20

Measurement Models

Christopher J. Fariss, Michael R. Kenwick and Kevin Reuning

INTRODUCTION

Measurement models in general, and latent variable models in particular, are now common in political science research. This is because political scientists are increasingly focused on improving the measurement of unobservable concepts and understanding the relationships and potential biases between different pieces of observable information and the measurement procedures that link this information to theoretical concepts. Recent methodological and computational advances have led to a flourishing of new latent variable modeling applications. These new tools provide researchers with a means of measuring difficult to observe concepts based on events, ratings or other pieces of observable information that are assumed to be a result of the underlying unobservable latent trait.[1]

Latent variable models are built on the idea that observable variables are manifestations of an underlying conceptual process that is not perfectly observable or knowable and includes increasingly computationally sophisticated probability models (e.g., Imai et al., 2016; Jackman, 2000, 2001; Martin and Quinn, 2002; Plummer, 2017; Carpenter et al., 2017) and computationally simply additive scales (e.g., Guttman, 1949; van Schuur, 2003). In this chapter, we review the scientific measurement process and the assumptions needed to construct models of unobservable theoretical concepts.

The scientific process of measurement occurs in three iterative stages: *conceptualization* of the sociological or physical system being studied, *operationalization* of the data generating process that approximates the system and *empirical analysis* of the data generated by that system. The relationship between each of these steps is assessed using construct validity tools.[2] Because the measurement process is iterative, it is incumbent on the researcher to (1) acknowledge the starting point of the measurement process and (2) provide an assessment of the quality of the links between these steps. We provide more details about these recommendations throughout this chapter, although our focus here is on how latent variable models can be used to assess these steps.

Latent variable models allow for the empirical assessment of how the different observed pieces of data relate to one another through their association with the estimated latent trait. Even computationally simple additive scales are models that represent an underlying latent concept. Additive scales require the same process of assessment as more computationally difficult latent variable approaches (van Schuur, 2003). We discuss these additive scaling models as a starting point for thinking about estimating latent variable models more generally, because these models share the same set of assumptions. New computationally sophisticated latent variable models allow the researcher to relax these assumptions in conceptually meaningful ways.

The particular examples of latent variable models that we review in this chapter have been applied across a variety of subfields, encompassing the study of political ideology (Barbera, 2015; Bond and Messing, 2015; Martin and Quinn, 2002; Martin et al. 2005; Caughey and Warshaw, 2015; Kōnig et al., 2013; Pan and Xu, 2018; Treier and Hillygus, 2009; Windett et al., 2015), political attitudes, knowledge and preferences (Blaydes and Linzer, 2008; Pérez, 2011; Jessee, 2017; Stegmueller, 2011, 2013), regime institutions (Treier and Jackman, 2008; Pemstein et al., 2010; Kenwick, 2018, Gandhi and Sumner, 2019), UN voting positions (Voeten, 2000), human rights abuse (Schnakenberg and Fariss, 2014; Fariss, 2014, 2019; Fariss et al., 2020), human rights treaty embeddedness (Fariss, 2018b,a), judicial independence (Linzer and Staton, 2016), demographic variables (Anders et al., forthcoming), and institutional transparency (Hollyer et al., 2014). We discuss several latent variable models that are capable of accommodating different forms of conceptual dependencies between units, in particular temporal interdependence in time-series cross-sectional data. We provide examples that build on insights from a recently published article on temporal dependence and sudden temporal changes in time-series cross-sectional data (Reuning et al., 2019).[3]

After discussing the measurement process and construct validity in more detail and laying out different dynamics of latent variables, we highlight places that we believe are ripe for future research. In particular, we discuss new ways to theoretically include time in latent variable models, ways to scale expert surveys, the use of Multiple-Indicator-Multiple-Causes models and issues with different model fit statistics. Finally, we end with a list of recommendations for the applied researcher using latent variable models.

THE MEASUREMENT PROCESS

The process of measurement can be broadly characterized as having three steps.[4] The process of measurement allows the researcher to think explicitly about each of these three steps and the relationships between them because it links theories, the *concept*, with operational procedures, the *construct*, which generate observable information, the *data*. We discuss each of these steps here.

In the first step, a researcher generates a systematized definition of a concept in which they are interested. The systematized definition should be specific enough to have intellectual traction, but sufficiently broad so that it can be meaningfully applied to a set of objects across time, space or both (Shadish, 2010). What does this mean in practice? That there is necessarily a trade-off between specificity and generalizability and, when applied, the researcher must clarify the boundary conditions that define the set of objects for which the measurement procedure operates and the set for which it does not. At the extreme, the conceptual process should cover more than one object, but less than all objects. Specifying these boundary conditions is part of the conceptual step in the measurement process. However, because the measurement process is iterative, the researcher can and should return to this first step in order to make refinements to the systematized definition based on information obtained in the second or third step of the process.

Often in political science, even a well-defined concept cannot be directly observed in the real world. In the second step, the researcher must therefore begin to identify how the latent trait relates to observable information, thereby creating a data generating process from the latent trait to the observed indicators. A researcher interested in democracy might, e.g., identify whether a country holds competitive elections, whether there is a representative legislature with the ability to effectively pass legislation and whether there has been alternation in power among competing political groups. Thus, this second step involves the critical task of designing the data generating procedures used to collect information that relates to the underlying concept of interest for the objects under study.

Once the data generating procedures are defined, the researcher proceeds to the third step, which involves collecting observational information about a set of objects and the categorization or scoring of those objects. This process maps the observed information collected about the objects in the second step back to the concept of interest defined in the first step through a defined categorization or scoring procedure. The definitional rules of the operational procedure should be consistent with the conceptual definition defined in the first step. The creation and use of any operational protocol requires that researchers make decisions about how to weight each piece of information and how they individually or jointly inform the researcher's beliefs about an object's score for the underlying trait.

In sum, the three steps are: (1) define theoretical concept and scope; (2) identify how observational data connects to the theoretical concept by defining the data generating process; (3) use the operational procedure to categorize or score cases which are the subjects or units of study. Most of our discussion from here focuses on the second and third steps. This procedure highlights the fact that all measurement inherently involves the creation of a measurement model, which is the second step of the measurement process, but with links to both the first and third steps. Like all other models in social science, those used in measurement require careful validation about the relationships between steps.

At the broadest level, measurement validation centers upon what is known as construct validity, which is an assessment of both the theoretical content of the operationalization protocol and the empirical content that is believed to be captured by this construct (e.g., Adcock and Collier, 2001; Jackman, 2008; Shadish, 2010; Shadish et al., 2001). Construct validity encompasses a variety of different ways to evaluate a measure and operationalization.

Two important parts of construct validity are translation validity and measurement validity. Translation validity is an evaluation of the match between the theoretical construct and the proposed data generating procedure, which generates the observed pieces of information. Measurement validity is an evaluation of the fit between the proposed data generating procedure and the actual data obtained from it.

Translation errors occur when the operational protocol does not match the theory of the concept. Measurement errors occur when the fit between representation of the data generating procedure (the measurement model) and the data is poor. As researchers validate their measures along these two related criteria, they may choose to (1) update the types of information to collect, (2) modify the method for linking this information into scores on the latent trait or (3) modify the theoretical concept that the data generating procedure is derived from. The measurement process is an inherently iterative process between each of the three steps outlined above.

Thus, to generate good estimates of a theoretical concept of interest, the research must understand the relationship between each part of the measurement process.

MEASUREMENT MODELING ASSUMPTIONS

All measurement models, regardless of their complexity, require assumptions about the underlying trait. In this section we provide an overview of these assumptions for some of the measurement models that are most commonly used in the social sciences (additive scales and IRT models). We begin by discussing the assumptions of additive scales, proceed to identify assumptions of latent variable models and finally provide an overview of latent variable model assumptions about dynamics and their relationship to local independence.

Before proceeding, it is useful to provide a brief overview of the notation we will use in the following section. We denote the latent trait as θ, which is observed across units indexed with i, which takes on values of 1, 2, …, N, where N is the total number of units in the sample. We observe θ indirectly through observable pieces of information often referred to as 'items' or 'manifest indicators', each of which is indexed using k with values 1, 2, …, K, where K is the total number of manifest indicators. The realized values of these indicators are y, with Y acting as the manifest indicator yet to be observed. This notation lets us refer empirically to both the potential observed realization of data Y and the actual realization of data y. Formally, we let Y_{ik} denote the score of subject i on item k, a random variable with realization $y_{ik} = \{0, 1\}$. For simplicity, we assume that each indicator is binary. In the next section, we will continue to build towards an additive scale as a latent variable representation of a concept. We also discuss the assumptions underlying this model and the standard unidimensional item response theory models which we review later in the chapter.

Assumptions of Additive Scale Measurement Models

To make the notation and formalizations presented in this section more clear, we introduce a small deterministic example that illustrates the relationships between the different model parameters and data. As we mentioned above, we let k take on integer values from 1, 2, 3, which represents three distinct questions of varying ability that we will ask of five hypothetical subjects. These are the items which generate responses (i.e., the item responses) from each subject. We first introduce a new parameter α_k which represents a feature of the items. In a testing setting, α_k parameters represent the difficulty of a particular question as it relates to the ability of the test-takers or subjects, which is represented by θ. In additive scales it is assumed that if the latent trait for unit i is greater than α_k then we will observe $y_i = 1$. More generally, α_k accounts for the variation in how high (or low) a unit has to be on the latent trait to achieve a positive outcome for indicator y_k. For this example, we are supposing that we know the true values of this parameter in our measurement model. Later on, we will estimate these parameters.

In our example we consider the following latent traits for five units ($\theta_1 = -2$, $\theta_2 = -1$, $\theta_3 = 0$, $\theta_4 = 1$, $\theta_5 = 2$) and three items ($\alpha_1 = -1.5$, $\alpha_2 = 0$, $\alpha_3 = 1.5$), which are all arrayed along the same unidimensional line. The relationship between the five units and the three items are displayed visually in Figure 20.1. The unidimensional line represents values of the unobservable theoretical concept of interest but the substantive meaning of the entities along the line differ because some are subjects and the others are the data generating objects (i.e., the item or test questions).

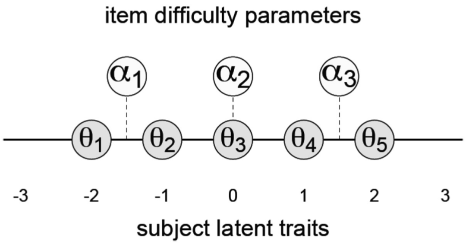

Figure 20.1 Latent variables and item parameters

Note: This plot displays latent traits for 5 units ($\theta_1 = -2$, $\theta_2 = -1, \theta_3 = 0, \theta_4 = 1, \theta_5 = 2$) and 3 items ($\alpha_1 = -1.5, \alpha_2 = 0$ $\alpha_3 = 1.5$) all arrayed along the same unidimensional line. The unidimensional line represents values of the unobservable theoretical concept of interest but the substantive meaning of the entities along the line differ because some are subjects and others are the data generating indicators (i.e., the item responses generated by the subjects). The subjects and items are comparable in this space however. In particular, the comparison of the distance between subject and object determines the observed binary item responses for each subject–object pairing.

The relationships displayed visually in Figure 20.1 are unobserved. What we actually observed are binary responses (e.g., the answers to questions generated by subjects or the categorical values created to compare country-year units). Our measurement goal is to create a test or categorization scheme that relates the observed data back to the unobserved latent traits. This is done by assuming a data generating process from the latent trait to the indicators. Here we will use a deterministic function for the relationship between each subject–item pairing, which is displayed in Equation 1. Later on we will introduce a probability model for accomplishing this task.

$$y_{ik} = \begin{cases} 1 & \text{if } \theta_i > \alpha_k \\ 0 & \text{if } \theta_i \le \alpha_k \end{cases} \quad (1)$$

Equation 1 represents the data generating function for the binary item responses produced for each subject–item pair. For the illustrative example,

$$y_i^+ = \sum_k^K (y_{ik}) \quad (2)$$

Equation 2 represents the observed additive scale value for each subject i, which is determined by the value of the logical proposition in equation 1. Table 20.1 presents the additive scale values for y_i^+ based on the pairwise comparisons between the five subjects and the three items. The additive scale is a deterministic, continuous scale, which satisfies the conditions outlined by Guttman (e.g., Guttman, 1949; van Schuur, 2003). In words, the first subject's ability is always less than the value of the item. To reiterate, the values are substantively distinct but are comparable together on the same latent scale.

The additive scale can also be rewritten as a function of just the values of the latent trait and the difficulties. This is the function in Equation 3, where the additive value is found by checking the latent trait's value against the ordered alphas. This emphasizes that in additive scales there is an assumption that all items can be ordered in such a way that

Table 20.1 Example of additive scale function

Latent Trait	*Items*			*Additive Scale*
θ_i	$\alpha_1 = -1.5$	$\alpha_2 = 0$	$\alpha_3 = 1.5$	y_i^+
$\theta_1 = -2$	$\theta_1 \le \alpha_1 \Rightarrow +0$	$\theta_1 \le \alpha_2 \Rightarrow +0$	$\theta_1 \le \alpha_2 \Rightarrow +0$	$y_1^+ = 0$
$\theta_2 = -1$	$\theta_2 > \alpha_1 \Rightarrow +1$	$\theta_2 \le \alpha_2 \Rightarrow +0$	$\theta_2 \le \alpha_3 \Rightarrow +0$	$y_2^+ = 1$
$\theta_3 = -0$	$\theta_3 > \alpha_1 \Rightarrow +1$	$\theta_3 \le \alpha_2 \Rightarrow +0$	$\theta_3 \le \alpha_3 \Rightarrow +0$	$y_3^+ = 1$
$\theta_4 = 1$	$\theta_4 > \alpha_1 \Rightarrow +1$	$\theta_4 > \alpha_2 \Rightarrow +1$	$\theta_4 \le \alpha_3 \Rightarrow +0$	$y_4^+ = 2$
$\theta_5 = 2$	$\theta_5 > \alpha_1 \Rightarrow +1$	$\theta_5 > \alpha_2 \Rightarrow +1$	$\theta_5 > \alpha_3 \Rightarrow +1$	$y_5^+ = 3$

Note: The additive scale values are based on the status of the logical propositions for each subject-item comparison.

they are monotonically increasing in difficulty. Once ordered, a researcher can identify where a unit is on the additive scale based on when its indicators switch from 1 to 0.

$$y_i^+ = \begin{cases} 3 & \text{if } \theta_i > \alpha_3 \\ 2 & \text{if } \theta_i > \alpha_2 \quad \text{and} \quad \theta_i \le \alpha_3 \\ 1 & \text{if } \theta_i > \alpha_1 \quad \text{and} \quad \theta_i \le \alpha_2 \\ 0 & \text{if } \theta_i \le \alpha_1 \end{cases} \tag{3}$$

We can visually represent the relationship between the values of the additive scale, the latent trait, and the items in Equation 3. We do this in Figure 20.2.

Up until now, we have assumed a deterministic model between the observed items and the latent trait, which are consistent with the assumptions from Guttman (1949). In later measurement research, Mokken (1971) developed a stochastic version under the assumptions of a unidimensional latent variable, latent monotonicity and local independence. Under these assumptions, the

Figure 20.2 Example of additive scale function

Note: This plot displays latent traits for 5 units ($\theta_1 = -2$, $\theta_2 = -1, \theta_3 = 0, \theta_4 = 1, \theta_5 = 2$) and 3 items ($\alpha_1 = 1.5, \alpha_2 = 0$, $\alpha_3 = 1.5$) all arrayed along the same unidimensional line displayed in Figure 20.1. The additive scale values on the y-axis are based on the status of the logical propositions for each subject-item comparison in Table 20.1.

proportion of 'correct' answers by subject i to item k is nondecreasing in the sum of all the items. These assumptions also imply that all of the items are positively correlated across all subsets of subjects (Mokken, 1971). Under these assumptions the unweighted sum of the variables increase as θ increases. Mokken Scaling Analysis (MSA) is simply a stochastic version of a Guttman scale, in which items measure a single latent construct and can be ordered by difficulty (Guttman, 1949) but are not assumed to be generated without error (van Schuur, 2003).

The assumptions made by Mokken (1971) are common across many latent variable models and so are worth exploring in more depth. The first assumption is that θ is a *unidimensional latent variable*, which means that the values of the latent trait reside on a single axis. This assumption can be tested using parameters from the Mokken Scaling Analysis (MSA) model (van Schuur, 2003). If this assumption fails, it means that the latent trait cannot be collapsed into a single dimension but that units can be high in one dimension and low on another.

The second assumption is of *latent monotonicity*, which means that the item step response function is strictly increasing on $\theta; \theta_1 \le \theta_2 \Rightarrow P(Y_{ik} \ge y_{ik} \mid \theta_1) \le P(Y_{ik} \ge y_{ik} \mid \theta_2)$. This implies that as a unit increases in the latent variable, the probability of observing a positive indicator also increases.

The third assumption is of *local independence*, which means that the item responses are not deterministically related to each other outside of their relationship to the latent trait. This implies that the probability of the set of each subject's item responses is $P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \cdots Y_{iK} = y_{iK} \mid \theta_i)$ $= \prod_{k=1}^{K} P(Y_{ik} = y_{ik} \mid \theta_i)$(van Schuur, 2003). The only relationship between items is through their relationship with the latent variable. This can be violated in the testing environment when getting one answer correct depends on getting previous answers correct.

To summarize, additive scaling is a data generating procedure that maps the latent trait to an additive index. In order to estimate a stochastic additive scale, researchers must make assumptions about unidimensionality, monotonicity and local independence. As we discuss next, these assumptions are also present in more complicated latent variable models which also allow more variation in how the latent trait relates to the observed indicators.

Identification Assumptions of Latent Variable Models

We now move to estimate θ itself because, up until this point, this parameter has been entirely conceptual. We do this through the Item Response Theory (IRT) framework which allows us to estimate θ as well as other parameters in the data generating process. In addition, using this framework we can add an additional layer of complexity of cross-sectional time-series data (i.e., country-year units) instead of the five hypothetical subjects from before.

In principal, IRT models are rooted in the same assumptions as the additive scale above; that is, we assume that θ is a *unidimensional latent variable* and that its relationship with its associated items is characterized by *latent monotonicity* and *local independence*.

Under the IRT framework, the latent trait is θ_i where the subscript $i = 1, \ldots, N$ indicates multiple units. y_{ik} is the observed value for item k for unit i. For each item α_k and β_k are also estimated. α_k continues to act as a 'difficulty' parameter, or a threshold that benchmarks how likely an indicator is to be observed relative to the values of the latent trait. In our formulation, this is analogous to an intercept in a traditional logistic regression model. β_k is often referred to as the 'discrimination' parameter and is the analogue of a slope coefficient.

The relationship between θ_i and our indicator y_{ik} is:

$$P(y_{ik} = 1) = \Lambda(\alpha_k - \beta_k \theta_i) \tag{4}$$

where Λ is the logistic function. Unlike in the case of the additive scale, this is necessarily probabilistic.[5] The likelihood function encompassing the latent trait, realizations of the manifest indicators and item-specific parameters take the following form:

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{k=1}^{K} \Lambda(\alpha_k - \beta_k \theta_i)^{y_{ik}} \left(1 - \Lambda(\alpha_k - \beta_k \theta_i)\right)^{1-y_{ik}}$$

The model estimates the placement of one unit relative to all the other units based on the values of the observed items. Without additional information such models are not identified, which means that estimation is not possible because multiple sets of values for the parameter estimates will fit the data equally well. There are generally three types of identification problem that most applied researchers will encounter: additive, scale and rotational. In each of these cases the likelihood is invariant across multiple parameter estimates. To prevent this situation, the researcher must make several benign assumptions that provide additional information to the model and prevent invariance.

The issues of scale and additive invariance are often the easiest to solve. In the case of additive invariance, $\theta + \delta$ and $\alpha - \delta$ lead to equivalent likelihood for any δ. Scale invariance is similar except is a result of multiplication: $\delta \cdot \theta$ and $\frac{\theta}{\delta}$ would again produce equivalent likelihoods. This invariance is commonly solved by providing information to θ through a standard normal distribution as the prior. This is useful as it leads to estimates of θ that are mean 0 with a standard deviation of 1.

Rotational invariance can be more complicated. Rotational invariance is the result of equivalent likelihoods that result when θ is multiplied by –1 or 'flipped'. In the context of a latent variable for ideology, estimates with negative numbers as conservative and positive numbers as liberal are the same as when negative numbers or liberal and positive numbers are conservative. Put differently, the model has no way of knowing whether to order the units from liberal to conservative, or from conservative to liberal ideologies.[6]

One simple strategy for resolving rotational invariance is to fix the values of the latent trait for two or more units. In the political ideology example, this could be achieved by assigning values of the latent trait for a very liberal and a very conservative individual. An alternative strategy imposes assumptions about the relationship between the manifest indicators and the latent trait through the discrimination parameters, β_k. For example, Fariss (2014) relies on a series of indicators believed to positively correlate with respect for human rights, and therefore restricts the β parameters to take on positive values. In practice, this can be done through the use of truncated distributions (e.g., half-normal) or strictly positive distributions (e.g., gamma).

As a demonstration of issues of invariance, consider the simple single dimensional model for five units. We plot these five units along a single dimension in Figure 20.3. The first row shows the baseline, placing all five units in order. The second row shows

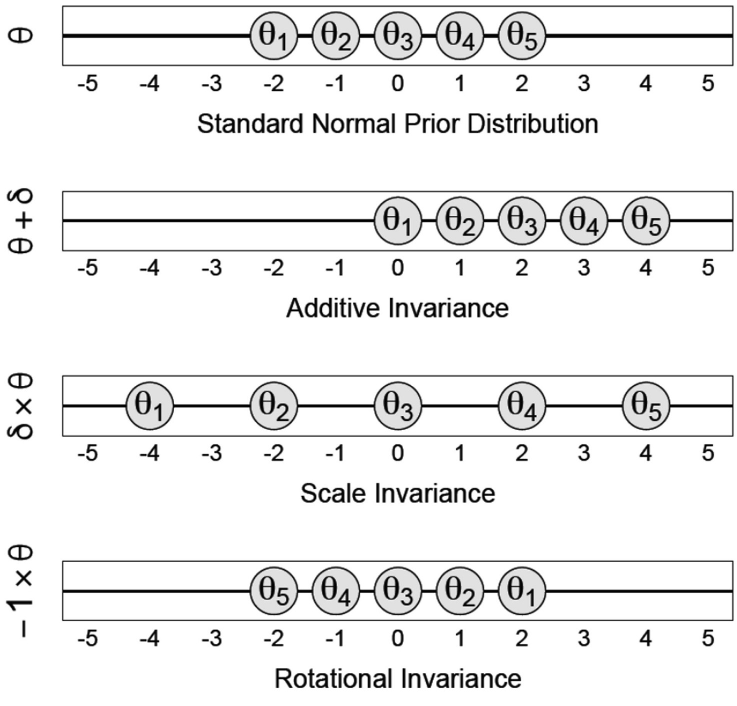

Figure 20.3 Identification issues in latent variables

Note: This plot displays latent traits from four idealized models. The top row displays the 5 units scaled so that the mean value is 0. The other rows show the consequence of the values of the latent trait when adding a constant (row 2), multiplying a constant (row 3), and multiplying by –1 (row 4). These models each provide the same values for comparisons of the value of one unit relative to any other or to the mean value of all of the units. Since we do not know the true absolute value of the concept we wish to make inferences about, it is useful to constrain the values of the latent trait to occupy the standard normal density function. By constraining the model in this way, we ensure that we are not mixing and therefore comparing values from the other models represented in this visualization.

a rightward shift of all five units (additive invariance). Since the latent dimension is arbitrary, this move does not matter as long as all units move in a similar way and there are no assumptions made about where the center of the latent space is.

In row 3 we demonstrate the issue of scale invariance. Here, the latent trait has been multiplied by 2, expanding the latent scale. Again, because each unit moves equally the end result is no different from the initial placement in row 1 if there is no constraint placed on the scale of the latent trait. Finally, row 4 shows rotational invariance. The latent traits have been reversed so that θ_1 moves from 2 to −2. This is equivalent to the first row if there is no constraint placed on the direction of the scale.

In our running examples, we place normal priors on the latent trait and resolve the issues of location and scale invariance.[7] To resolve rotational invariance, we constrain β_k to be greater than zero, such that increasing values of each manifest indicator are associated with increasing values of the latent trait. Finally, we place weakly informative normal priors on the difficulty parameters. The prior assignments can therefore be expressed as:

$$\theta_{it} \sim \mathrm{N}(0,1) \quad \forall i = 1,\ldots,N$$
$$\beta_k \sim \mathrm{HN}(0,3)$$
$$\alpha_k \sim \mathrm{N}(0,3)$$

where HN is the half-normal distribution, with support on [0, ∞).

Local Independence and Assumptions about Dynamics

The model described above can be expanded to include units over multiple time periods. In the above equations, this is accommodated by replacing θ_i with θ_{it} where t indexes time periods from 1, …, T. There is no requirement that all units must be observed over all time periods.

This does lead to some methodological questions. Latent variable models, including simple additive and cumulative scales, are built on the assumption that each observed variable for a unit is generated independently of the other observed pieces of information about that unit. This is the assumption of *local independence*. For the type of cross-sectional time-series data that we consider in this chapter, the assumption of local independence means that any two observed variables are *only* related because of the fact that they are each an observable outcome of the same latent variable.

There are three relevant local independence assumptions: (1) local independence of different indicators within the same country-year; (2) local independence of indicators across countries within years; and (3) local independence of indicators across years within countries. Priors are a useful and common means of addressing potential violations of the lattermost type of local independence violations. Applied researchers in international relations are likely to encounter problems where they are attempting to estimate a measure of multiple units observed over time. The dependencies within a unit across time can be modeled as part of the prior on the latent variable. In this section we discuss three broad approaches in the field. Two of these are relatively common, while the last has been recently introduced. In each case we discuss the assumptions that the model makes, the benefits of it and the costs.

Static model

The three modeling strategies we present are differentiated by the prior information assigned to the latent variable. We start here with the simplest model, the static model. The static model places a standard normal prior on all units for all time periods:

Static model prior

$$\theta_{it} \sim \mathrm{N}(0,1) \quad \forall i = 1,\ldots,N \qquad \forall t = 1,\ldots,T$$

The standard normal prior, as discussed above, prevents additive and scale invariance. Estimates for the latent trait for each unit in each time period are differentiated exclusively by the values of the indicators for that unit at that time period. This model treats each unit-time period as independent, which is a bold assumption to make in most applied research. In addition, this limits the information that is being used to estimate the latent trait and so is likely to increase credible intervals. In the case where the indicator variables contain sufficient information on the latent trait, this modeling strategy may not be problematic. Unfortunately, this is seldom the case when using social science data, where indicators are often coarse or missing. As a result, these indicators often do not contain sufficient information to differentiate between theoretically distinct units. The benefit to this approach is that it does not force any atheoretical 'memory' on the latent trait allowing sudden changes in the latent trait across time-periods.

Standard dynamic model

To address temporal non-independence in the data, many researchers have used a dynamic prior for the latent trait, where the latent trait for unit i in time t is related directly to the latent trait for unit i at time $t-1$ (Martin and Quinn, 2002; Schnakenberg and Fariss, 2014; Fariss, 2014; Caughey and Warshaw, 2015; Kōnig et al., 2013). The choice of a 'random walk' prior on the latent variable is particularly common.

The random walk approach begins with the use of a standard normal prior on the latent trait in the first observation period for every unit. Then for each subsequent time period, the prior is normally distributed with mean $\theta_{i(t-1)}$, and a standard deviation σ which is either assigned by the researcher or, more commonly, estimated from the data.[8] Here, we assign a weakly informative prior to σ by using a half-normal distribution with standard deviation of 3 and mean 0.

Standard dynamic model priors

$$\theta_{i1} \sim \text{N}(0,1) \quad \forall i = 1,\ldots,N$$
$$\theta_{it} \sim \text{N}(\theta_{i(t-1)},\sigma) \quad \forall i = 1,\ldots,N \quad \forall t = 2,\ldots,T$$
$$\sigma \sim \text{HN}(0,3)$$

This strategy trades the assumption that observations are independent with the assumption that the latent trait will be correlated over time and will follow a random walk. As a result, estimates from dynamic models typically have less uncertainty because more information is used to estimate each latent variable. This also induces smoothing over time because changes between time periods are constrained. When researchers have theoretical reasons to expect that the latent trait is relatively slow-moving over time, both modeling features can be desirable. If, however, the latent trait is subject to rapid fluctuations or state changes between time periods, this temporal smoothing can produce biased estimates. The modeling strategy we introduce below is designed to address this problem while still accounting for temporal dynamics.

Robust dynamic modeling

We recently proposed an alternative strategy that drew on the robust modeling literature to implement a robust version of the dynamic modeling (Reuning et al., 2019). In the Bayesian framework, robust models alternate normal distributions with the Student's t-distribution to account for outliers (Gelman et al., 2014; Lange and Sinsheimer, 1993; Lange et al., 1989; Geweke, 1993; Fonseca et al., 2008). In the context of dynamic latent variables, potential outliers are the 'shocks' where values of the true latent variable change suddenly within a unit's time series.

The robust dynamic model continues to use a standard normal distribution for the first observation in a unit's time series.[9] In subsequent years, the prior follows a Student's t-distribution with four degrees of freedom. Setting the degrees of freedom to a relatively low value increases the density of the tails of the distribution, which allows 'extreme values' to be estimated from time period to time period. Thus, the model smooths estimates across time during periods of stability, but also allows for rapid changes in the latent trait during periods of volatility. It is possible to estimate the degrees of freedom, but this can lead to identification problems, which we explore in more detail in the appendix to Reuning et al. (2019). Setting a low degree of freedom of 4 has been recommended in other contexts (Gelman et al., 2014) and so we believe that it will be useful in most latent variable cases.

Robust dynamic model priors

$$\theta_{i1} \sim \mathrm{N}(0,1) \quad \forall i = 1,\ldots,N$$
$$\theta_{it} \sim \mathrm{T}_4(\theta_{i(t-1)},\sigma) \quad \forall i = 1,\ldots,N \quad \forall t = 2,\ldots,T$$
$$\sigma \sim \mathrm{HN}(0,3)$$

EXTENSIONS OF LATENT VARIABLE MODELS AND SUGGESTIONS FOR FUTURE RESEARCH

In this final section, we highlight different fruitful paths for research using latent variable models. We discuss new ways that theory has informed particular modeling strategies and how this can provide new insights. We then present Multi-Rater/Aldrich-McKelvey Scaling models, which allow researchers to use latent variable models to reduce the impact of rater preferences when trying to develop uniform scales from expert surveys. We go on to introduce Multiple-Indicator Multiple-Causes models. These models are relatively common in psychology but are rarely used in published political science research, even though they provide a principled way to test what drives change in a latent variable. We then discuss problems with different model fit statistics. We close with a set of best practices useful for guiding future research.

The Seriousness with Which One Must Take Time

The modeling structures outlined above identify only a few ways in which researchers may care to model temporal dynamics. In practice, researchers are beginning to identify a variety of new strategies to address different forms of temporal non-independence. At times, for example, researchers have reason to suspect that the relationship between a manifest indicator and the latent trait may change over time. Kenwick (2018), for example, is interested in civilian control of regime institutions and argues that the strength of this control increases over time, with civilian control expected to be higher in a state where civilians have ruled for several decades than in one that had previously experienced a military takeover. He therefore structures the prior distribution on the latent trait for civilian regimes as a random walk with drift, allowing the values of the latent trait to systematically increase (or decrease) over time. Fariss (2014) faces a different type of temporal non-independence in the study of human rights violations, and argues that the standards with which human rights reports are written has changed over time. To accommodate these potential biases, Fariss (2018b) allows the item discrimination parameters linking standards based indicators to latent trait to vary over time to mitigate temporal biases.

In each case, the specific modeling structure used to generate estimates of the latent trait was informed by prior theory and the results are empirically validated against competing models. These examples demonstrate how the choice of modeling structure can fundamentally alter the estimates of the latent trait itself, and the theoretical inferences one draws from the measurement analysis. These insights are often nontrivial and must be treated with the same care with which other forms of hypothesis testing are conducted. Nevertheless, these examples demonstrate how the proliferation of dynamic variable modeling techniques offers fertile new testing grounds for the theoretical evaluation of concepts of interest.

Models of Other Unit Dependences: Multi-Rater/ Aldrich–McKelvey Scaling

Latent variable approaches can also be useful in the context of expert and non-expert survey when there is concern over how individuals will respond to survey items. This question was first approached in research on surveys of voters in the United States (Aldrich and McKelvey, 1977; Hare et al., 2015), but has also recently been used in the context of expert surveys to quantify country level attributes (Marquardt and Pemstein, 2018). The benefits of these approaches, which we will refer to here as multi-rater IRT, is that in using them, researchers can place answers from survey participants that might view underlying concepts on different scales onto a single scale.

As an example, take the work of Marquardt and Pemstein (2018), in which the authors use a multi-rater model to place expert surveys about democratic practices within a country on a single scale. They start with a survey of experts, asking them to rate several countries on a variety of democratic attributes. The problem with using these ratings directly is that different experts might have different opinions about how democratic a country must be to be considered the most democratic, and may also vary in their general understanding of the question. This is a form of *differential item functioning* where the relationship between an item (a response to a particular survey question) and the latent trait varies.

To account for differential item functioning the β (discrimination) and α (difficulty) parameters are estimated for each survey participant but held constant across the countries that they rated. For example, if Y_{ic} is expert i's response to a question on country c then it would be estimated as a function of $\alpha_i + \beta_i\theta_c$.

This technique is fruitful not only in the context of expert surveys but also for non-expert surveys where there are varying perceptions. Hare et al. (2015) use this to identify ideological placement of US senators from a survey of voters. The multi-rater method accounts for the fact that more liberal voters are likely to see the same senator as being more conservative than a moderate voter.

Nevertheless, in order for measures to be made comparable, there has to be a degree of overlap in the units that survey participants rate. This returns to the problem of bridging discussed above. Without overlap, the latent estimates will not be comparable across units. Overlap allows us to identify the degree of differential item functioning and so provide estimates of latent variables that are comparable when there is significant differential item functioning.

Adding Even More Structure: MIMIC Models

The final extension we consider is less focused on particular latent models and more on the use of estimates from the latent models. Latent models produce estimates of the latent traits that include error. The error needs to be a part of any future models that use the latent variable. When the latent variable estimates are used as an independent variable, estimation that incorporates error can be achieved relatively easily. All that is necessary is to take *N* draws from the posterior of the latent variable, estimate N models that use the latent variable as an IV and then combine those estimates using the same process that is used to combine multiple imputations.[10]

Estimating models where the latent variable is the dependent variable requires more care, but there are methods that are commonly used outside of political science that can accomplish this goal. Multiple-Indicator Multiple-Causes (MIMIC) models were developed starting in the 1970s to allow researchers to use multiple measures of a trait when estimating the impacts of exogenous variables on that trait (Jöreskog and Goldberger, 1975; Muthén, 1989). The MIMIC model approach is commonly employed in psychology (Krishnakumar and Nagar, 2008) and was more recently introduced to political science in the context of political psychology (Pérez, 2011).

In brief, MIMIC models include covariates for the latent variable that is being estimated. These covariates are included in the initial estimation process and so capture the error that is inherent in measuring a latent variable. Covariates are included by modeling θ directly as a function of the covariates instead of just setting a simple prior on it.[11] In addition to providing better estimates of the covariates on the underlying latent trait, MIMIC models can be modified to identify differential item functioning that is correlated with one of the covariates (Pérez, 2011).

One caveat for MIMIC models is that we are unaware of anyone who has connected the MIMIC approach to the dynamic latent variable approaches discussed here. Both approaches involve modifying the modeling of the latent variable (either through an informative prior or a regression setup) and so connecting the two will require additional work.

Assessing Model Fit: WAIC for Hierarchical and IRT Models

WAIC (the Watanabe–Akaike or widely applicable information criterion) is currently one of the more preferred model diagnostics for Bayesian models (e.g., Gelman et al., 2014). However, several open research questions remain under-explored when using WAIC with hierarchical or IRT models.

WAIC is an approximation of leave-one-out validation, but approximating leave-one-out validation leads to a problem in IRT data over what ought to be 'left out' when validating models. That is, should individual items be left out for all unit-time periods, for units from a panel or for all unit-years? Or should all the items be left out for one of these unit structures? Newly published research extends WAIC to cases in which items are clustered within an observation (Furr, 2017) as well as other work incorporating time dynamics (Li et al., 2016). Another recent area of work is diagnostics, and best practices for WAIC and other models (Vehtari et al., 2017).

When there is concern over the validity of WAIC statistics, it is useful to also estimate a K-fold cross validation. This of course also requires removing a set of data and estimating the model. We suggest that researchers randomly sample indicators to remove so that each unit-time is still in the model. This allows estimates of latent traits for each unit-time and those estimates can be used to calculate a held-out log-likelihood.

We suggest that while this area of research continues, researchers should provide multiple checks of model fit. Posterior predictive checks are another very powerful way to test how well an IRT model fits data (Gelman and Hill, 2007). Overall, fit statistics, posterior predictive checks and visual analysis of the temporal patterns of well-known cases allow for the evaluation of competing models without relying on a single statistical tool.

Best Practices for Applied Measurement Research

Finally, as researchers use these methodologies, we propose a few useful suggestions on how to best approach modeling latent variables. It is our intention that these suggestions are consistent with the statistical modeling choices made when selecting the component parts of latent variable models, and that these choices will be made with reference to the two main types of construct validity also discussed. Recall that the process of measurement occurs in three iterative stages: *conceptualization* of the sociological or physical system being studied; *operationalization* of the data generating process that approximates the system; and *empirical analysis* of the data. The specific terms we use for each of these three stages is *concept*, *construct*, *data*. Construct validity is an overarching term for assessing the relationship between one or more of the entities represented in each of these steps.[12]

- **Validate by letting the theoretical concept drive the measurement specification:** We have referred to this type of validation as translation validity and it is concerned with the link between the theoretical concept and the operationalized construct. It is not possible to consider a measure of an unobserved concept without referencing a theoretical concept. For a construct to be valid, it needs to translate the theoretical concept into an operational procedure that will generate data consistent with the theory. Thus, the first step for any research on latent variables is to outline the assumed relationships between the data generating process and the concept to be measured. Will the data generating process produce indicators that reflect the underlying concept of interest? Are the proposed items manifest of the underlying concept? Are the proposed items substitutes for each other? How are proposed items measured over time?
- **Validate by assessing the assumptions of the measurement model as they relate to theoretical concept of interest.** This is also a suggestion about translation validity. How does the specification of the measurement model translate the theoretical concept into the operational procedure that generates the observed data? Every measurement model has underlying assumptions and it is important that any empirical patterns are the result of the underlying data and not of the assumptions. In the case of latent measurement models, researchers must pay close attention to any parameters that are set without reference to theory of their concept of interest.
- **Validate the fit of the measurement model as it relates to the observed data.** How does the model of the data generating process, the latent variable, fit the observed data? This is an assessment of measurement validity. Measurement validity is an evaluation of the fit between the proposed data generating procedure and the actual data obtained from it. WAIC (the Watanabe–Akaike or widely applicable information criterion) and other statistical tools are useful ways to test model fit, but researchers should not just select a model based on a single statistical tool. One useful way to test competing models is to focus on divergent estimates and use *a priori* knowledge about the world to validate which one is the best.

There is no guarantee that any single modeling strategy will be equally well-suited for use with all data types or for estimating all types of latent concepts. The assumptions of the measurement model will influence the conclusions researchers draw about the underlying theoretical concept of interest, as well as the empirical linkages between these concepts and other political phenomena.

CONCLUSION

The assessment of theories about political institutions and behaviors often requires measuring concepts that are not directly observable. Thus, for science to proceed, measurement is essential, because without a clearly articulated link between the empirical content of a study and the theoretical structure that gives rise to that content, it is not possible to make claims about the relationship between data and the world. Yet, despite the necessity for valid measurement, research in the social sciences still often tends to ignore the construct validity of most measures and usually takes existing data, especially experimental data, for granted or at least as good enough. Thus, one of the critical steps in evaluating theoretical concepts is the development, formalization and validation of measurement models. This is because there is no model-free way to measure unobservable or difficult to observe concepts. And many of the concepts of interest to the political science community are often by definition difficult to observe. As we have discussed in this chapter, construct validity – and measurement models in general, and latent variable models in particular – are tools which are useful for systematically evaluating the relationship between concepts, operational procedures (e.g., the data generating process) and data.

Notes

1 For the purposes of this chapter, we focus exclusively on unidimensional measurement models that are explicitly created in an effort to link observed data to an unobservable concept.
2 The development of the concept of construct validity has occurred over many decades. Primary contributors include: Campbell and Fiske (1959); Campbell (1960); Campbell and Ross (1968); Cook and Campbell (1979); Shadish (2010); Shadish et al. (2001). However, the conceptual meaning of the terms used in these article have evolved over time. As Jackman (2008) notes, 'there are several species of measurement validity. But at least in the context of latent variables, the term "construct validity" has lost much of the specificity it once had, and today is an umbrella term of sorts' (122). We use the term construct validity in this way and point out specific subtypes where appropriate. We note further that different fields and subfields use the various construct validity terms in different ways, which has led to some confusion when translating across terms. Adcock and Collier (2001) review this issue in brief, but like them, we leave a full accounting for the agreement and disagreement of overlapping validity concepts to future work.
3 Reuning et al. (2018) provide a complete and detailed set of replication files that demonstrate how to use these particular latent variable models using both applied examples and a set of simulation-based models: https://doi.org/10.7910/DVN/SSLCFF.
4 We build on ideas covered in Adcock and Collier (2001) and elsewhere (e.g., Jackman, 2008; Shadish, 2010; Shadish et al., 2001).
5 The additive scale can be seen as a result of rewriting this to $\beta_k\ (\theta_i - \alpha_k)$ and fixing $\beta = \infty$. This creates the step function that can be seen in Figure 20.2.
6 As the number of dimensions for the latent variable increases there is an increasing number of invariant rotations. For one dimension there are only two equivalent estimates; with two dimensions that number increases to eight (e.g., Jackman, 2001).
7 In the following section we will continue to leverage the normal prior for identification constraints, but we will introduce modifications to accommodate temporal dynamics.
8 The σ parameter is sometimes referred to as the innovation parameter.
9 In practice, one can also substitute a Student's t-distribution with a very high degree of freedom (e.g., 1,000), which closely approximates the normal distribution.
10 Mislevy (1991), Bolck et al. (2004) and Schnakenberg and Fariss (2014) each provide arguments and detailed suggestions on how to incorporate the uncertainty from latent variable estimate using the multiple imputation equation formula from Rubin (1987).
11 For more detailed discussion of estimations of MIMIC models see Fahrmeir and Raach (2007).
12 Two important parts of construct validity are translation validity and measurement validity. Translation validity is an evaluation of the match between the theoretical construct and the proposed data generating procedure which generates the observed pieces of information. Measurement validity is an evaluation of the fit between the proposed data generating procedure and the actual data obtained from it (Fariss and Dancy, 2017).

REFERENCES

Adcock, Robert and David Collier. 2001. 'Measurement Validity: A Shared Standard for Qualitative and Quantitative Research.' *American Political Science Review* 95(3):529–546.

Aldrich, John H. and Richard D. McKelvey. 1977. 'A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections.' *American Political Science Review* 71(1):111–130.

Anders, Therese, Christopher J. Fariss, and Jonathan Markowitz. Forthcoming. 'Bread before guns or butter: Introducing Surplus Domestic Product (SDP)' *International Studies Quarterly*.

Barbera, Pablo. 2015. 'Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.' *Political Analysis* 23(1):76–91.

Blaydes, Lisa and Drew A. Linzer. 2008. 'The Political Economy of Women's Support for Fundamentalist Islam.' *World Politics* 60(July):576–609.

Bolck, Annabel, Marcel Croon and Jacques Hagenaars. 2004. 'Estimating Latent

Structure Models with Categorical Variables: One-Step versus Three-Step Estimators.' *Political Analysis* 12(1):3–27.

Bond, Robert M. and Solomon Messing. 2015. 'Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook.' *American Political Science Review* 109(1):62–78.

Campbell, Donald T. and Donald W. Fiske. 1959. 'Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.' *Psychological Bulletin* 56(2):81–105.

Campbell, Donald T. 1960. 'Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity.' *American Psychologist* 15(August):546–553.

Campbell, Donald T. and H. Laurence Ross. 1968. 'Analysis of Data on the Connecticut Speeding Crackdown as a Time-Series Quasi-Experiment.' *Law and Society Review* 3(1):33–54.

Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. 'Stan: A Probabilistic Programming Language.' *Journal of Statistical Software* 76(1):1–32.

Caughey, Devin and Christopher Warshaw. 2015. 'Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model.' *Political Analysis* 23(2):197–211.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin.

Fahrmeir, Ludwig and Alexander Raach. 2007. 'A Bayesian Semiparametric Latent Variable Model for Mixed Responses.' *Psychometrika* 72(3):327. URL: *https://doi.org/10.1007/s11336-007-9010-7*

Fariss, Christopher J. 2014. 'Respect for Human Rights Has Improved over Time: Modeling the Changing Standard of Accountability.' *American Political Science Review* 108(2):297–318.

Fariss, Christopher J. 2018*a*. 'Are Things Really Getting Better? How to Validate Latent Variable Models of Human Rights.' *British Journal of Political Science* 48(1):275–282.

Fariss, Christopher J. 2018*b*. 'The Changing Standard of Accountability and the Positive Relationship between Human Rights Treaty Ratification and Compliance.' *British Journal of Political Science* 48(1):239–272.

Fariss, Christopher J. 2019. 'Yes, Human Rights Practices Are Improving over Time.' *American Political Science Review* 113(3):868–881.

Fariss, Christopher J. and Geoff Dancy. 2017. 'Measuring the Impact of Human Rights: Conceptual and Methodological Debates.' *Annual Review of Law and Social Science* 13:273–294.

Fariss, Christopher J., Michael R. Kenwick and Kevin Reuning. 2020. 'Estimating One-Sided Killings from a Robust Measurement Model of Human Rights.'

Fonseca, Thaís C. O., Marco A. R. Ferreira and Helio S. Migon. 2008. 'Objective Bayesian Analysis for the Student-*t* Regression Model.' *Biometrika* 95(2):325–333.

Furr, Daniel C. 2017. Bayesian and frequentist cross-validation methods for explanatory item response models. PhD thesis, University of California, Berkeley.

Gandhi, Jennifer and Jane Lawrence Sumner. 2019. 'Measuring the Consolidation of Power in Non-Democracies'. *Journal of Politics.* Forthcoming.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: CRC Press.

Geweke, John. 1993. 'Bayesian Treatment of the Independent Student-*t* Linear Model.' *Journal of Applied Econometrics* 8(S1):S19–S40.

Guttman, Louis. 1949. *The Basis for Scalogram Analysis*. Indianapolis: Bobbs-Merrill.

Hare, Christopher, David A. Armstrong, Ryan Bakker, Royce Carroll and Keith T. Poole. 2015. 'Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions.' *American Journal of Political Science* 59(3):759–774.

Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2014. 'Measuring Transparency.' *Political Analysis* 22(4):413–434.

Imai, Kosuke, James Lo and Jonathan Olmsted. 2016. 'Fast Estimation of Ideal Points with Massive Data.' *American Political Science Review* 110(4):631–656.

Jackman, Simon. 2000. 'Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo.' *American Journal of Political Science* 44(2):375–404.

Jackman, Simon. 2001. 'Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking.' *Political Analysis* 9(3):227–241.

Jackman, Simon. 2008. Measurement. In *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady and David Collier, 119–151. Oxford University Press.

Jessee, Stephen A. 2017. '"Don't Know" Responses, Personality and the Measurement of Political Knowledge.' *Political Science Research and Methods* 5(4):711–731.

Jöreskog, Karl G. and Arthur S. Goldberger. 1975. 'Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable.' *Journal of the American Statistical Association* 70(351a):631–639.

Kenwick, Michael R. 2018. 'Self-Reinforcing Civilian Control: A Measurement-Based Analysis of Civil-Military Relations.' *Working Paper*.

Kōnig, Thomas, Moritz Marbach and Moritz Osnabrügge. 2013. 'Estimating Party Positions Across Countries and Time – A Dynamic Latent Variable Model for Manifestos Data.' *Political Analysis* 21(4):468–491.

Krishnakumar, Jaya and A. Nagar. 2008. 'On Exact Statistical Properties of Multidimensional Indices based on Principal Components, Factor Analysis, MIMIC and Structural Equation Models.' *Social Indicators Research* 87:481–496. ID: UNIGE:41664. URL: *https://archive-ouverte.unige.ch/unige:41664*

Lange, Kenneth and Janet S. Sinsheimer. 1993. 'Normal/Independent Distributions and Their Applications in Robust Regression.' *Journal of Computational and Graphical Statistics* 2(2):175–198.

Lange, Kenneth L., Roderick J. A. Little and Jeremy M. G. Taylor. 1989. 'Robust Statistical Modeling Using the *t* Distribution.' *Journal of the American Statistical Association* 84(408):881–896.

Li, Longhai, Shi Qiu, Bei Zhang and Cindy X. Feng. 2016. 'Approximating Cross-Validatory Predictive Evaluation in Bayesian Latent Variable Models with Integrated IS and WAIC.' *Statistics and Computing* 26(4):881–897.

Linzer, Drew A. and Jeffrey K. Staton. 2016. 'A Global Measure of Judicial Independence, 1948–2012.' *Journal of Law and Courts* 3(2):223–256.

Marquardt, Kyle L. and Daniel Pemstein. 2018. 'IRT Models for Expert-Coded Panel Data.' *Political Analysis* 26(4):431–456.

Martin, Andrew D. and Kevin M. Quinn. 2002. 'Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999.' *Political Analysis* 10(2):134–153.

Martin, Andrew D., Kevin M. Quinn, and Lee Epstein. 2005. 'The Median Justice on the U.S. Supreme Court.' *North Carolina Law Review* 83(5):1275–1321.

Mislevy, Robert J. 1991. 'Randomization-Based Inference about Latent Variables from Complex Samples.' *Psychometrika* 56(2):177–196.

Mokken, R. J. 1971. *A Theory and Procedure of Scale Analysis*. Mouton.

Muthén, Bengt O. 1989. 'Latent Variable Modeling in Heterogeneous Populations.' *Psychometrika* 54(4):557–585. URL: *https://doi.org/10.1007/BF02296397*

Pan, Jennifer and Yiqing Xu. 2018. 'China's Ideological Spectrum.' *Journal of Politics* 80(1):254–273.

Pemstein, Daniel, Stephen A. Meserve and James Melton. 2010. 'Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type.' *Political Analysis* 18(4):426–449.

Pérez, Efrén O. 2011. 'The Origins and Implications of Language Effects in Multilingual Surveys: A MIMIC Approach with Application to Latino Political Attitudes.' *Political Analysis* 19(4):434–454.

Plummer, Martyn. 2017. JAGS (Just Another Gibbs Sampler) 4.3.0. URL: *http://mcmc-jags.sourceforge.net/*

Reuning, Kevin, Michael R. Kenwick and Christopher J. Fariss. 2018. 'Replication Data for: Exploring the Dynamics of Latent Variable Models.' URL: *https://doi.org/10.7910/DVN/SSLCFF*

Reuning, Kevin, Michael R. Kenwick and Christopher J. Fariss. 2019. 'Exploring the Dynamics of Latent Variable Models.' *Political Analysis* 27(4):503–517.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: J. Wiley & Sons.

Schnakenberg, Keith E. and Christopher J. Fariss. 2014. 'Dynamic Patterns of Human Rights Practices.' *Political Science Research and Methods* 2(1):1–31.

Shadish, William R. 2010. 'Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings.' *Psychological Methods* 15(1):3–17.

Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth.

Stegmueller, Daniel. 2011. 'Apples and Oranges? The Problem of Equivalence in Comparative Research.' *Political Analysis* 19(4):471–487.

Stegmueller, Daniel. 2013. 'Modeling Dynamic Preferences: A Bayesian Robust Dynamic Latent Ordered Probit Model.' *Political Analysis* 21(3):314–333.

Treier, Shawn and D. Sunshine Hillygus. 2009. 'The Nature of Political Ideology in the Contemporary Electorate.' *Public Opinion Quarterly* 73(4):679–703.

Treier, Shawn and Simon Jackman. 2008. 'Democracy as a Latent Variable.' *American Journal of Political Science* 52(1):201–217.

van Schuur, Wijbrandt H. 2003. 'Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory.' *Political Analysis* 11(2):139–163.

Vehtari, Aki, Andrew Gelman and Jonah Gabry. 2017. 'Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC.' *Statistics and Computing* 27(5): 1413–1432.

Voeten, Erik. 2000. 'Clashes in the Assembly.' *International Organization* 54(2):185–215.

Windett, Jason H., Jeffrey R. Harden and Matthew E. K. Hall. 2015. 'Estimating Dynamic Ideal Points for State Supreme Courts.' *Political Analysis* 23(3):461–469.