Extended Model Descriptions

Factored Structural Equation Modeling in Blimp

Craig K. Enders

University of California, Los Angeles

Brian T. Keller

University of Missouri

Contents

1. Probit Regression

Blimp defaults to probit regression method for binary and ordered categorical outcomes, and it also uses a multinomial probit model for multicategorical exogenous variables. The probit link for binary and ordered categorical outcomes introduces a normally distributed latent response variable representing a continuous dimension of the measured characteristic. To illustrate, consider the model in Figure 4a from the paper, where M is a binary dummy code, and the factor indicators are five-point rating scales. The four measurement models feature latent response dependent variables, as does the empty model for the binary predictor. The regression equations below use an asterisk superscript to denote the underlying latent variables.

$$\begin{gathered} X_{p,i}^{*} = \beta_{0X_p^*} + \beta_{1X_p^*}(\eta_i) + \varepsilon_{X_p^*,i} = \bar{X}_{p,i}^{*} + \varepsilon_{X_p^*,i} \\ M_i^* = \beta_{0M^*} + \varepsilon_{M^*,i} = \bar{M}_i^* + \varepsilon_{M^*,i} \end{gathered} \tag{S1}$$

The figure uses ellipses to represent latent responses and rectangles to denote discrete variables. For identification, both residuals are defined as standard normal variables (i.e., latent response variables are standardized within subgroups defined by the same predicted value).

The probit model also introduces threshold parameters that segment the latent response distribution. The areas under the normal curve between these thresholds define the discrete category probabilities. The binary covariate requires a single threshold, and five-point indicators require four cutpoints each. The following expression gives the relationship between a latent response variable V^* and a discrete variable V with response options numbered $c = 1$ to C.

$$V_i = c \text{ if } \tau_{c-1} < V_i^* \leq \tau_c \tag{S2}$$

In words, the equation says that the latent response scores for discrete category c fall between two adjacent cutpoints in the normal curve, τ_{c-1} and τ_c. The lowest category $c = 1$ is bounded by an imaginary threshold at negative infinity (i.e., $\tau_0 = -\infty$), and the highest category $c = C$ is similarly bounded at positive infinity (i.e., $\tau_C = +\infty$). To identify the mean structure, Blimp fixes τ_1 to 0. Thresholds are estimated following the procedure in Cowles (1996).

Returning to the factorization in the bottom line of Equation 9, the bivariate normal distribution in the rightmost term now describes a latent factor and a latent response variable.

$$f(\eta, M) \rightarrow \begin{cases} \mathcal{N}\left(\begin{pmatrix} \bar{\eta}_i \\ \bar{M}_i^* \end{pmatrix}, \begin{pmatrix} \sigma_{\varepsilon_\eta}^2 & \sigma_{\varepsilon_\eta \varepsilon_M} \\ \sigma_{\varepsilon_M \varepsilon_\eta} & 1 \end{pmatrix} \right) \\ M = \begin{cases} 1 \text{ if } M_i^* \geq \tau_1 \\ 0 \text{ if } M_i^* < \tau_1 \end{cases} \end{cases} \tag{S3}$$

Notice that $\sigma_{\varepsilon_M}^2$ in the lower diagonal element of the covariance matrix is fixed at 1 for identification (because M^* has no predictors, the variance of ε_M is the total variance). The second line of the expression enforces Equation S2, ensuring that latent responses are confined to the correct region of normal curve. That is, the latent score for a person with $M = 0$ must fall between negative infinity and τ_1, and the score for $M = 1$ is bounded by τ_1 and positive infinity.

The distributions of the conditionally independent indicators, X_1 and X_2, are univariate normal curves. Predicted latent response scores define the conditional means, and the residual variances are again fixed at 1 for identification.

$$\begin{aligned} f(X_1|\eta) &\rightarrow \begin{cases} \mathcal{N}\left(\bar{X}_{1,i}^*, 1\right) \\ X_{1,i} = c \text{ if } \tau_{1,c-1} < X_{1,i}^* \leq \tau_{1,c} \end{cases} \\ f(X_2|\eta) &\rightarrow \begin{cases} \mathcal{N}\left(\bar{X}_{2,i}^*, 1\right) \\ X_{2i,} = c \text{ if } \tau_{2,c-1} < X_{2,i}^* \leq \tau_{2,c} \end{cases} \end{aligned} \tag{S4}$$

As before, the first thresholds (τ_{11} and τ_{21}) are fixed at 0 for identification, and the categorization functions ensure that latent response scores are confined to the correct area of normal curve. Finally, the correlated indicators, X_3 and X_4, share a bivariate normal distribution like the one in Equation S3. In this case, both diagonal elements of the residual covariance matrix are constrained to 1 for identification, and the off-diagonal is the residual correlation on the latent response metric.

2. Multinomial Probit Model for Multicategorical Exogenous Predictors

For multicategorical exogenous variables, Blimp uses a multinomial model that assigns a latent response variable to each discrete category (Aitchison & Bennett, 1970). These so-called indicants or utilities represent a continuous propensity for category membership. Following the logic of dummy coding, the full set of utilities is replaced by one fewer latent difference scores. Due to its suitability for data augmentation, the multinomial probit model is a popular tool for missing data imputation (Carpenter et al., 2023; Carpenter et al., 2011; Enders, 2022; Enders et al., 2018; Goldstein et al., 2009; Quartagno & Carpenter, 2019). Blimp currently uses the model exclusively for this purpose.

To illustrate the multinomial model, consider the model in Figure 5 in the main paper. Assume the factor indicators are ordinal rating scales, and M_1 and M_2 are dummy codes representing three nominal categories, coded 0, 1, and 2. The factorization in the bottom line of Equation 6 of the main paper also applies to this model. The term $f(M_1, M_2)$ now references the multivariate submodel for the incomplete dummy codes (or equivalently, the multicategorical variable, M). The multinomial probit model includes two latent difference scores that contrast an individual's propensities for belonging to groups $M_1 = 1$ and $M_2 = 1$ relative to the reference group with $M_1 = 0$ and $M_2 = 0$.

For this example, Blimp uses a multinomial model that assigns a latent response variable to each of the three discrete categories: U_0^*, U_1^*, and U_2^*. Paralleling the logic of dummy coding, the predictor model uses one fewer latent difference scores, with the difference scores subtracting the reference group utility from the other two.

$$\begin{aligned} M_{1,i}^* &= U_{1,i}^* - U_{0,i}^* \\ M_{2,i}^* &= U_{2,i}^* - U_{0,i}^* \end{aligned} \tag{S5}$$

The multinomial model does not use threshold cutpoints. Instead, an individual's largest latent utility determines their group membership. For example, a reference group member would have $U_{0,i}^* > U_{1,i}^*$ and $U_{0,i}^* > U_{2,i}^*$ (or equivalently, $M_{1,i}^* < 0$ and $M_{2,i}^* < 0$). Similarly, a person with $M_{1,i} =$

1 would have $U^*_{1,i} > U^*_{0,i}$ and $U^*_{1,i} > U^*_{2,i}$ (or equivalently, $M^*_{1,i} > 0$ and $M^*_{1,i} < M^*_{2,i}$), and an individual with $M_{1,i} = 2$ would have $U^*_{2,i} > U^*_{0,i}$ and $U^*_{2,i} > U^*_{1,i}$ (or equivalently, $M^*_{2,i} > 0$ and $M^*_{2,i} < M^*_{1,i}$).

$$M = \begin{cases} 0 \text{ if } U^*_{0,i} > U^*_{1,i} \text{ and } U^*_{0,i} > U^*_{2,i} \text{ or } M^*_{1,i} < 0 \text{ and } M^*_{2,i} < 0 \\ 1 \text{ if } U^*_{1,i} > U^*_{0,i} \text{ and } U^*_{1,i} > U^*_{2,i} \text{ or } M^*_{1,i} > 0 \text{ and } M^*_{1,i} < M^*_{2,i} \\ 2 \text{ if } U^*_{2,i} > U^*_{0,i} \text{ and } U^*_{2,i} > U^*_{1,i} \text{ or } M^*_{2,i} > 0 \text{ and } M^*_{2,i} < M^*_{1,i} \end{cases} \tag{S6}$$

The latent difference scores share the multivariate normal distribution shown below.

$$\begin{aligned} f(M_1, M_2) \rightarrow\ & \mathcal{N}\left(\begin{pmatrix} \bar{M}^*_{1,i} \\ \bar{M}^*_{2,i} \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right) \\ & \times \big(I(M^*_{1,i} < 0 \;\&\; M^*_{2,i} < 0) \times I(M_{1,i} = 0 \;\&\; M_{2,i} = 0) \\ & + I(M^*_{1,i} > 0 \;\&\; M^*_{1,i} > M^*_{2,i}) \times I(M_{1,i} = 1) \\ & + I(M^*_{2,i} > 0 \;\&\; M^*_{2,i} > M^*_{1,i}) \times I(M_{2,i} = 1)\big) \end{aligned} \tag{S7}$$

Because the latent difference scores are unobserved, the covariance matrix elements are fixed for identification. The values in the covariance matrix follow from adopting a diagonal covariance matrix for the latent utilities. Enders (2022, pp. 244–248) provides additional details. The $I(\cdot)$ terms are indicator functions that work like true or false statements. In this case, the indicator functions ensure that the sign and magnitude of the difference scores is logically consistent with the individual's dummy codes. That is, the indicator functions restrict latent difference score imputations to the region of the multivariate normal distribution defined by Equation S6.

The Blimp syntax excerpt below specifies the model in Figure 5. To simplify specification, the software creates a set of dummy codes for binary and multicategorical variables listed on the `NOMINAL` command line, defining the lowest numeric code as the reference group. Listing the categorical variable as a predictor in the `MODEL` section introduces these dummy codes into the regression equation.

```
# model 6 excerpt
NOMINAL: m;
ORDINAL: x1:x4;
LATENT: eta;
MODEL:
eta ~~ eta@1;
eta ~ m;
eta -> x1@lo1 x2:x4;
x3 ~~ x4;
y ~ eta m;
```

The dummy codes can also be manually entered into an equation by appending their corresponding numeric codes to the predictor's name (e.g., `eta ~ m.1 m.2`). To facilitate model-wide missing data handling, the software automatically constructs the multinomial probit model without user input, and it iteratively updates dummy codes to reflect the most current imputations. The software omits these unnecessary supporting models when appropriate, in which case the dummy codes are fixed across iterations.

3. Logistic Regression

Blimp also provides logistic regression models for binary and multicategorical outcomes. MCMC estimation uses a data augmentation strategy like that of the probit model, but the composition of the latent response scores is more complex. To illustrate binary logistic regression, consider a modified version of the model in Figure 5 with a dichotomous outcome coded $Y = 0$ or $Y = 1$. The factorization in the bottom line of Equation 6 also applies to this model. In logistic regression, the transformed variable is a log-odds or logit instead of a normally distributed latent variable. The regression equation in the first line below shows the predicted log-odds as linear function of the regressors.

$$
\begin{gathered}
\ln\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right) = \beta_{0Y} + \beta_{1Y}(\eta_i) + \beta_{2Y}\left(M_{1,i}\right) + \beta_{3Y}\left(M_{2,i}\right) = \bar{Y}_i^* \\
P(Y_i = 1) = \frac{1}{1 + \exp(-\bar{Y}_i^*)} \\
f(Y|\eta, M) \rightarrow \text{Bernoulli}\big(P(Y_i = 1)\big)
\end{gathered}
\tag{S8}
$$

The inverse link function on the second line converts the continuous outcome to a predicted probability. Finally, the third line defines the conditional distribution of the binary outcome as a Bernoulli function that depends on the predicted probability.

Blimp uses the data augmentation method outlined by Polson et al. (2013) and Windle et al. (2014) for logistic regression. Asparouhov and Muthén (2021) provide an accessible description of this approach, which is also available in M*plus*. A key challenge for MCMC estimation is that a normal prior distribution for regression coefficients is not conjugate to the logistic data model (likelihood) in the bottom line of Equation S8. This mismatch requires inefficient estimation algorithms that introduce unique computational challenges. Instead, data augmentation introduces an auxiliary latent variable, A, that transforms the Bernoulli data distribution into normal, but heteroscedastic, distribution for the latent response variable, Y^*. This modified data distribution is compatible with a normal prior, and the logistic coefficients can be estimated using simple Gibbs sampling routines for linear regression. At a high level, the procedure is essentially a mathematical trick that uses latent response scores to transform a complex generalized linear model estimation problem into a more straightforward linear model problem. The latent response scores themselves do not have an intuitive interpretation like they do in the probit model.

The key feature of Polson's data augmentation method is that it expresses the original logistic model as a linear regression with a normally distributed latent response outcome. The equations below show the latent response variable, the reparameterized logistic model, and its corresponding normal conditional distribution.

$$Y_i^* = \frac{Y_i - 0.5}{A_i}$$

$$Y_i^* = \beta_{0Y} + \beta_{1Y}(\eta_i) + \beta_{2Y}(M_{1,i}) + \beta_{3Y}(M_{2,i}) + \varepsilon_{Y,i} = \bar{Y}_i^* + \varepsilon_{Y^*,i} \tag{S9}$$

$$f(Y|\eta, M, A) \rightarrow \mathcal{N}\left(\bar{Y}_i^*, \frac{1}{A_i}\right)$$

In the first line, the latent response variable is a function of the binary outcome and auxiliary latent variable, which is imputed in a prior step (see below). Like the probit model, the latent response variable serves as the outcome in a linear regression model, as shown in the second line. The normal curve function in the third line indicates that errors are normal, but heteroscedastic; the spread of each person's error distribution is determined by their auxiliary latent variable[1]. To reiterate, this specification is simply a mathematical device that transforms a complex generalized linear model estimation problem into a more straightforward linear model problem. Across many realizations of the auxiliary variable, the coefficients from the transformed model reproduce those of the target logistic model, $f(Y|\eta, M)$.

[1] Asparouhov and Muthén (2021b) show an equivalent version of this specification where the residual variance is constant at 1 and the data are rescaled by multiplying Y_i^*, η_i, and M_i by $\sqrt{A_i}$.

Polson et al. (2013) define the auxiliary variable A as a Pólya-gamma random variable, which is a weighted sum of an infinite number of gamma random variables. Visually, the Pólya-gamma is right-skewed distribution with a shape parameter that controls symmetry and a second parameter that determines peakedness. In a binary logistic model, the Pólya-gamma distribution's shape is a function of the predicted log-odds, $\bar{Y}_i^*$. Each MCMC iteration updates (imputes) the auxiliary variable scores using the current regression coefficients. Imputing the auxiliary variable scores is computationally intensive because each person's value involves a sum over a massive (infinite) number of random numbers. Windle et al. (2014) describe multiple strategies for simulating Pólya-gamma variables, but Blimp uses the method from Polson et al. (2013). The discrete imputations themselves are sampled from a Bernoulli distribution.

The Blimp syntax excerpt below specifies the logistic structural model. Listing an outcome variable on the `NOMINAL` command line triggers logistic regression. All other aspects of the script remain unchanged.

```
# model 7 excerpt
NOMINAL: m y;
ORDINAL: x1:x4;
LATENT: eta;
MODEL:
eta ~~ eta@1;
eta ~~ m;
eta -> x1@lo1 x2:x4;
x3 ~~ x4;
y ~ eta m;
```

To reiterate, the underlying factored regression specification, which now integrates logit, probit, and normal submodels, is entirely hidden from view.

Next, consider a multicategorical outcome variable with K groups, numbered $k = 1$ to K. The factorization in the bottom line of Equation 6 also applies to this model. The previous Blimp script remains unchanged; the software automatically applies multinomial logistic regression, treating the lowest numeric code as the reference group. The first equation below shows the predicted log-odds of group k (for $k > 1$) versus the reference group ($k = 1$) as a linear function of the regressors. The second equation gives group k's predicted probability[2], and the third equation represents a multinomial conditional distribution with the predicted probabilities as its parameters.

$$\ln\left(\frac{P(Y_i = k | k \in 2, \dots, K)}{P(Y_i = 1)}\right) = \beta_{0Y_k} + \beta_{1Y_k}(\eta_i) + \beta_{2Y_k}(M_{1,i}) + \beta_{3Y_k}(M_{2,i}) = \bar{Y}^*_{k,i}$$

$$P(Y_i = k | k \in 2, \dots, K) = \frac{\exp(\bar{Y}^*_{k,i})}{1 + \sum_{k=2}^{K} \exp(\bar{Y}^*_{k,i})} \tag{S10}$$

$$f(Y|\eta, M) \rightarrow \text{Multinomial}\big(P(Y_i = 1), P(Y_i = 2), \dots, P(Y_i = K)\big)$$

Blimp again uses Pólya-gamma auxiliary variables to reparameterize the K – 1 logistic regression equations into a corresponding set of linear models with latent response variables as the outcomes. The coefficients are estimated one equation at a time using a sequential approach, which applies the binary logistic regression procedure to each of the K – 1 equations in turn. The technical supplement from Polson et al. (2013) describes a modified form of the multinomial distribution that supports this sequential estimation, as do Asparouhov and Muthén (2021)

[2] The predicted probability for $k = 1$ is obtained by subtracting the other probabilities from 1.

4. Negative Binomial Regression

Blimp also offers negative binomial regression models for count outcomes—dependent variables that take nonnegative integer values representing the number of times an event occurs (e.g., the number of heavy drinking days per month or aggressive acts during a play period). Researchers frequently use Poisson regression for this type of data. The Poisson distribution is defined by a single parameter, μ, which represents both the mean and variance of the counts. In practice, however, count data often display greater variability than the Poisson model allows (a phenomenon known as overdispersion). Overdispersion can arise from factors such as model misspecification or dependent events within individuals (Coxe et al., 2009). Negative binomial regression provides a more flexible alternative by adding a dispersion parameter that captures heterogeneity among individuals with the same predicted count. When the dispersion parameter equals 0, the negative binomial model reduces to the Poisson model.

To illustrate negative binomial regression, consider a modified version of the model in Figure 5 of the main paper where Y is a count outcome. The negative binomial regression equation below shows the logarithm of the predicted count (μ_i) as linear function of the regressors.

$$\begin{gathered}\ln(\mu_i) = \beta_{0Y} + \beta_{1Y}(\eta_i) + \beta_{2Y}\left(M_{1,i}\right) + \beta_{3Y}\left(M_{2,i}\right) = \bar{Y}_i^* \\ f(Y|\eta, M) \rightarrow \text{NB}(\bar{Y}_i^*, \phi)\end{gathered} \tag{S11}$$

The factorization in the bottom line of Equation 6 also applies here, but the model-implied conditional distribution of Y changes from a normal curve to a negative binomial distribution with the predicted log-count and dispersion as its parameters ($\bar{Y}_i^*$ and ϕ, respectively).

Blimp once again uses data augmentation with an auxiliary latent variable to simplify estimation of the regression coefficients. Pillow and Scott (2012) and Neelon (2019) detail this approach for negative binomial regression models, as do Asparouhov and Muthén (2021). The procedure is essentially the same as binary logistic regression with two modifications. First, the Pólya-gamma distribution used to generate (impute) the auxiliary latent variable has different shape and peakedness parameters. Second, the latent response variable definition from the top line of Equation S9 changes to incorporate the modified auxiliary variable and dispersion parameter. The latent response model and normal conditional distributions on the second and third line of Equation S9, respectively, are unchanged (see Neelon, 2019, pp. 834–835). To reiterate, the latent variable specification is merely a mathematical device that simplifies estimation of the regression coefficients. Finally, the dispersion parameter is iteratively updated using conditional distributions found in Asparouhov and Muthén (2021)

The Blimp syntax excerpt below specifies a structural model with a count outcome. Listing an outcome variable on the `COUNT` command line triggers negative binomial regression. All other aspects of the script remain unchanged.

```
# model 8 excerpt
NOMINAL: m;
ORDINAL: x1:x4;
COUNT: y;
LATENT: eta;
MODEL:
eta ~~ eta@1;
eta ~~ m;
eta -> x1@lo1 x2:x4;
x3 ~~ x4;
y ~ eta m;
```

As mentioned elsewhere, the only current limitation with negative binomial regression is that incomplete count variables cannot influence downstream variables. Other generalized linear models do not share this limitation.

5. Heterogeneous Variation

Nonnormal data can arise for many reasons. One common source is heterogeneous residual variation that occurs when the spread of a variable differs systematically across individuals or groups. Even if residuals are conditionally normal, combining observations with different variances produces an overall distribution that departs from normality, often appearing skewed or heavy-tailed. Rather than transforming variables, an alternative strategy is to model this heterogeneity directly by allowing the residual variance to depend on covariates. This idea has a long history in the statistical literature on variance modeling (Cepeda & Gamerman, 2000; Goldstein, 2005; Harvey, 1976) and has recently gained attention in the SEM framework as a tool for capturing measurement or structural heterogeneity, most notably in moderated nonlinear factor analysis (MNLFA) applications (Bauer, 2017, 2023).

To illustrate, reconsider the model from Figure 5 in the main paper and the factorization in the bottom line of Equation 6. In addition to capturing group mean differences on the latent variable, we can expand the model to include factor variation differences. The first equation below is the structural model, and the second equation is a regression model linking the logarithm of the factor variance to the dummy codes—this is akin to a multiple group model where each group has a unique disturbance variance. Modeling the log-variance ensures that the predicted spread is always positive.

$$\begin{gathered}\eta_i = \beta_{0\eta} + \beta_{1\eta}\left(M_{1,i}\right) + \beta_{2\eta}\left(M_{2,i}\right) + \varepsilon_{\eta,i} = \bar{\eta}_i + \varepsilon_{\eta,i} \\ \ln\left(\sigma^2_{\varepsilon_{\eta,i}}\right) = \beta_{0\sigma^2_{\varepsilon_\eta}} + \beta_{1\sigma^2_{\varepsilon_\eta}}\left(M_{1,i}\right) + \beta_{2\sigma^2_{\varepsilon_\eta}}\left(M_{2,i}\right) \\ f(\eta|M_1, M_2) \rightarrow \mathcal{N}\left(\bar{\eta}_i, \sigma^2_{\varepsilon_{\eta,i}}\right)\end{gathered} \tag{S12}$$

On the second line, $\beta_{0\sigma^2_{\varepsilon_\eta}}$ is the expected log-variance for the reference group (a covariate profile comprised of 0s), and $\beta_{1\sigma^2_{\varepsilon_\eta}}$ and $\beta_{2\sigma^2_{\varepsilon_\eta}}$ represent group differences in the log-variance. The intercept is fixed at 0 for identification, which sets the baseline group's variance equal to 1 on the raw metric. This part of the model does not have a random error term; observation-specific variation is deterministic. Finally, notice that the normal distribution function in the third line features heteroscedastic variation where each observation or combination of covariate scores has a unique residual variance, $\sigma^2_{\varepsilon_{\eta},i}$. Manifest outcomes like the factor indicators can also have dispersion models.

Dispersion or scale models are difficult to implement in some software programs because their regression coefficients must be specified as nonlinear constraints on the variance parameters. In Blimp, the submodels for the log-variance are simply additional regression equations in the `MODEL` section. To illustrate, the syntax excerpt below specifies heteroscedastic factor variance by listing the latent variable's name within the `var()` function.

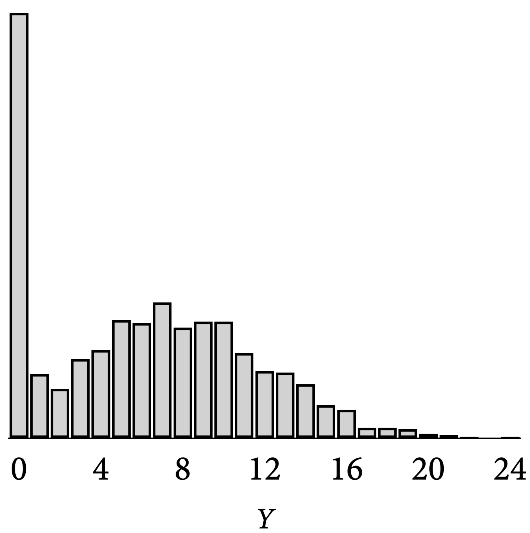
```
# model 10 excerpt
ORDINAL: x1:x4;
NOMINAL: m;
MODEL:
eta ~ m;
var(eta) ~ intercept@0 m;
eta -> x1@x1lo x2:x4;
x3 ~~ x4;
y ~ eta m;
```

Notice that the intercept of the latent variable's log-variance model (the $\beta_{0\sigma^2_{\varepsilon_{\eta}}}$ term in Equation S12) is constrained to 0 for identification (1 on the raw metric). Finally, recall that listing the multicategorical predictor on the `NOMINAL` command line automatically introduces dummy codes, with the lowest numeric code as the reference.

6. Two-Part Models

Two-part models are a strategy for nonnormal data resulting from floor or ceiling effects (often called semicontinuous data). The literature describes numerous variants of two-part models, both for single-level and multilevel regression models (Blozis et al., 2020; Muthén et al., 2024; Neelon & O'Malley, 2019; Olsen & Schafer, 2001; Smith et al., 2017). To illustrate, suppose an indicator has many score values at its minimum value, perhaps because it is measuring an extreme aspect of the construct that isn't relevant to all respondents. The histogram below illustrates this scenario.

Figure S1.



As its name implies, a two-part model represents an outcome, Y, as two separate variables. The first variable, U, indicates whether an observation is above the floor. In the graph above, minimum value of Y is 0, so U equals 0 if Y equals 0 and 1 otherwise. The second variable is a recoded version of the Y that is missing for every value at the floor. In the graph above, the recoded variable Y_i^* is missing if Y is 0 but otherwise equals Y. The two parts of the variable are as follows.

$$U_i = \begin{cases} 0 & \text{if } Y_i = \min(Y) \\ 1 & \text{if } Y_i > \min(Y) \end{cases} \qquad Y_i^* = \begin{cases} \text{NA} & \text{if } U_i = 0 \\ Y_i & \text{if } U_i = 1 \end{cases} \tag{S13}$$

These variables can be created outside of Blimp, or they can be generated by the software's `TRANSFORM` command as follows.

```
TRANSFORM:
u = ifelse(y == min(y), 0, 1);
yr = missing(u == 0, y);
ORDINAL: u;
```

The fitted models include a linear regression for the observations above the floor and a probit regression for the binary indicator. To illustrate, consider the model in Figure S1 above. The models below feature the same predictors in both equations but that need not be the case.

$$Y_i^* = \beta_{0Y^*} + \beta_{1Y^*}(\eta_i) + \beta_{2Y^*}(M_{1,i}) + \beta_{3Y^*}(M_{2,i}) + \varepsilon_{Y^*,i} = \bar{Y}_i^* + \varepsilon_{Y^*,i}$$

$$U_i^* = \beta_{0U^*} + \beta_{1U^*}(\eta_i) + \beta_{2U^*}(M_{1,i}) + \beta_{3U^*}(M_{2,i}) + \varepsilon_{U^*,i} = \bar{U}_i^* + \varepsilon_{U^*,i}$$

$$f(Y|\eta) \rightarrow \begin{cases} Y_i^* \sim \mathcal{N}\left(\bar{Y}_i^*, \sigma^2_{\varepsilon_{Y^*}}\right) \\ U_i^* \sim \mathcal{N}\left(\bar{U}_i^*, 1\right) \\ U = \begin{cases} 1 \text{ if } U_i^* > \tau_1 \\ 0 \text{ if } U_i^* > \tau_1 \end{cases} \end{cases} \tag{S14}$$

The syntax excerpt below assigns the latent variable as a predictor of the manifest outcomes listed on the right of the right-facing arrow. The model for U is a probit regression because the variable was listed on the `ORDINAL` line in the previous excerpt.

```
# model 20 excerpt
TRANSFORM:
```

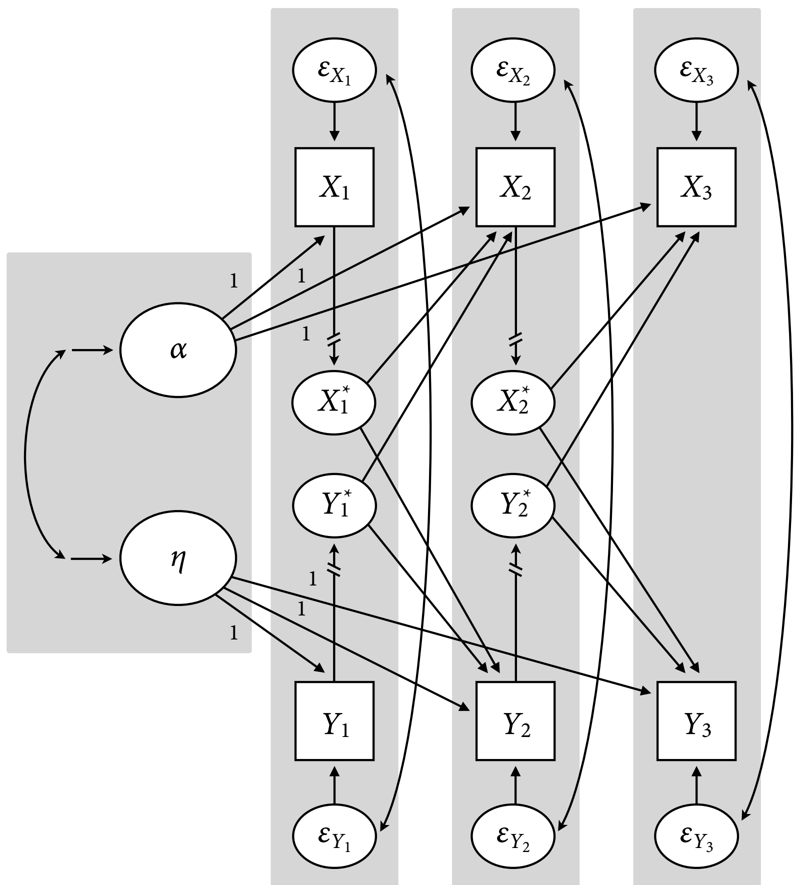
```
u = ifelse(y == min(y), 0, 1);
yr = missing(u == 0, y);
NOMINAL: m;
ORDINAL: u x1:x4;
LATENT: eta;
MODEL:
eta ~ m;
eta -> x1:x4;
x3 ~~ x4;
y ~ eta m;
u ~ eta m;
```

Following Olsen and Schafer (2001), the continuous part of the outcome could be log transformed if the variable is skewed. The two-part composition can also be paired with the Yeo–Johnson normalizing transformation.

7. Random Intercept Cross-Lagged Panel Model (RI-CLPM)

Keller and Enders (2021Ch. 5.19) illustrate the RI-CLPM extensions described in Mulder and Hamaker (2021). This section illustrates a basic random intercept cross-lagged panel model (RI-CLPM) with three repeated measurements on a pair of variables, X and Y. Figure S2 shows a path diagram of the model.

Figure S2.



In the RSEM version of the RI-CLPM, a latent variable at each measurement occasion t represents within-person variation that remains after subtracting out an individual's predicted value. Each predicted score is calculated as the sum of an occasion-specific mean and random intercept latent variable. The path diagram and the equations below denote these residual latent variables as X_t^* and Y_t^*. The square brackets below enclose the predicted values.

$$
\begin{aligned}
X_{t,i} &= \left[\mu_{X_t} + \eta_{X,i}\right] + X_{t,i}^* \\
Y_{t,i} &= \left[\mu_{Y_t} + \eta_{Y,i}\right] + Y_{t,i}^*
\end{aligned}
\tag{S15}
$$

The autoregressive ($\beta_{1X_t^*}$ and $\beta_{1Y_t^*}$) and cross-lagged ($\beta_{2X_t^*}$ and $\beta_{2Y_t^*}$) slopes in the structural model involve the within-person latent variables, as shown below.

$$
\begin{aligned}
X_{t,i}^* &= \beta_{1X_t^*}\left(X_{t-1,i}^*\right) + \beta_{2X_t^*}\left(Y_{t-1,i}^*\right) + \varepsilon_{X_t^*,i} \\
Y_{t,i}^* &= \beta_{1Y_t^*}\left(Y_{t-1,i}^*\right) + \beta_{2Y_t^*}\left(X_{t-1,i}^*\right) + \varepsilon_{Y_t^*,i}
\end{aligned}
\tag{S16}
$$

Importantly, the structural slopes represent pure within-person associations, free from stable individual differences. The occasion-specific correlated residuals reflect unexplained variation from the same source.

Blimp accommodates residualized associations, but it does so without residual latent variables. One of the unique features of Blimp's factorization is that a variable can exist in two states. For example, a binary predictor functions as a latent response variable in the exogenous variable submodel but as a dummy code in the structural model. In essence, variables with both incoming and outgoing arrows can occupy different metrics depending on the direction of the arrow. Residual associations can be incorporated by exploiting this property. In the RI-CLPM, the repeated measurements have incoming arrows from the random intercept latent variables in which the raw X_t and Y_t scores serve as the dependent variables. In first two manifest indicators also have an outgoing arrow to the next indicator in which X_t and Y_t. These are the autoregressive and cross-lagged paths. Importantly, the outgoing arrows reflect manifest variables that are centered at their predicted values (the bracketed terms in Equation S20). This centering defines the outgoing arrows as the effect of X_{t-1}'s residual on X_t and Y_t and the effect of Y_{t-1}'s residual on X_t and Y_t.

The top two lines of the following equation show the predicted values for occasion t – 1 (where t = 2 or 3). The second two lines show the structural regression equations. Although the raw X_t and Y_t variables are the outcomes, we continue to use an asterisk on the coefficients and residuals to highlight equivalence with the RSEM.

$$\begin{aligned}
\bar{X}_{t-1,i} &= \mu_{X_{t-1}} + \eta_{X,i} \\
\bar{Y}_{t-1,i} &= \mu_{Y_{t-1}} + \eta_{Y,i} \\
X_{t,i} &= \mu_{X_t} + \eta_{X,i} + \beta_{1X_t^*}\left(X_{t-1,i} - \bar{X}_{t-1,i}\right) + \beta_{2X_t^*}\left(Y_{t-1,i} - \bar{Y}_{t-1,i}\right) + \varepsilon_{X_t^*,i} \\
Y_{t,i} &= \mu_{Y_t} + \eta_{Y,i} + \beta_{1Y_t^*}\left(Y_{t-1,i} - \bar{Y}_{t-1,i}\right) + \beta_{2Y_t^*}\left(X_{t-1,i} - \bar{X}_{t-1,i}\right) + \varepsilon_{Y_t^*,i}
\end{aligned} \tag{S17}$$

Despite its somewhat different model specification, the FSEM parameters are fully equivalent to those of the RSEM. Because within-person residuals are correlated, the FSEM factorization involves the product of four bivariate conditional distributions.

$$\begin{gathered}
f(X_1, X_2, X_3, Y_1, Y_2, Y_3, \eta_X, \eta_Y) = \\
f(X_3, Y_3 | X_2, Y_2, X_1, Y_1, \eta_Y, \eta_M) \times f(X_2, Y_2 | X_1, Y_1, \eta_Y, \eta_M) \times f(X_1, Y_1 | \eta_Y, \eta_M) \times f(\eta_Y, \eta_M) = \\
f(X_3, Y_3 | X_2, Y_2, \eta_Y, \eta_M) \times f(X_2, Y_2 | X_1, Y_1, \eta_Y, \eta_M) \times f(X_1, Y_1 | \eta_Y, \eta_M) \times f(\eta_X, \eta_Y)
\end{gathered} \tag{S18}$$

The syntax excerpt below corresponds to the RI-CLPM from Figure S2. The three quantities needed to center each lagged variable—the raw scores, occasion-specific means, and random intercepts—are grouped together inside parentheses, forming composite functions that define the residualized predictors. During each MCMC iteration, the software substitutes the current values of these quantities into the within-equation function to estimate the regression model parameters. Attaching labels to the regression intercepts (e.g., @mu_x1, @mu_y1, etc.) provides access to the occasion-specific means.

```
# model 21a excerpt
LATENT: alpha eta;
MODEL:
alpha ~~ eta;
x1 ~ intercept@mux1 alpha@1;
x2 ~ intercept@mux2 alpha@1 (x1 - (mux1 + alpha)) (y1 - (muy1 + eta));
x3 ~ intercept@mux3 alpha@1 (x2 - (mux2 + alpha)) (y2 - (muy2 + eta));
y1 ~ intercept@muy1 eta@1;
y2 ~ intercept@muy2 eta@1 (y1 - (muy1 + eta)) (x1 - (mux1 + alpha));
y3 ~ intercept@muy3 eta@1 (y2 - (muy2 + eta)) (x2 - (mux2 + alpha));
x1 ~~ y1;
x2 ~~ y2;
x3 ~~ y3;
```

To simplify model specification, you can instead define the within-equation functions to define the residualized variables, as shown in the first four lines of the `MODEL` command as shown below.

```
# model 21b excerpt
LATENT: alpha eta;
MODEL:
# within-equation residual definitions
x1res = x1 - (mux1 + alpha);
y1res = y1 - (muy1 + eta);
x2res = x2 - (mux2 + alpha);
y2res = y2 - (muy2 + eta);
# ri-clpm equations
alpha ~~ eta;
x1 ~ intercept@mux1 alpha@1;
x2 ~ intercept@mux2 alpha@1 x1res y1res;
x3 ~ intercept@mux3 alpha@1 x2res y2res;
```

```
y1 ~ intercept@muy1 eta@1;
y2 ~ intercept@muy2 eta@1 y1res x1res;
y3 ~ intercept@muy3 eta@1 y2res x2res;
x1 ~~ y1;
x2 ~~ y2;
x3 ~~ y3;
```

Importantly, these definitions are not new variables created in the data set. Instead, they serve as shorthand that Blimp substitutes into the regression equations, making the specification more compact.

8. Dynamic Structural Equation Models (DSEM)

Multilevel models with lagged effects are central to a broader category of analyses known as dynamic structural equation models (DSEM; Asparouhov et al., 2018; Hamaker et al., 2021; Hamaker et al., 2018; McNeish & Hamaker, 2020). DSEM is an important innovation for the analysis of intensive repeated measures data. A prototypical a pair of repeated measures variables, X and Y. The level-1 models include autoregressive effects that connect X and Y scores at the current occasion to their lagged values from the previous occasion. They also feature cross-lagged effects linking the current value of X to the previous value of Y, and vice versa. Blimp can fit a variety of DSEM models with lagged effects, although it does so without decomposing lower-level outcome variables into within- and between-cluster latent variables. Instead, the "full" dependent variable always serves as the outcome. Decomposition is achieved by centering lower-level and lagged predictors at their latent group means.

To illustrate a basic DSEM, consider a two-level model featuring a pair of repeated measures variables, X and Y, with observation t nested within individual i. The level-1 models below include autoregressive effects, β_{1X} and β_{1Y}, which connect scores at occasion t to their previous values at occasion t – 1. They also feature cross-lagged effects, β_{2X} and β_{2Y}, linking the current value of X to the previous value of Y, and vice versa. For simplicity, these slopes are fixed, but they could be (and often are) random.

$$\begin{gathered}
X_{ti} = \alpha_i + \beta_{1X}\left(X_{t-1,i} - \alpha_i\right) + \beta_{2X}\left(Y_{t-1,i} - \eta_i\right) + \varepsilon_{X,ti} = \bar{X}_{ti} + \varepsilon_{X,ti} \\
Y_{ti} = \eta_i + \beta_{1Y}\left(Y_{t-1,i} - \eta_i\right) + \beta_{2Y}\left(X_{t-1,i} - \alpha_i\right) + \varepsilon_{Y,ti} = \bar{Y}_{ti} + \varepsilon_{Y,ti} \\
f(X_t, Y_t|\alpha, \eta) \to \mathcal{N}\left(\begin{pmatrix}\bar{X}_{ti} \\ \bar{Y}_{ti}\end{pmatrix}, \begin{pmatrix}\sigma^2_{\varepsilon_X} & \sigma_{\varepsilon_X\varepsilon_Y} \\ \sigma_{\varepsilon_Y\varepsilon_X} & \sigma^2_{\varepsilon_Y}\end{pmatrix}\right)
\end{gathered} \tag{S19}$$

An important feature of the previous expression is that the lagged predictors are centered at their latent group means (random intercepts), α_i and η_i, rather than their arithmetic means. This type of centering helps eliminate well-documented biases associated with multilevel lagged effects

(Asparouhov et al., 2018; Hamaker & Grasman, 2015; Zyphur et al., 2020). At level-2, the latent group means each have their own regression equation, and they share a bivariate normal conditional distribution.

$$
\begin{gathered}
\alpha_i = \beta_{0\alpha} + \varepsilon_{\alpha,i} = \bar{\alpha}_i + \varepsilon_{\alpha,i} \\
\eta_i = \beta_{1\eta} + \varepsilon_{\eta,i} = \bar{\eta}_i + \varepsilon_{\eta,i} \\
f(\alpha,\eta) \rightarrow \mathcal{N}\left(\begin{pmatrix}\bar{\alpha}_i \\ \bar{\eta}_i\end{pmatrix}, \begin{pmatrix}\sigma^2_{\varepsilon_\alpha} & \sigma_{\varepsilon_\alpha\varepsilon_\eta} \\ \sigma_{\varepsilon_\eta\varepsilon_\alpha} & \sigma^2_{\varepsilon_\eta}\end{pmatrix}\right)
\end{gathered}
\tag{S20}
$$

The Blimp syntax excerpt below shows the DSEM model specification. The model can include other multilevel modeling features such as random slopes and random within-cluster variation.

```
# model 22 excerpt
CLUSTERID: l2id;
TIMEID: time;
LATENT: l2id = alpha eta;
MODEL:
x ~ intercept@alpha (x.lag - alpha) (y.lag - eta);
y ~ intercept@eta (y.lag - eta) (x.lag - alpha);
x ~~ y;
intercept -> alpha eta;
alpha ~~ eta;
```

The `TIMEID` line identifies a level-1 index variable that encodes the temporal ordering of the observations (required for lagged effects). Next, the `LATENT` command defines level-2 latent variables that represent the latent cluster means (random intercepts). In the `MODEL` section, the first two lines are the level-1 regressions from Equation S24. The `@` symbols fix the intercept coefficients to the value of the latent group means. Appending `.lag` to a predictor's name creates a lagged variable that shift scores down by one row (time index) within each cluster. Within-cluster effects are created by subtracting the latent group means from the lagged predictors. This is another example of a function serving as a predictor (i.e., an equation within an equation). The within-cluster correlation is requested with a double tilde. Finally, the last two lines specify the multivariate distribution for the latent group means in Equation S25.

9. References

Aitchison, J., & Bennett, J. A. (1970). Polychotomous quantal response by maximum indicant. *Biometrika*, *57*(2), 253–262. https://doi.org/10.2307/2334834

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359-388.

Asparouhov, T., & Muthén, B. (2021). Expanding the Bayesian structural equation, multilevel and mixture models to logit, negative-binomial and nominal variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 622–637. https://doi.org/doi.org/10.1080/10705511.2021.1878896

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507.

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *British Journal of Mathematical and Statistical Psychology*, *Advance online publication*, 1–27. https://doi.org/10.1111/bmsp.12316

Blozis, S. A., McTernan, M., Harring, J. R., & Zheng, Q. (2020). Two-part mixed-effects location scale models. *Behavior Research Methods*, *52*, 1836–1847. https://doi.org/doi.org/10.3758/s13428-020-01359-7

Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Quartagno, M., & Kenward, M. G. (2023). *Multiple Imputation and its Application (2nd. Ed.)*. Wiley.

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software*, *45*(5), 1–14. https://doi.org/10.18637/jss.v045.i05

Cepeda, E., & Gamerman, D. (2000). Bayesian modeling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics*, 207–221.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, *6*(2), 101–111. https://doi.org/10.1007/Bf00162520

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *J Pers Assess*, *91*(2), 121–136. https://doi.org/10.1080/00223890802634175

Enders, C. K. (2022). *Applied Missing Data Analysis* (2nd ed.). Guilford Press.

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, *23*(2), 298–317. https://doi.org/10.1037/met0000148

Goldstein, H. (2005). Heteroscedasticity and complex variation. *Encyclopedia of statistics in behavioral science*, *2*, 790-795.

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, *9*(3), 173–197. https://doi.org/10.1177/1471082x0800900301

Hamaker, E., Asparouhov, T., & Muthén, B. (2021). Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling. In R. H. Hoyle (Ed.), *The Handbook of Structural Equation Modeling* (pp. 576–596). Guilford Press.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research*, *53*(6), 820–841. https://doi.org/doi.org/10.1080/00273171.2018.1446819

Hamaker, E. L., & Grasman, R. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*, 1492. https://doi.org/10.3389/fpsyg.2014.01492

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, 461–465.

Keller, B. T., & Enders, C. K. (2021). Blimp user’s guide (Version 3) [Computer Software]. https://www.appliedmissingdata.com/blimp

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

Mulder, J. D., & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 638–648. https://doi.org/10.1080/10705511.2020.1784738

Muthén, B., Asparouhov, T., & Shiffman, S. (2024). Dynamic structural equation modeling with floor effects. *Submitted for publication*.

Neelon, B. (2019). Bayesian zero-inflated negative binomial regression based on pólya—gamma mixtures. *Bayesian Analysis*, *14*(3), 829–855.

Neelon, B., & O'Malley, A. J. (2019). Two-part models for zero-modified count and semicontinuous data. In A. Levy, S. Goring, C. Gatsonis, B. Sobolev, E. v. Ginneken, & R. Busse (Eds.), *Health Services Evaluation* (pp. 695-716). Springer, New York, USA.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, *96*(454), 730–745.

Pillow, J., & Scott, J. (2012). Fully Bayesian inference for neural models with negative-binomial spiking. *Advances in neural information processing systems*, *25*, 1–9.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Quartagno, M., & Carpenter, J. R. (2019). Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical Journal*, *61*, 1003–1019. https://doi.org/10.1002/bimj.201800222

Smith, V. A., Neelon, B., Maciejewski, M. L., & Preisser, J. S. (2017). Two parts are better than one: modeling marginal means of semicontinuous data. *Health Services and Outcomes Research Methodology*, *17*, 198–218. https://doi.org/10.1007/s10742-017-0169-9

Windle, J., Polson, N. G., & Scott, J. G. (2014). Sampling Pólya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*.

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., & Koval, P. (2020). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods*, *23*(4), 688–716. https://doi.org/10.1177/1094428119847280