the greatest influence on model fit are plotted in Figure 25. Notice there are cases whose removal would improve model fit and others whose removal would worsen fit. The removal of the case with the largest influence on the model fit (highlighted in green) would result in a reduction (improvement) in the χ^2 value of 2.34. This participant had an unusual pattern of values on

the measured variables, scoring 1.58 *SD*s above the sample mean on the *meaning* composite despite scoring 1.70 *SD*s below average on *life satisfaction* and well above the sample means on *loneliness* (+2.70 *SD*s) and *sadness* (+1.90 *SD*s). Importantly, these unusual combinations are theoretically plausible. One could, for example, maintain high levels of meaning despite having poor life satisfaction and high levels of sadness and loneliness. A big discrepancy in *meaning-satisfaction* and *loneliness-sadness* composite scores was common among participants for whom the model performed poorly. This is an indication that we may need to further refine our theory and model. It may be, for example, preferable to treat *life satisfaction* and *meaning* as separate latent variables rather than as indicators of a single latent. To save space, we do not model that in this paper. Rather we merely note model problems highlighted with `flexplavaan` to demonstrate how one might use these sorts of visuals for model refinement.

This example demonstrates how `flexplavaan`'s visualization functions can facilitate deeper evaluation of LVMs than numeric fit indices alone. Hopper plots overlaying the residual covariances from our hypothesized and alternative models showed plainly that our hypothesized model consistently better reproduced bivariate associations than the alternative model. Additionally, the Hopper plot showed whether discrepancies were due to over- vs. under-estimation of associations. Trail, disturbance dependence, and measurement plots alerted us to potentially problematic distributional properties (i.e., floor and ceiling effects) and high-leverage outliers. These suggest possible modifications to the model, including nonlinear indicator effects, nonlinear latent effects, and/or disaggregating indicators. This example is instructive as it utilized data from real-world instruments with unwieldy distributions common in psychology applications. When used in conjunction with the `influence.SEM` package (Pastore & Altoe', 2018), we identified cases for which the model performed poorly, giving us hypotheses about potential modifications to our theory and model.

Modelers relying exclusively on conventional cutoffs on the approximate fit indices – all of which had values indicative of a well-fitting model by conventional standards – would be unlikely to appreciate the potential shortcomings of the hypothesized model. Importantly, `fleplavaan` could be helpful in identifying problems in models even when the χ^2 test is not significant, such as when one is modeling just-identified models, or even models with smaller sample sizes.

Discussion

LVM scholars have long warned about the deficiencies of relying exclusively on the χ^2 fit test and approximate fit indices when judging model adequacy (e.g., Tomarken & Waller, 2003). It is well-known that models meeting conventional standards for model adequacy can have serious problems that go undetected, potentially leading to biased estimates and faulty inferences (Hayduk, 2014; Tomarken & Waller, 2003). The χ^2 test, though more sensitive to misspecifications than approximate fit indices, cannot detect all problems (Hayduk, 2014). Therefore, regardless of whether there is evidence of misspecification, in-depth model evaluation strategies are needed to identify problems (including those flying under the fit index radar) and to inform subsequent model development. When models perform poorly, numeric fit indices provide, at best, vague hints about what went wrong. A model could provide a poor match to the data because of misspecifications in the causal structure, violations of distributional assumptions, or a combination of both (Bollen, 1989).

In this paper, we have introduced `fleplavaan`, an open-source R package designed to provide easy-to-use and intuitive visualizations of latent variables. Throughout we have shown various instances where model fit indices failed to capture important misspecification that `flexplavaan` easily captured. We have also showed instances where fit indices signaled problems, but provided very little guidance on how to proceed. `flexplavaan`, on the other hand, revealed these problems strikingly.

In addition to providing intuitive diagnostic tools, `flexplavaan` also overcomes (or at least mitigates) various problems endemic in LVMs. For example, with equivalent models, fit statistics will be identical, while visualizations will not. Likewise, just/under-identified models do not allow researchers to evaluate fit using traditional indices. The visuals provided by `flexplavaan`, on the other hand, do provide a means to evaluate model adequacy under these conditions.

Beyond helping LVM modelers, we believe that `flexplavaan` can improve model transparency and allow reviewers and research consumers to better scrutinize LVMs than they could if they had to rely exclusively on summaries of aggregate statistics. In our applied example with the NIH Toolbox data, for example, the visualizations clearly alert the user to potential model shortcomings even when the approximate fit indices would generally be considered to be indicative of a close fit. Thus, data visualizations can make it more difficult for model problems to go undetected by readers.

Hopefully, throughout this paper, we have convinced the reader that existing methods for evaluating model fit have very poorly communicated model strengths and weaknesses. Without these visuals, we have historically been modeling blind, both literally and figuratively. It is our hope that the tools we provide through `flexplavaan` will increase transparency, improve communication, and provide a means to refine the models we use for psychological research.

References

Asparouhov, T., & Muthén, B. (2014). Using mplus individual residual plots for diagnostic and model evaluation in sem. *Mplus Web Notes*, (20).

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*, *10*(3), 305–316. https://doi.org/10.1037/1082-989X.10.3.305

Bollen, K. A. (1989). *Structural equations with latent variables.* John Wiley & Sons.

Bollen, K. A. (2019). Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling. *Multivariate Behavioral Research*, *54*(1), 31–46.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology*, 235–262.

Bollen, K. A., & Bauldry, S. (2011). Three cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265.

Browne, M. W., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–421. https://doi.org/10.1037//1082-989x.7.4.403

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. https://doi.org/10.1038/s41562-018-0399-z

Chalmers, R. P., & Flora, D. B. (2015). Faoutlier: An r package for detecting influential cases in exploratory and confirmatory factor analysis. *Applied Psychological Measurement*, *39*(7), 573–574. https://doi.org/10.1177/0146621615597894

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research*, *36*(4), 462–494.

https://doi.org/10.1177/0049124108314720

Correll, M. A. (2015). *Visual Statistics* (Doctoral Dissertation). University of Wisconsin-Madison.

Cyranowski, J. M., Zill, N., Bode, R., Butt, Z., Kelly, M. A., Pilkonis, P. A., ... Cella, D. (2013). Assessing social support, companionship, and distress: National institute of health (nih) toolbox adult social relationship scales. *Health Psychology*, *32*(3), 293.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629–634.

Fife, D. A. (2019). A Graphic is Worth a Thousand Test Statistics: Mapping Visuals onto Common Analyses. Retrieved from http://rpubs.com/dustinfife/528244

Fife, D. A. (2020a). Flexplot: Graphical-Based Data Analysis. *PsyArxiv*. https://doi.org/10.31234/osf.io/kh9c3

Fife, D. A. (2020b). The Eight Steps of Data Analysis: A Graphical Framework to Promote Sound Statistical Analysis. *Perspectives on Psychological Science*, *15*(4), 1054–1075. https://doi.org/10.1177/1745691620917333

Fife, D. A., Longo, G., Correll, M., & Tremoulet, P. D. (2021). A graph for every analysis: Mapping visuals onto common analyses using flexplot. *Behavioral Research Methods*. https://doi.org/10.3758/s13428-020-01520-2

Fife, D. A., & Rodgers, J. L. (2019). Exonerating EDA, Expanding CDA: A Pragmatic Solution to the Replication Crisis. *PsyArxiv*. https://doi.org/10.31234/osf.io/5vfq6

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology*, *3*, 55.

Gershon, R. (2016). *NIH Toolbox Norming Study* (Version V4) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/FF4DI7

Goodboy, A. K., & Kline, R. B. (2017). Statistical and Practical Concerns With Published Communication Research Featuring Structural Equation Modeling. *Communication Research Reports*, *34*(1), 68–77. https://doi.org/10.1080/08824096.2016.1214121

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430.

Hallgren, K. A., McCabe, C. J., King, K. M., & Atkins, D. C. (2019). Beyond path diagrams: Enhancing applied structural equation modeling research through data visualization. *Addictive Behaviors*, *94*(March 2018), 74–82. https://doi.org/10.1016/j.addbeh.2018.08.030

Hayduk, L. A. (2014). Seeing Perfectly Fitting Factor Models That Are Causally Misspecified: Understanding That Close-Fitting Models Can Be Worse. *Educational and Psychological Measurement*, *74*(6), 905–926. https://doi.org/10.1177/0013164414527449

Hayduk, L. A. (2018). Review essay on rex b. Kline's principles and practice of structural equation modeling: Encouraging a fifth edition. *Canadian Studies in Population*, *45*(3-4), 154–178.

Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, *12*(1), 1–17.

Healy, K., & Moody, J. (2014). Data Visualization in Sociology. *Annual Review of Sociology*, *40*(1), 105–128. https://doi.org/10.1146/ANNUREV-SOC-071312-145551

Hipp, J. R., & Bollen, K. A. (2003). Model fit in structural equation models with censored,

ordinal, and dichotomous variables: Testing vanishing tetrads. *Sociological Methodology*, *33*(1), 267–305.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). *semTools: Useful tools for structural equation modeling*. Retrieved from https://CRAN.R-project.org/package=semTools

Kline, R. B. (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research* (p. 325). Washington D.C.: American Psychological Association.

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when rmsea and cfi disagree. *Multivariate Behavioral Research*, *51*(2-3), 220–239.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research*, *25*(3), 313–334.

Lee, T., & MacCallum, R. C. (2015). Parameter influence in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 102–114.

Levine, S. S. (2018). Show us your data: Connect the dots, improve science. *Management and Organization Review*, *14*(2), 433–437. https://doi.org/10.1017/mor.2018.19

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling*, *13*(4), 497–519.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*(1), 185–199.

McDonald, R. P. (2013). *Test theory: A unified treatment*. psychology press.

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, *42*(5), 859–867. https://doi.org/10.1016/j.paid.2006.09.020

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1708274114

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science. *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pastore, M., & Altoe', G. (2018). *Influence.SEM: Case influence in structural equation models*. Retrieved from https://CRAN.R-project.org/package=influence.SEM

Pek, J., & MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: Cases and their influence. *Multivariate Behavioral Research*, *46*(2), 202–228.

Raykov, T., & Penev, S. (2014). Latent growth curve model selection: The potential of individual case residuals. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 20–30.

Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research*, *54*(3), 429–443.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *The American Psychologist*, *65*(1), 1–12. https://doi.org/10.1037/a0018326

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Salsman, J. M., Butt, Z., Pilkonis, P. A., Cyranowski, J. M., Zill, N., Hendrie, H. C., ... others. (2013). Emotion assessment using the NIH toolbox. *Neurology*, *80*(11 Supplement 3), S76–S86.

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement*, *80*(3), 421–445.

Smith, T. D., & McMillan, B. F. (2001). A primer of model fit indices in structural equation modeling.

Steiger, J. H. (1996). Dispelling some myths about factor indeterminancy. *Multivariate Behavioral Research*, *31*(4), 539–550.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893–898. https://doi.org/10.1016/j.paid.2006.09.017

Tay, L., Parrigon, S., Huang, Q., & LeBreton, J. M. (2016). Graphical Descriptives: A Way to Improve Data Transparency and Methodological Rigor in Psychology. *Perspectives on Psychological Science*, *11*(5), 692–701. https://doi.org/10.1177/1745691616663875

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods*, *23*(1), 27.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with" well fitting" models. *Journal of Abnormal Psychology*, *112*(4), 578.

Tomarken, A. J., & Waller, N. G. (2005). Structural Equation Modeling: Strengths, Limitations, and Misconceptions. *Annual Review of Clinical Psychology*, *1*(1), 31–65. https://doi.org/10.1146/annurev.clinpsy.1.102803.144239

Wang, C.-P., Hendricks Brown, C., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, *100*(471), 1054–1076.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, *54*(8), 594–601.

Yuan, K.-H., & Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnosis using two scatter plots. *Psychological Methods*, *15*(4), 335.

Appendix

Getting Started with Flexplavaan

`flexplavaan` was designed to be easy-to-use with one-line functions for most visualization strategies. To install `flexplavaan`, users must first download the `devtools` package:

```
install.packages("devtools")
```

Then, using the devtools package, the `flexplavaan` package can be installed from github:

```
devtools::install_github("dustinfife/flexplavaan")
```

`flexplavaan` comes pre-loaded with several datasets, including the jedi dataset from this paper. The dataset is called `jedi_jedi` in the `flexplavaan` package. To load and model the dataset, the user could, for example, use the following `lavaan` code:

```
require(flexplavaan)


model = "
Force =~ fitness + saber + midichlorian + force_history
Jedi =~ exam_one + exam_two + exam_three
Jedi ~ Force
"


# Fit the models ----------------------------------------------------------
force_fit = sem(model, jedi_jedi)
```

Once the user has the fitted `lavaan` object, they may use visuals in `flexplavaan`, as in the following code:

```
# show the hopper plots, only including those variables
# that have a residual larger than 0.01
residual_plots(force_fit, max_val = .01)

# trail/ddp of the first 4 variables
visualize(force_fit, subset=1:4)

# alternatively, specify which variables should be plotted
visualize(force_fit, subset=c("fitness", "midichlorian", "exam_one"))

# show the measurement plot for the Force latent variable
implied_measurement(force_fit, latent="Force")

# show the structural plot, with Force on the Y axis.
visualize(force_fit, plot="latent", formula = Force~Jedi)
```

The source code for `flexplavaan` is available on github at github.com/dustinfife/flexplavaan, as well as the source code that generated this document.

Table 1

Global Fit Indices for the Original and Nonlinear Model for the Jedi Dataset.

	χ^2	*df*	*p*	CFI	TLI	BIC	RMSEA	SRMR
Original	19.4	13	0.110	0.996	0.994	42,616.2	0.026	0.025
Nonlinear	26.3	31	0.705	1.000	1.004	50,388.6	0.000	0.022

Table 2

Global Fit Indices for the Alternative and Hypothesized Model for the NIH Toolbox Dataset.

Model	χ^2	*df*	*p*	RMSEA	RMSEA_l	RMSEA_u	TLI	SRMR
Alternative	148.5	7	0.000	0.120	0.110	0.140	0.910	0.040
Hypothesized	20.2	6	0.000	0.040	0.020	0.060	0.990	0.010

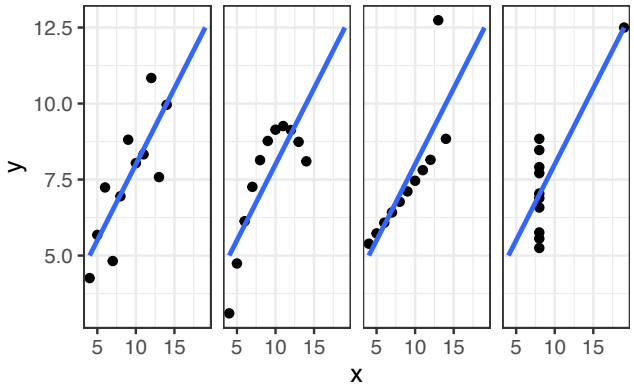

Figure 1. A reproduction of Anscombe's quartet. Each plot has identical regression lines/correlations, even though the underlying data are vastly different.

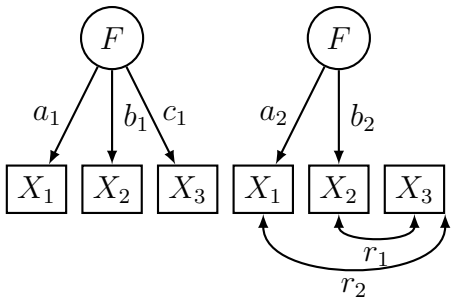

Figure 2. The model on the left is the user-specified model, while the model on the right is the data-generating model. These two models make very different theoretical statements, but have the same implied correlation between the variables. Visualizing the raw data from these models can help identify structural problems for the model on the left.

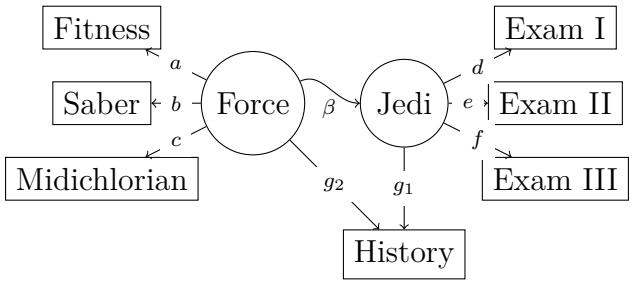

Figure 3. Simulated dataset of Jedi selection and training. This model features a crossloading on the history indicator, as well as a nonlinear relationship between force and Jedi. This is the data-generating model.

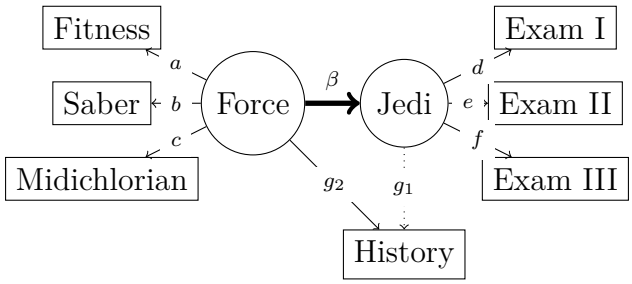

Figure 4. Actual model fitted to the Jedi dataset. This posits a linear (rather than nonlinear) relationship between the latent variables Jedi and force (indicated by the thick black line labeled β) and fails to model the Jedi to history relationship (indicated by the dotted line labeled $g1$).

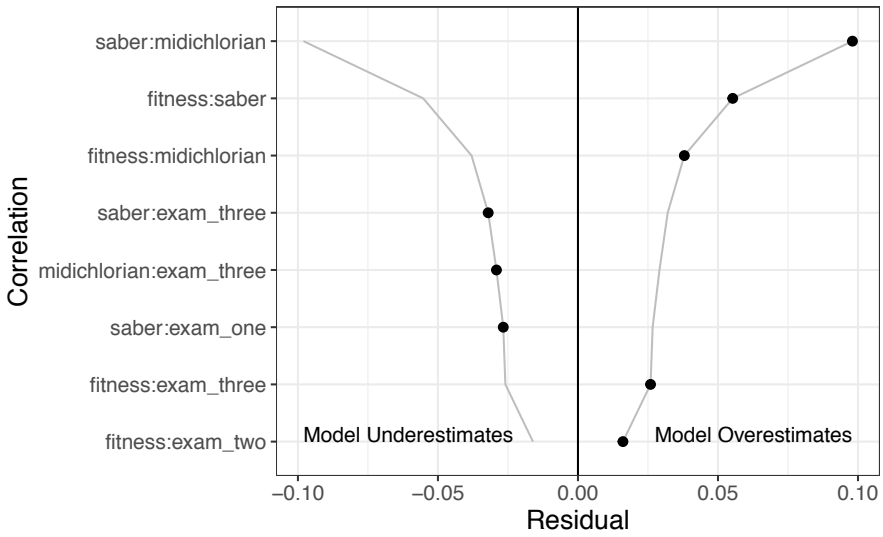

Figure 5. An example of a "hopper plot," which graphs the discrepancy between the implied and observed correlation matrix. The dots represent observed residuals, while the lines represent the absolute value/negative absolute value.

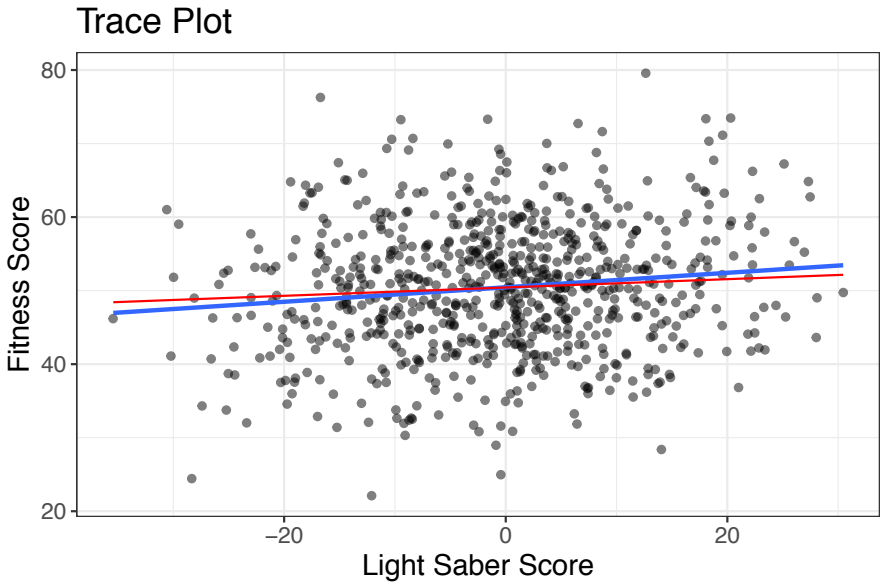

Figure 6. The LVM-implied fit between fitness score and light saber score, shown in red. The blue line represents the regression line between the two variables. The more closely the model-implied fit line resembles the regression line, the better the fit of the LVM.

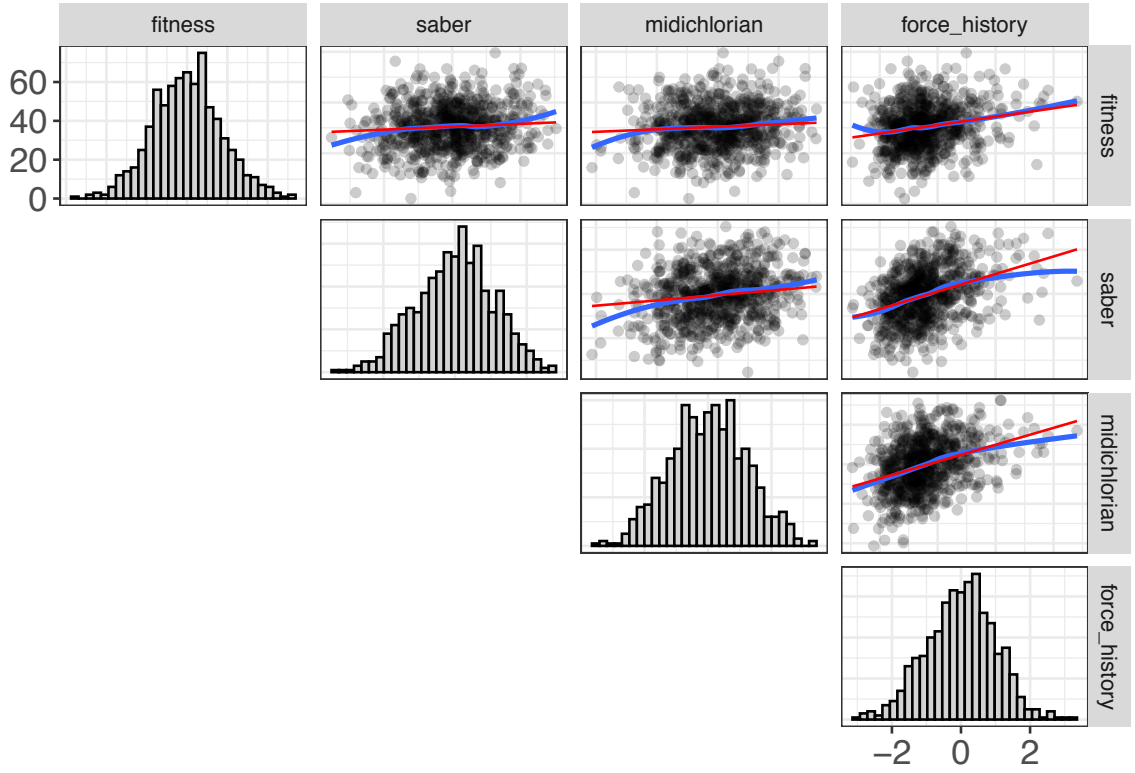

Figure 7. Scatterplot matrix showing the model-implied fit (red) and loess lines (blue) between four simulated indicator variables. These four variables are the indicators for the force latent variable. The diagonals show the histograms of the ICRs.

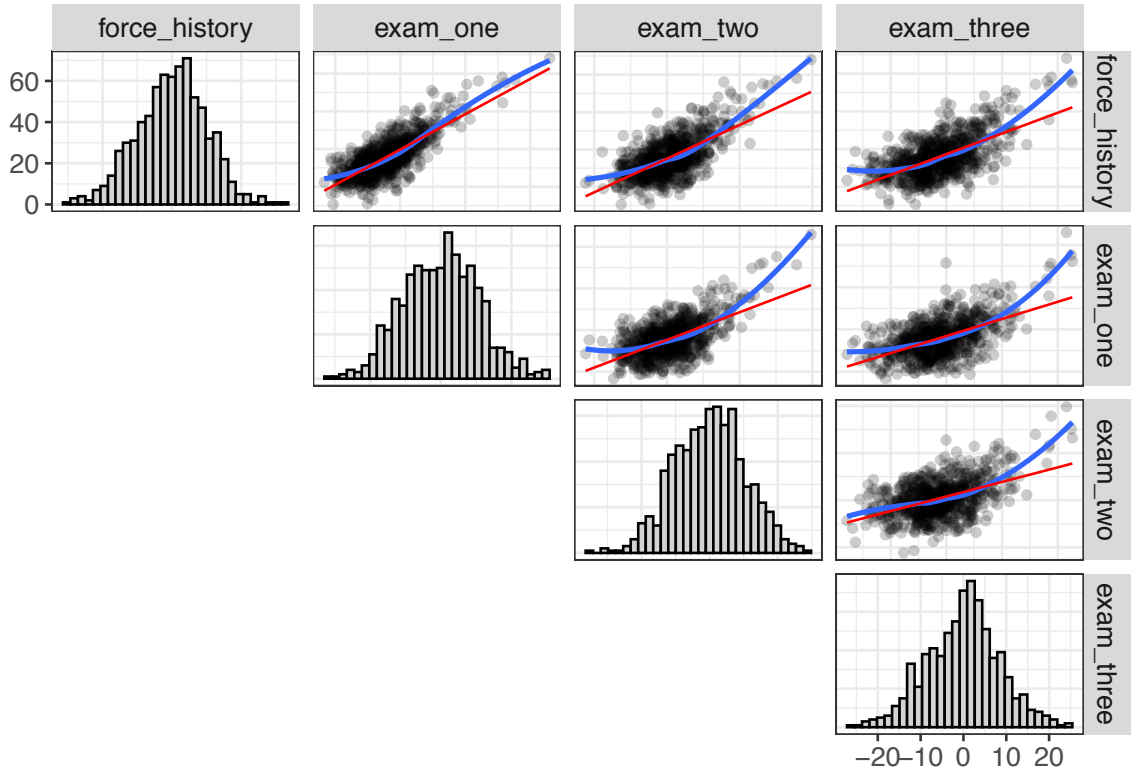

Figure 8. Scatterplot matrix showing the model-implied fit (red) and regression-implied fit (blue) between four simulated indicator variables. These four variables are the indicators for the Jedi latent variable. The diagonals show the histograms.

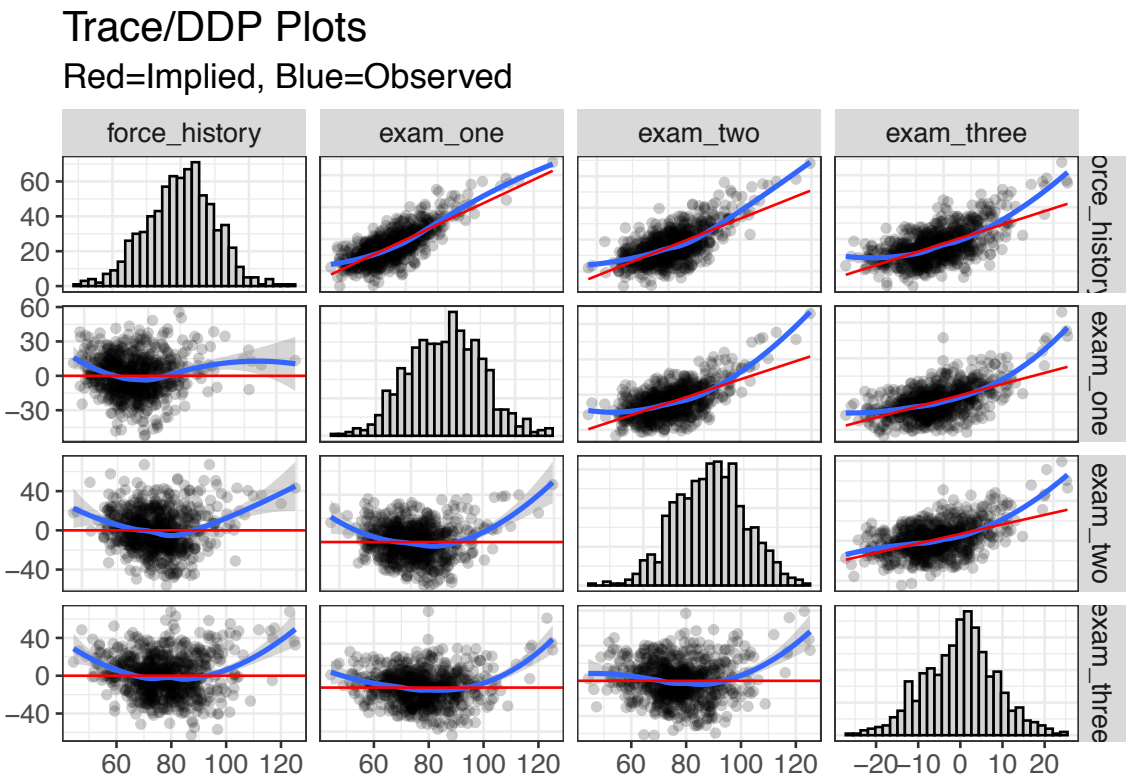

Figure 9. The upper triangle of this plot is the same as the plot shown in Figure 8. However, the lower triangle adds the disturbance-dependence plots.

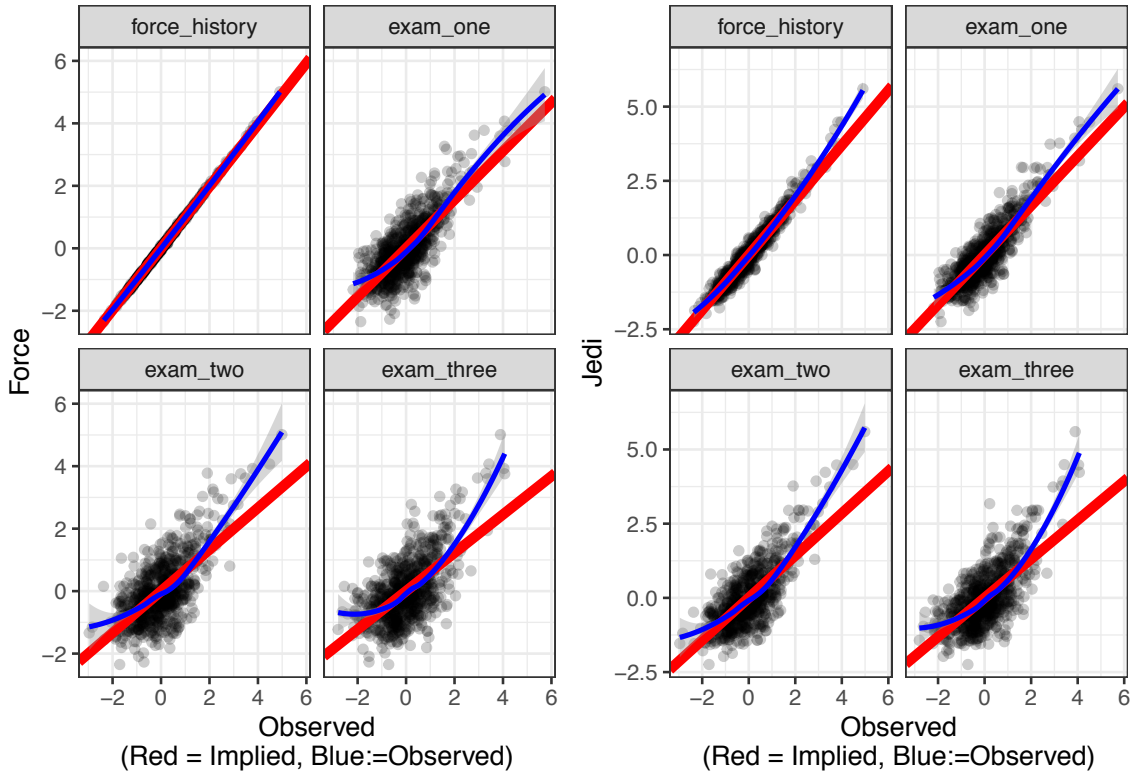

Figure 10. This image, called a measurement plot, shows the relationship between the latent variables (Y axis) and each standardized indicator. The blue lines are loess lines, while the red lines are the model-implied fits of the model.

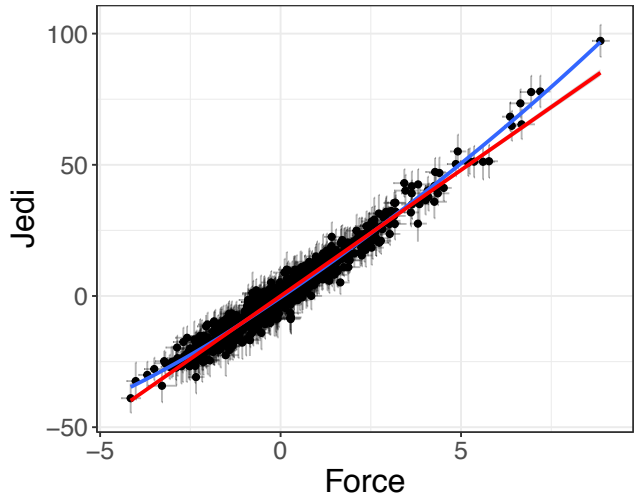

Figure 11. Structural or "cross-hair" plot of the relationship between the latent variables force and Jedi. The crosshairs represent the prediction intervals for the factor scores of the latent variables.

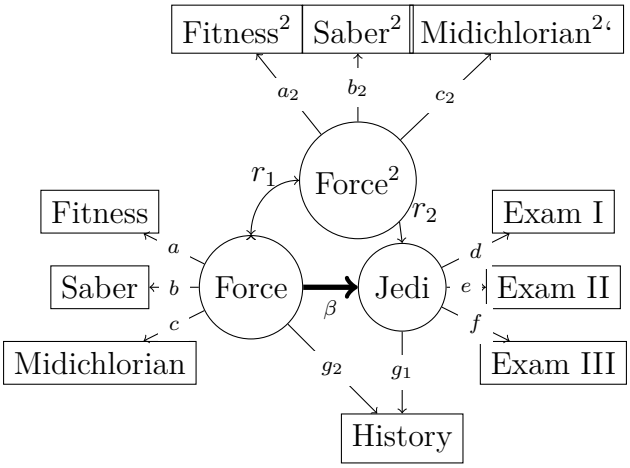

Figure 12. Nonlinear model for the Jedi dataset. This model proposes a new latent variable (Force^2) that has saber, fitness, and midichlorian as squared indicators. This model also allows history to load onto both force and Jedi.

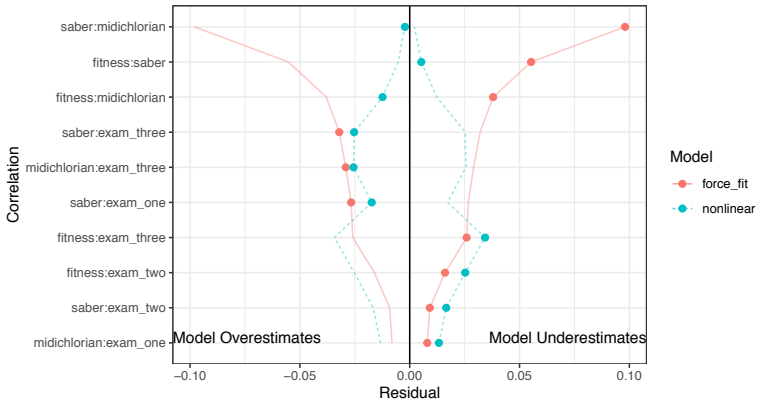

Figure 13. A hopper plot, but this time two models are being compared. The red line shows the residuals for the original model, while the blue lines show the residuals for the nonlinear model.

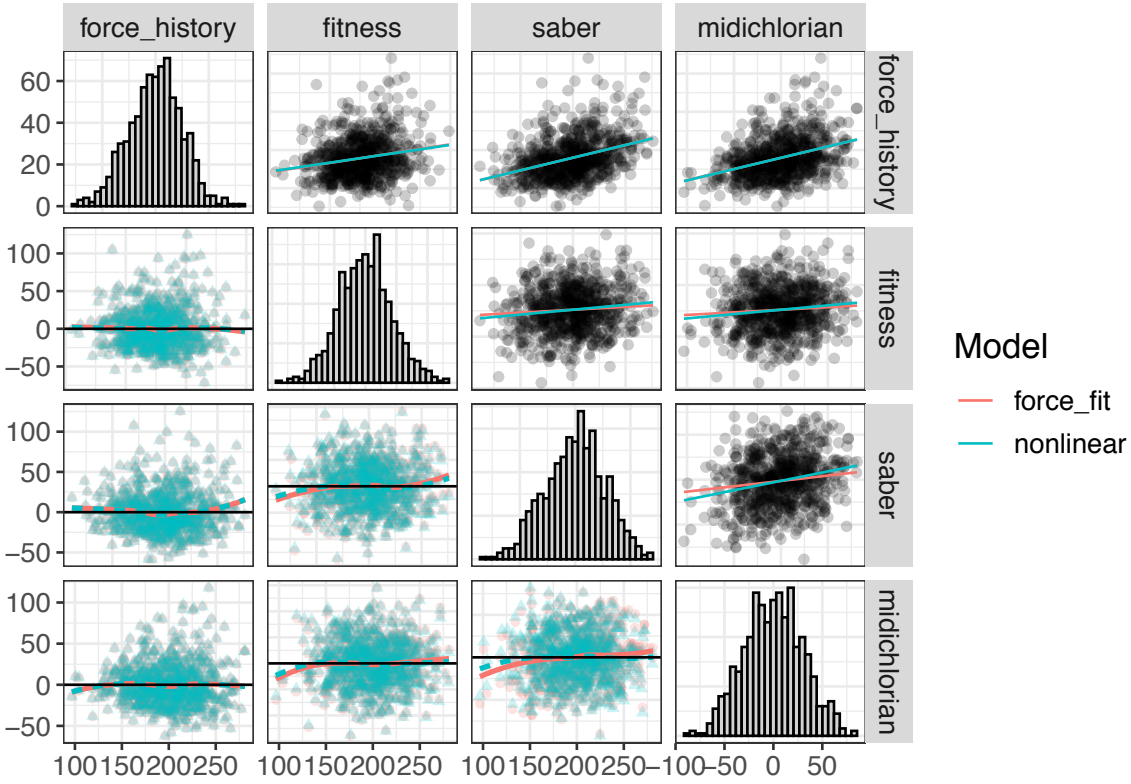

Figure 14. Trail plots and DDPs for the two models. The red shows the original model, while the blue shows the fit for the nonlinear model. Trail/DDPs show very little difference between the two models, indicating the two models seem to have similar implied variance/covariance matrices.

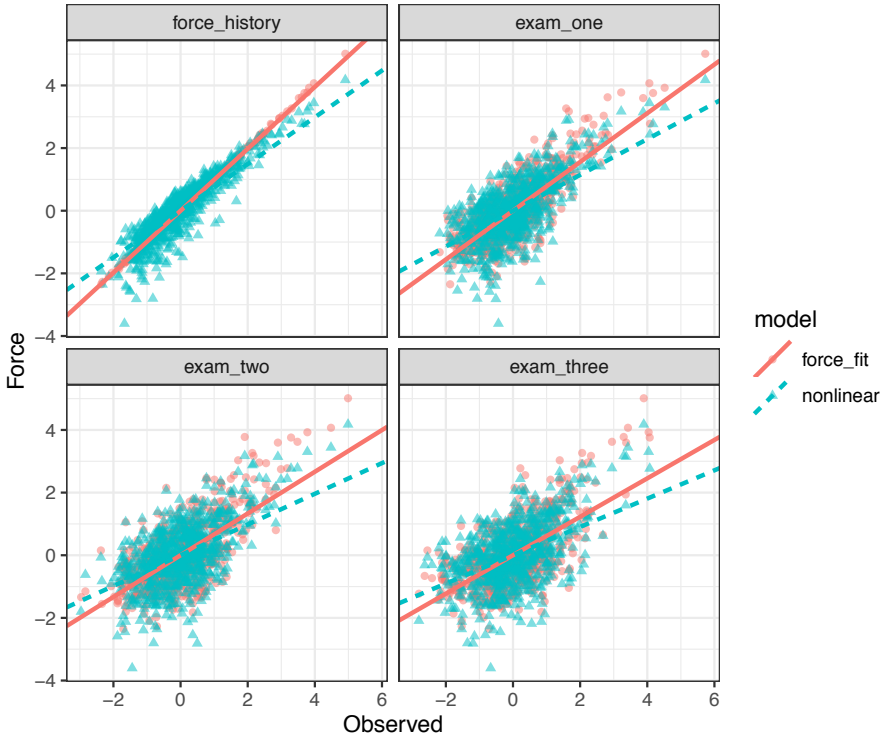

Figure 15. Measurement plots for the force latent variable, with both the nonlinear, red, model and original, blue, model.

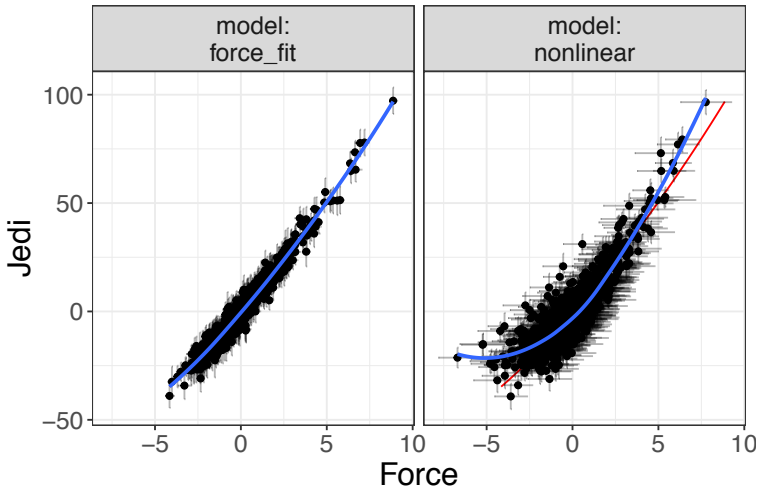

Figure 16. Relationship between the two latent variables for the original model (left) and the nonlinear model (right). The red line is a 'ghost line,' which simply repeats the line from the left plot to the right plot. This line makes it easier to compare the fits of the models across panels.

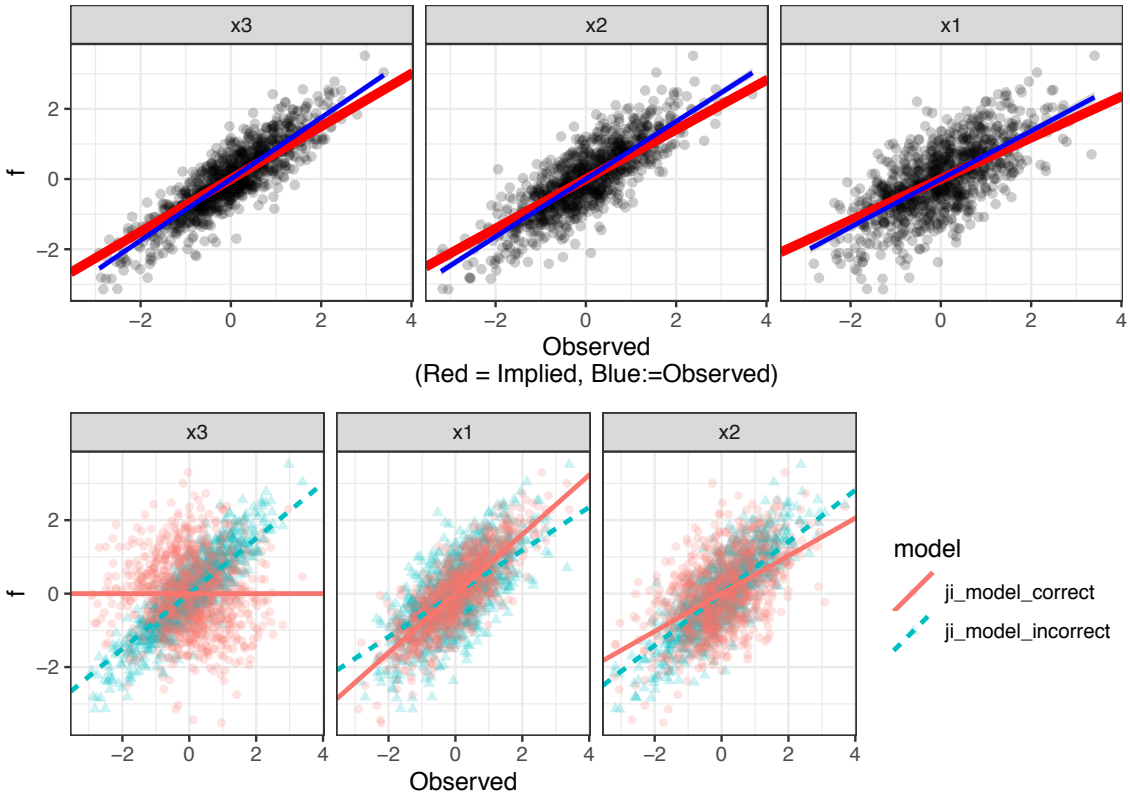

Figure 17. Measurement plots of the user-specified model in Figure 2. While fit statistics will not reveal any problems, the measurement plots do. The top plot shows that the model-implied fit consistently underestimates the relationship between the observed and latent variable. The bottom plot compares the true model to the proposed model and reveals the two models say very different things about the relationships between the observed and latent variable.

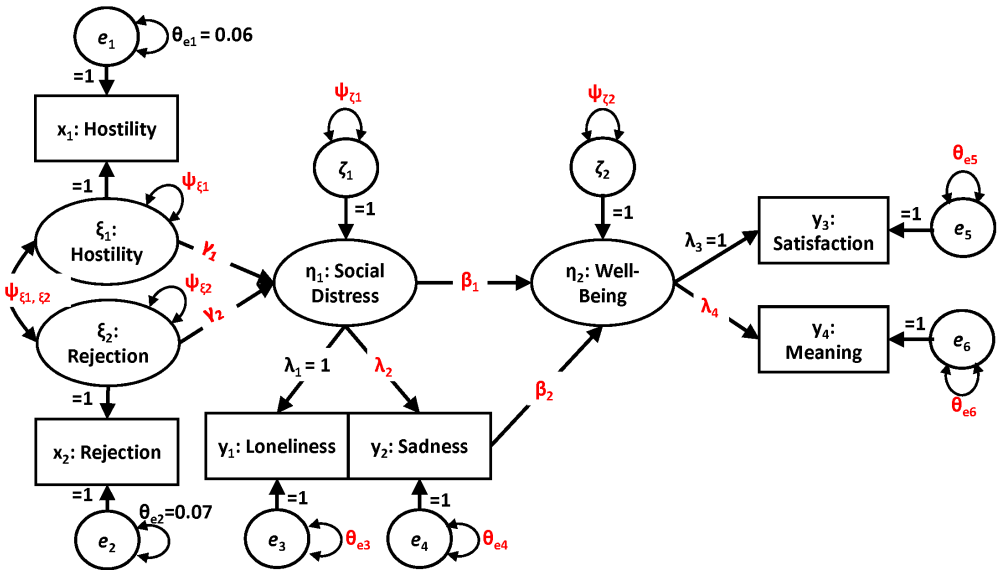

Figure 18. Hypothesized Model of the Effect of Social Distress on Emotional Well-Being with the NIH Toolbox Norming Data Set. In this model, Perceived Rejection and Perceived Hostility are treated as causal indicators of the Social Distress latent factor. Estimated parameters have red labels.

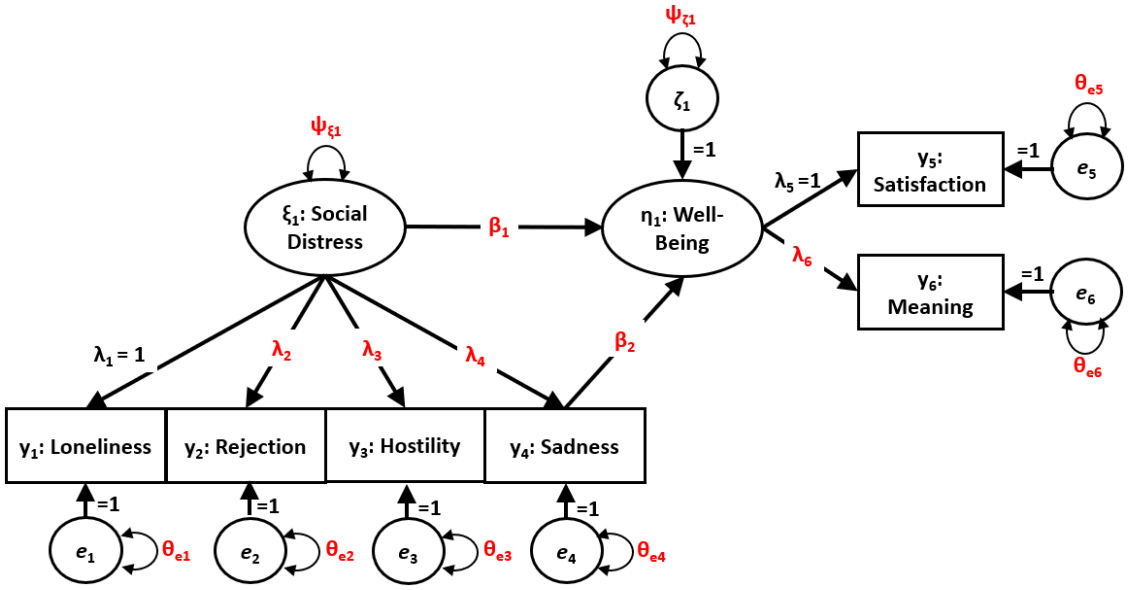

Figure 19. Alternative Model of the Effect of Social Distress on Emotional Well-Being with the NIH Toolbox Norming Data Set. In this model, Perceived Rejection and Perceived Hostility are treated as reflective indicators of the Social Distress latent factor rather than causal indicators. Estimated parameters have red labels.

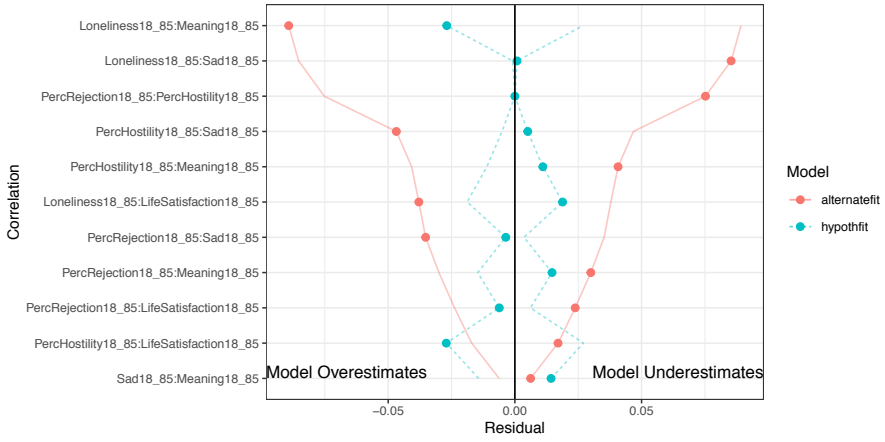

Figure 20. Hopper plot of the hypothesized and alternative models.

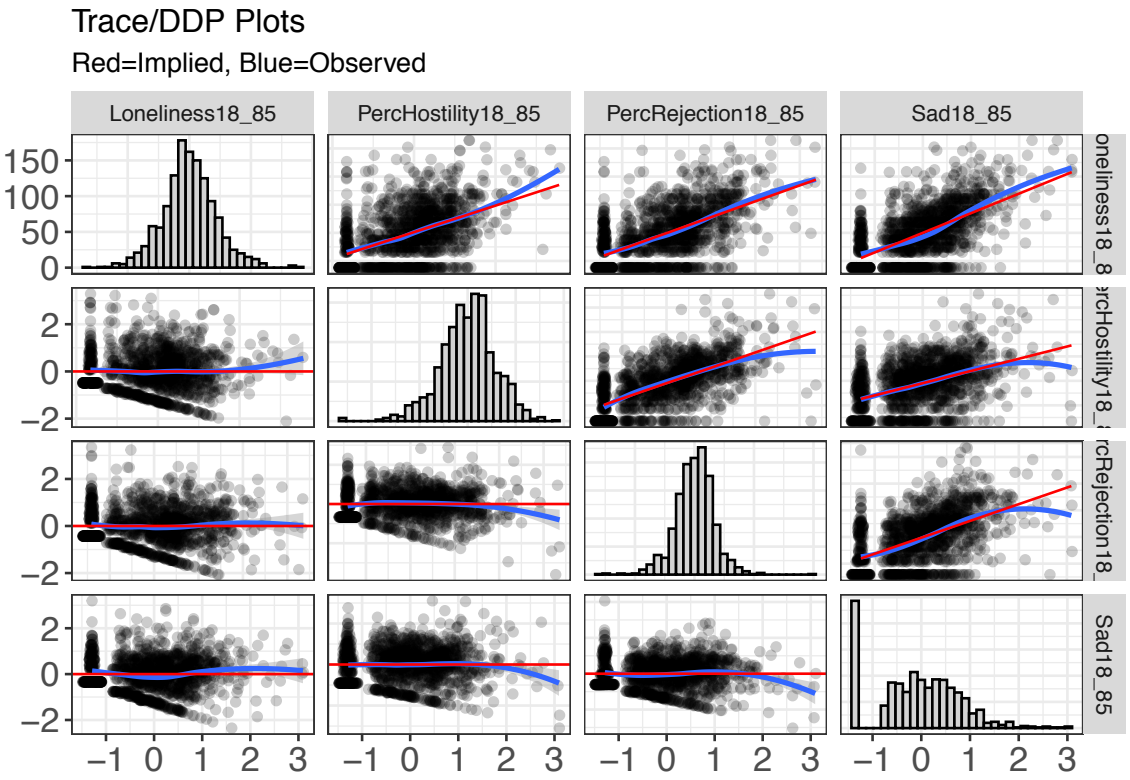

Figure 21. Trail/DDPs of the hypothesized model for the indicators of the social distress latent factor.

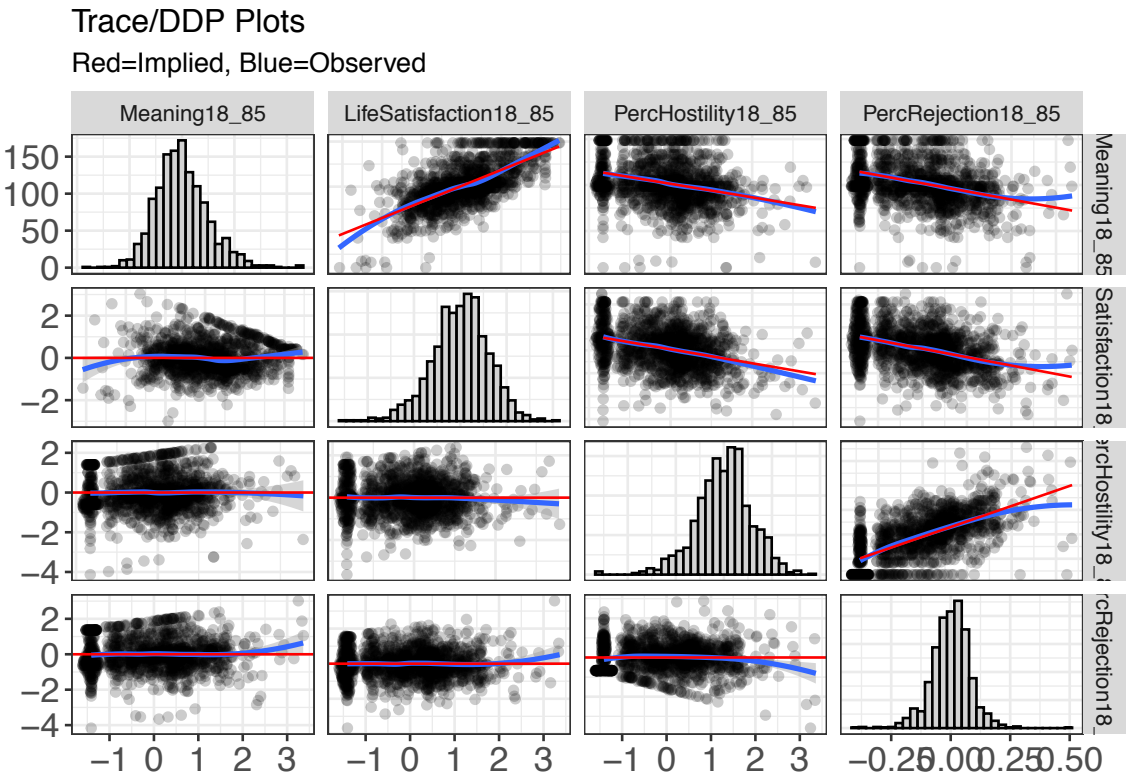

Figure 22. Trail/DDPs of the hypothesized model for the indicators of the social distress latent factor.

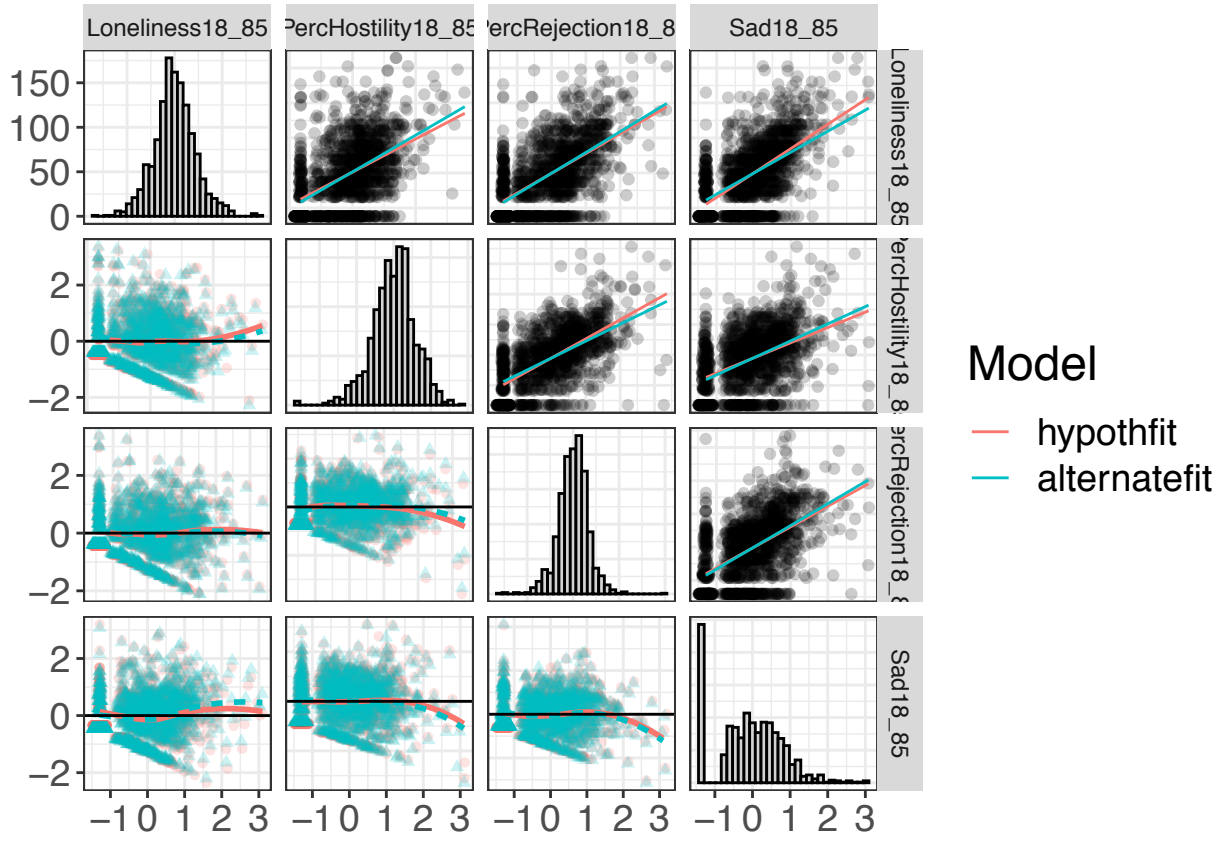

Figure 23. Trail/DDPs for both the hypothesized and altnerative model for the distress latent factor.

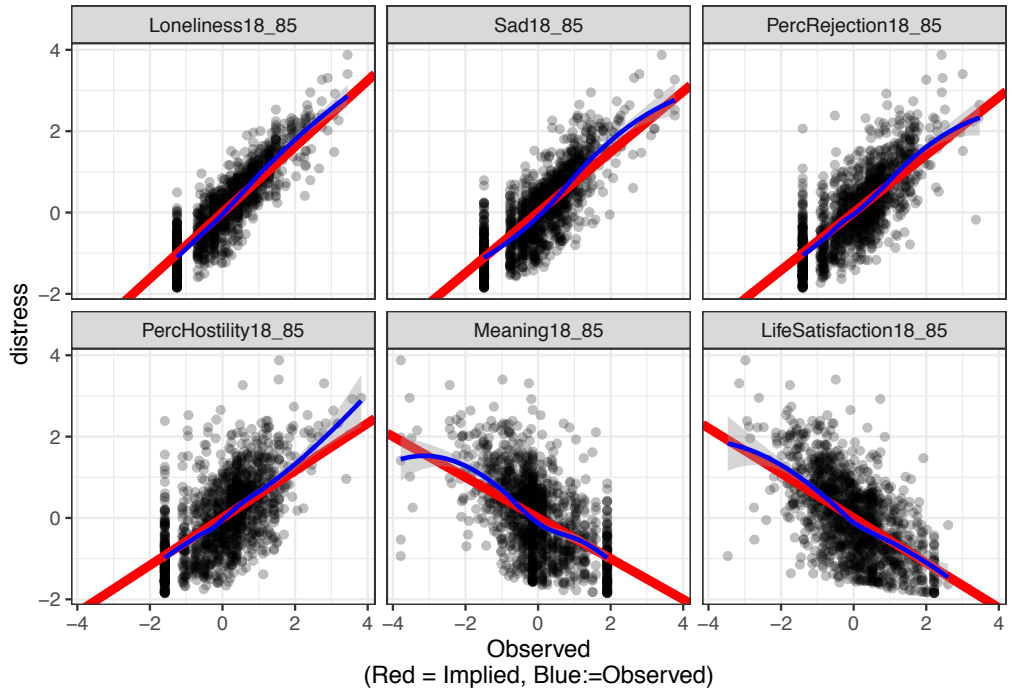

Figure 24. Measurement plot for the hypothesized model, showing the relationship between several indicators and the distress latent factor.

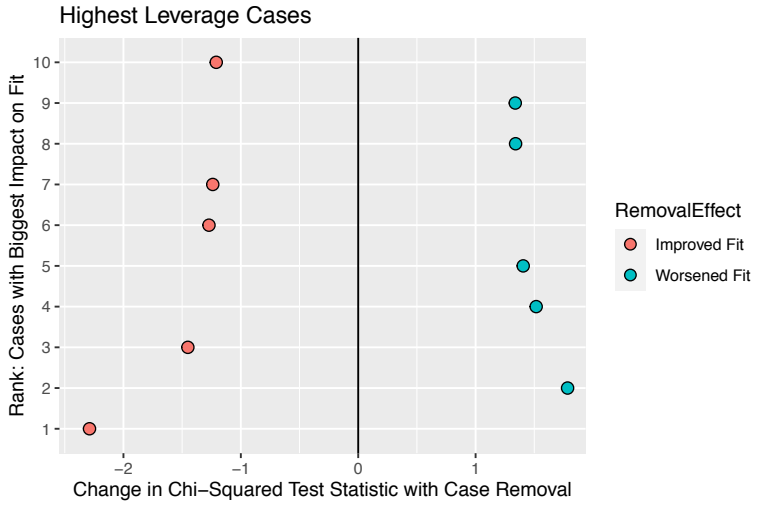

Figure 25. Influence plot for the hypothesized model.