doi:10.1017/S000305542400039X

Measurement That Matches Theory: Theory-Driven Identification in Item Response Theory Models

MARCO MORUCCI *New York University, United States*
MARGARET J. FOSTER *Duke University, United States*
KAITLYN WEBSTER *Independent Scholar, United States*
SO JIN LEE *Harvard University, United States*
DAVID A. SIEGEL *Duke University, United States*

Measurement is the weak link between theory and empirical test. Complex concepts such as ideology, identity, and legitimacy are difficult to measure; yet, without measurement that matches theoretical constructs, careful empirical studies may not be testing that which they had intended. Item response theory (IRT) models offer promise by producing transparent and improvable measures of latent factors thought to underlie behavior. Unfortunately, those factors have no intrinsic substantive interpretations. Prior solutions to the substantive interpretation problem require exogenous information about the units, such as legislators or survey respondents, which make up the data; limit analysis to one latent factor; and/or are difficult to generalize. We propose and validate a solution, IRT-M, that produces multiple, potentially correlated, generalizable, latent dimensions, each with substantive meaning that the analyst specifies before analysis to match theoretical concepts. We offer an R package and step-by-step instructions in its use, via an application to survey data.

INTRODUCTION

Modern social science is theoretically rich and methodologically sophisticated. Measurement, the bridge between theory and empirics, lags behind. That is understandable given the complexity of social-scientific theoretical concepts. Democracy, ideology, legitimacy, power, and identity: none of those are easy to pin down with simple proxies. Yet, without measurement that matches theory, careful empirical studies may not truly be testing what they were intended to test (e.g., Adcock and Collier 2001; Barber 2022; Pietryka and MacIntosh 2022; Reuning, Kenwick, and Fariss 2019). Moreover, without measures that maintain their meanings across time and place, scholars working with the same theoretical concepts but in different systems cannot build on each others' work. We aim to improve the connection between theory and measurement by proposing and validating a solution, IRT-M, that derives multiple, potentially correlated, generalizable, latent dimensions from data. Each dimension IRT-M produces possesses a substantive meaning that the analyst specifies before analysis to match theoretical concepts, ensuring the link between theory and measurement.

Ameliorating the measurement problem requires transparently constructed and improvable measures of theoretical concepts. Dimensional-reduction techniques, such as item response theory (IRT) models, assign to each unit in the data a position along each of one or more latent dimensions. Units can range from survey respondents, to legislators, to nations. Positions in latent space are chosen by the model to best predict each unit's responses to a series of items. Thus, positions along latent dimensions predict individuals' behavior. IRT models are transparently constructed as long as their inputs are public. They are improvable as new data become available, via Bayesian estimation. Perhaps for those reasons, IRT models are increasingly employed in the social sciences as measurement tools. While they most frequently have been applied to legislative or judicial ideal point estimation (Bailey and Voeten 2018; Martin and Quinn 2002; Poole and Rosenthal 1985), they have also been used to derive a variety of measures relating to regime traits (Marquardt et al. 2019) or state capacity (Hanson and Sigman 2021), qualities of human rights (Hill Jr. 2016; Schnakenberg and Fariss 2014) or wartime sexual violence (Krüger and Nordås 2020), interstate hostility

Corresponding author: Marco Morucci, Faculty Fellow, Center for Data Science, New York University, United States, marco.morucci@nyu.edu.
Margaret J. Foster, Post-Doctoral Fellow, Department of Political Science, Duke University, United States, margaret.foster@duke.edu.
Kaitlyn Webster, Independent Scholar, United States, kmwebster819@gmail.com.
So Jin Lee, Stanton Nuclear Security Post-Doctoral Fellow, Belfer Center for Science and International Affairs, Harvard University, United States, sojinlee@hks.harvard.edu.
David A. Siegel, Professor, Department of Political Science and Public Policy, Duke University, United States, david.siegel@duke.edu.

Received: November 23, 2021; revised: July 29, 2022; accepted: March 25, 2024.

(Terechshenko 2020), states' preferences over investor protection (Montal, Potz-Nielsen, and Sumner 2020), peace frameworks (Williams et al. 2021), leaders' willingness to use force (Carter and Smith 2020), state trade legislation (Lee and Osgood 2019), international norms (Girard 2021), media freedom (Solis and Waggoner 2020), and women's inclusion, rights, and security (Karim and Hill Jr. 2018).

Unfortunately, dimensional-reduction techniques do not fully solve the measurement problem because they do not intrinsically capture theoretical concepts: the latent dimensions that best predict the data need not match the theoretical concepts of interest, or even have any clear substantive meaning at all. One way the latter might occur would be if the latent dimensions found by an IRT model were complex combinations of clear theoretical concepts, so that each dimension would be difficult to interpret on its own. That is particularly likely in low-dimensional latent spaces that are intended to explain complicated social behavior, a point raised in Aldrich, Montgomery, and Sparks (2014) and to which we return later.

The problem of substantive interpretation is exacerbated by the lack of modeled correlations between dimensions in typical IRT models, since theoretical concepts are often correlated. For example, we may believe theoretically that voting is driven by ideological positions along economic and social dimensions. If so, in order to test the theory we would need a two-dimensional IRT model to return values along those two ideological dimensions. However, if partisanship strongly predicts voting behavior along both economic and social ideological dimensions, the IRT model might instead return one latent dimension that is a combination of economic and social ideological positions, and a second latent dimension, uncorrelated with the first, that either has no clear theoretical meaning or that carries a meaning unrelated to the theory being tested. We discuss that point further in the analysis of roll call data that forms part of our model validation.

The problem of mismatched theory and measure is particularly evident when traditional IRT methods are used to analyze responses to surveys designed with the specific intent of measuring theoretical concepts. In such cases, the survey design includes information about the manner in which survey questions tie to concepts of interest. Yet that information is not used when computing latent dimensions from survey responses in a traditional IRT model. In fact, traditional IRT methods can estimate latent dimensions that capture none of the carefully established links between theoretical concepts and survey responses, simply because there exists a model that fits the data better than one that takes the question design into account.

Prior solutions to the problem of substantive interpretation—that is, solutions intended to avoid incorrectly characterizing the substantive meaning of latent dimensions derived from IRT and other dimensional-reduction techniques—have generally taken three forms. The first involves the use of additional information about units in the data, information exogenous to the data source from which the latent dimensions will be derived. One common example of that occurs in the placement of legislators in a two-dimensional space. Legislative voting patterns are only one of many data sources that can speak to legislators' ideological positions; others include speeches and donor behavior. Analysts can use such exogenous information to fix the positions of a small subset of well-known legislators; voting behavior is then sufficient to discern the others' positions. The second form of solution uses complementary methods to extract more information from the same data source from which the latent dimensions will be derived. One example of that makes use of LDA topic models to associate bill, judicial decision, or metadata text with issues, and those issues with latent dimensions revealed by voting patterns (Gerrish and Blei 2011; 2012; Lauderdale and Clark 2014).

Both forms of solution represent advances over the standard IRT approach in that they enable substantive interpretation of the discovered latent dimensions. However, both are limited in their range of applicability, for related reasons.

First, the data requirements for the application of either solution are not always, or even often, met. The first solution requires exogenous information on units in the data, and as such is often not available. The second requires that the conditions necessary for application of the complementary method be met. In the case of LDA topic modeling, that means either long texts or a limited issue space. The typical anonymous survey—a common source of individual-level data—usually fails to satisfy either requirement.

Second, neither solution guarantees accurate characterization of the substantive meaning of the discovered latent dimensions. In the case of exogenous information, one might use public speeches, for example, to specify that two legislators are positioned at opposite extremes in each dimension of a two-dimensional latent space capturing economic and social ideology. However, both the IRT's output and the public speeches would also be consistent with a two-dimensional latent space in which one dimension corresponded to partisanship and the other to any other topic over which the two legislators had opposite positions. In that case, *neither* dimension might capture social or economic ideological positions. In the case of complementary methods, one is limited by the inherent limitations of the complementary method. For example, nearly all text models suffer from the difficulty of identifying stance and tone (Bestvater and Monroe 2023).

Third, the meanings of the dimensions suggested by both exogenous information and complementary methods can change over time or differ by place or data source. For example, in the context of exogenous information, two latent dimensions of ideology at some time and place might capture economic and civil rights concerns (Poole and Rosenthal 1985), security and religion (Schofield and Sened 2005), or agreement with the Western liberal order and North-South conflict (Bailey and Voeten 2018). But those can change over time, as did the link between the second latent dimension and civil rights in Congress or the North-South

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

dimension in U.N. voting (Bailey and Voeten 2018). Should there be common units across time, place, or data source, then one can standardize measures via those units, but that requires the additional assumption that those units are not changing latent positions over time, an assumption not available in many contexts. In the context of complementary methods, many such methods lack a way to standardize across contexts. For instance, topic models inductively label scales based on what the topic model finds in the data. The degree to which the models shape the uncovered topics varies based on the degree of supervision as well as other preprocessing decisions (Denny and Spirling 2018), and the topics that best predict texts might vary across time, space, and data source, leading to different impressions of ground truth (Foster 2023).

The third form of solution to the problem of substantive interpretation avoids many of those concerns by carefully selecting and employing response data that are thought to be related only to a single latent dimension. That approach has been used productively, such as in measuring leaders' willingness to use force (Carter and Smith 2020), the strength of norms (Girard 2021), respect for human rights (Schnakenberg and Fariss 2014), and peace agreements (Williams et al. 2021). Unfortunately, no matter how careful the selection, it comes with one significant limitation: analysts cannot allow their response data to be generated by more than one latent dimension. There are many cases in which one dimension might be sufficient, particularly as a first approximation. For example, if partisanship theoretically were to drive most voting behavior, a one-dimensional measure of partisan attachment might be sufficient to predict most voting behavior. In that case, a one-dimensional IRT model run on voting data might produce a good measure of partisan attachment. Fariss (2014) makes such an argument with respect to physical integrity rights.

However, very often the theories we want to test contain multiple theoretical concepts that are imperfectly correlated. One could attempt to measure multiple latent dimensions by estimating one dimension at a time, with each measurement using a different subset of the data. The subsets would have to be nonoverlapping in that case; otherwise, the assumption that one latent dimension is sufficient to predict the data would be violated for the overlapping indicators. We view that level of precision across subsets as rare. Further, even should there be such nonoverlapping subsets, that estimation procedure does not model correlation between latent dimensions, rendering it poorly able to capture imperfectly correlated theoretical concepts. As imperfectly correlated theoretical concepts are common in the social sciences—for example, legitimacy and democracy—we view the range of applicability of the one-dimensional solution as somewhat limited.

We offer a novel solution to the measurement problem via a model that enables transparent substantive interpretation across multiple, possibly correlated, latent dimensions. It does so without needing to leverage either exogenous information about individual units in the data or complementary methods. Our solution, which we call IRT-M, is a semi-supervised approach based on Bayesian IRT. In the spirit of the one-dimensional solution, it employs careful selection of response data. In IRT-M, the analyst identifies before the analysis how each latent dimension is supposed to affect each response to each item, and the model uses that information to output a set of theoretically defined latent dimensions that may also be imperfectly correlated. Thus, the range of applicability of IRT-M is considerably broader than that of existing solutions to the problem of substantive interpretation.

The IRT-M model takes as input data in which each of a set of units (e.g., survey respondents, legislators, and peace treaties) respond to an array of items (e.g., answers to survey questions, votes on bills, and elements of treaties). The present iteration of IRT-M requires dichotomous items, with any non-dichotomous item needing to be reformatted into a series of dichotomous ones, but extensions to more general data, as we discuss further, are straightforward. In the pre-analysis step, the analyst hand-codes each item according to its connection to each latent dimension, recalling that latent dimensions capture specific theoretical concepts from one's theory (e.g., perception of threat from immigration, ideological position, and degree to which a peace treaty captures minority rights and security). In that coding, each item–latent dimension pair is assigned a value of $1, -1, 0$, or NA. A 1 (-1) indicates that a unit with greater (lesser) values along that dimension is more likely to respond positively to that item. A 0 indicates that a unit's value along that dimension does not predict that unit's response to that item. An NA indicates that the analyst has no prior belief about whether that dimension influences one's response to that item. The M in IRT-M represents the constraint matrices that capture those coding rules; the rules capture the theory, ensuring that the measurement matches the theory.

Thus, one's theory plays a key role in the IRT-M approach. If the theory captures some aspect of the process that generated the data used to derive the latent dimensions, and if the coding is applied consistently, the model will produce measures of relevant theoretical concepts that are constant in meaning across disparate data sources and across time and place. For example, a second ideological dimension coded to capture civil rights will maintain that substantive meaning across time, even if it becomes less predictive of variation in the data. In contrast, if the theory does not capture the data generating process well, then the coding should reveal that by assigning values of 0 or NA to many item–latent dimension pairs. The model therefore ensures that measurement matches theory both positively, by constraining the IRT model to connect items to theoreticallydetermined latent dimensions, and negatively, by not using items in analysis that are unrelated to the theory at hand.

IRT-M can be used in any setting in which one would use an unsupervised IRT model, as coding all item–dimension pairs as NA when there is no theory turns IRT-M into an unsupervised model. However, it is particularly well suited when there are clear underlying

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

theoretical mechanisms driving the responses of units to items. We use a motivating example drawn from Kentmen-Cin and Erisen (2017) both to make that point and to provide an intuitive introduction and step-by-step guide to IRT-M. Kentmen-Cin and Erisen (2017) describe how in surveys of Europeans, attitudes toward immigration relate to perceptions of different threats that might be induced by immigration, including economic, cultural, and religious threats. We coded the February–March 2021 wave of Eurobarometer for questions related to those three threats, an additional health threat, since the survey took place during the COVID-19 pandemic, and the two outcome attitudes of support for immigration and support for the European Union (EU). We then employed IRT-M to derive posterior distributions of survey respondents' positions along each of the six latent dimensions. We show that, given sufficient related questions, our measures of latent dimensions possess construct validity and our latent threat dimensions are correlated with each other. We also find some correlation of threats to attitudes, particularly support for the EU.

The section following our motivating example offers a more technical presentation of IRT-M in brief and suggests how it may be easily extended to non-dichotomous items. Our fourth section presents two forms of validation of IRT-M. The first employs simulated data to show that IRT-M yields a reduction in error as compared to established approaches to dimensional reduction while still possessing similar convergence properties. Further, that error remains comparable or lower to established approaches even under substantial misspecification of the model. In other words, one need not perfectly code one's theoretical concepts into a constraint matrix in order to accurately measure them.

The second form of validation applies IRT-M to roll call votes in both the 85th and 109th House and Senate. We hand-code all votes in order to employ IRT-M, rather than use exogenous information on a subset of legislators, as does DW-NOMINATE. In doing so, we recover legislators' placements on two theoretically meaningful latent dimensions that are broadly consistent with positions derived from DW-NOMINATE, particularly along the first, theoretically clearer, latent dimension corresponding roughly to economic ideology. At the same time, we also produce a theoretically clear second latent dimension that is usually strongly correlated with the first. That correlation suggests a role for partisan voting behavior. We also, by varying our coding rules, illustrate the manner in which the theory drives the measurement. Together, our two validation exercises indicate that although one need not be perfect in coding the constraint matrix to make use of IRT-M, straying too far from the latent concepts in question risks measuring different concepts entirely.

Though using IRT-M entails an up-front cost in coding a constraint matrix based on one's theory, it produces measurements of positions on conceptually meaningful latent dimensions while eliminating the need for exogenous information about units in the data to identify the model. That opens up numerous applications, which we briefly discuss in the conclusion. We also provide an R package that will allow analysts to employ the IRT-M framework and model in their own analyses.[1]

APPLYING THE IRT-M MODEL

The IRT-M model produces a set of latent dimensions whose substantive meaning derives from the theoretical concepts underlying pre-analysis coding rules. Thus, applying the IRT-M model begins with identifying a theory that possesses a set of theoretical concepts that one desires to measure. IRT-M will translate each concept to a corresponding latent dimension and compute a latent position for each data unit on each dimension. To illustrate application of IRT-M, we use a motivating example drawn from Kentmen-Cin and Erisen (2017), which offers an overview of anti-immigration attitudes in Europe. Kentmen-Cin and Erisen (2017) also calls explicitly, in the abstract, for scholars to "employ methodological techniques that capture the underlying constructs associated with attitude and public opinion," a task for which IRT-M is ideally suited.

Kentmen-Cin and Erisen (2017) elaborate on the manner by which perceptions of threat partially determine one's attitudes toward both immigration and the EU. We draw three distinct dimensions of threat from their discussion of the literature: a sense of economic threat, a sense of cultural threat, and a sense of religious threat. Together with the two attitudes those threat perceptions are theorized to influence, we thus have five theoretical concepts: economic threat, cultural threat, religious threat, support of immigration, and support of the EU. Each of those five concepts corresponds to a substantively meaningful latent dimension, and the theory indicates that people's values on the first three latent dimensions influence their values on the last two.

The next step in applying the IRT-M model is to identify a data source comprising items that one believes would be informative about individual data units' positions along the latent dimensions identified by the theory. In the context of our example, we require data that are informative about each of the three threats and the two attitudes. We identified the Eurobarometer survey for that purpose, and specifically chose the February–March 2021 (94.3) wave due to availability (European Commission 2021). In reading through the survey codebook, it became clear that there was a fourth sense of threat, specific to the COVID-19 pandemic, for which the survey was informative and which might have also influenced attitudes toward immigration and the EU along the same lines as those discussed in Kentmen-Cin and Erisen (2017):

[1] R-package implementation of IRT-M is available at: https://github.com/dasiegel/IRT-M. Replication data are available on the APSR Dataverse (Morucci et al. 2024).

health threat. Thus, we added that latent dimension to our list, giving us six in all.

After identifying the latent theoretical dimensions of interest and a candidate data source, the next step is to produce the constraint matrix. We accomplish that here by coding the items in the data according to their expected relationships to the underlying theoretical dimensions. Each item in the data is an opportunity for the unit to express its positions along the latent dimensions specified in the theory. In the context of the Eurobarometer survey, an item is a possible answer to a survey question, and a positive response to that item is choosing that answer to the question. Each answer to a question is an item as all items must be dichotomous. Converting all non-dichotomous options to a set of dichotomous items is straightforward. For instance, a five-level feeling thermometer survey question would be broken down into five separate dichotomous items. That is known in the machine-learning literature as One Hot encoding. Similarly, a continuous variable would be made dichotomous through the use of thresholds. In the next section, we discuss how to generalize our model for more general items. In other contexts, responses to items may be votes on bills or resolutions, or the presence of particular elements of treaties, constitutions, or other documents.

After creating an array of dichotomous items, one assigns to each item–latent dimension pair one of four possible values: 1, −1, 0, or NA. One assigns a 1 if a positive response to that item would be predicted by a greater value along that latent dimension. Conversely, one assigns a −1 if a positive response to that item would be predicted by a lesser value along that latent dimension. If one believes that values on that latent dimension do not influence the likelihood of a positive response to that item, one assigns a 0. Finally, if one truly has no prior belief about whether and how values on that latent dimension influence the likelihood of a positive response to that item, IRT-M allows the user to assign NA to the item–latent dimension pair.

Returning to our motivating example, consider the latent dimension of economic threat. We would code a survey question that asks about one's personal job situation as relevant to that underlying dimension. If the survey question had four possible responses, ranging from "very good" to "very bad," then there would be four item–latent dimension pairs to code. We would code the two "bad" responses as 1 and the two "good" responses as −1, since they would be predicted by more and less perception of threat, respectively. In contrast, a survey question asking for one's opinion about a common European Asylum system would be coded as 0 along the economic threat dimension, since we do not view responses to that question as influenced appreciably by one's sense of economic threat.

In coding the Eurobarometer survey wave, we are implicitly assuming that the theoretical concepts we are trying to measure are constant across both space and time, at least during the time the survey was fielded. We are also assuming that the connection between those concepts and the survey questions is constant across both space and time. We view those assumptions as fair for our motivating example, but what would happen if either were violated?

The assumption of constant theoretical concepts is a necessary one for the application of IRT-M. Should concepts not apply equally across space or should they change over time, then the application of IRT-M should be restricted to those regions or times in which the concepts are uniform. Otherwise, it is not clear what concept one would be measuring. In contrast, it is not a problem for IRT-M if the tie between concepts and items varies across space or time, as long as the coding rules capture that variation. In some cases, there may be appreciable variation in coding rules across time and space (Fariss 2014; 2019). For example, a survey question regarding discrimination might change in meaning across time due to shifts in a country's demographics, or might mean different things in different countries. In such cases, the coding rules should assign a value to each item–latent dimension pair appropriate to each country, at each time. In other cases, particularly when items are presented to samples of the same population in the same context over time, as they often are for repeated surveys, coding rules may be constant across time. In other words, coding rules should be time and space dependent when necessary.

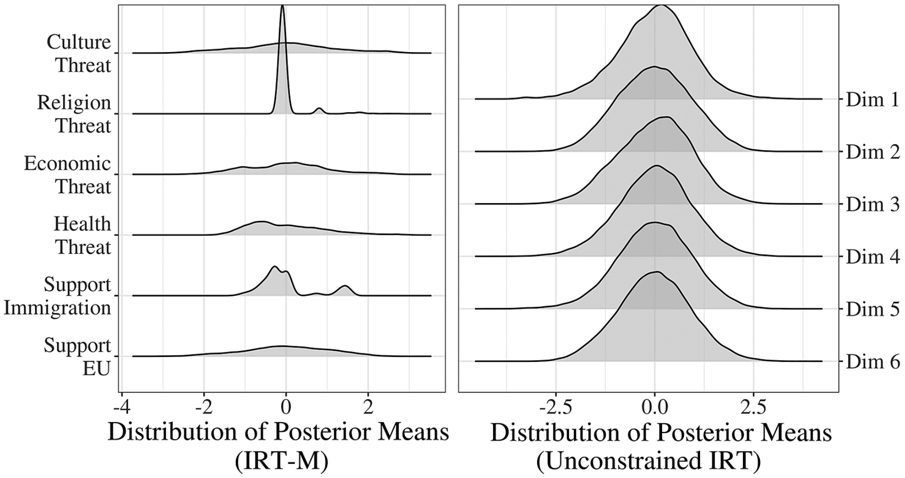
While we expect exact coding rules to vary between coders—connecting items to theoretical concepts allows for subjective assessment—our simulations, described subsequently, illustrate that IRT-M's performance is not significantly reduced even when the constraints are only 50%–75% correct. Thus, based on our simulation, the model can still return reasonable estimates even with partially misspecified constraints.

Our set of constraints captures coding for all item–latent dimension pairs in the model, and its use distinguishes IRT-M from unsupervised IRT models. Once coding is complete, one inputs the constraints (in the form of a matrix) and the data into the IRT-M package IRTM. The output of IRT-M is a user-definable number of draws from the posterior distribution of each unit's position along each latent dimension. In other words, IRT-M maps out the posterior probability distribution capturing each unit's position in the latent space defined by the theoretical concepts encoded in the constraints. One can average those draws—taking the expected value of the posterior distribution—along each latent dimension for each unit. Doing so produces a distribution of the units' expected positions in latent space, as in Figures 1 and 2. That captures the distribution of the theoretical concepts in the sample. One can then compare the distributions of different concepts within a sample, or of the same concepts across samples. Individual units' positions in latent space can also be used in subsequent analyses as individual-level direct measures of theoretical concepts.

We can use our motivating example to illustrate IRT-M's output. While there is a great deal one could do with our six latent dimensions—and we provide coding rules, data, and complete model output for those who might want to explore further—we focus on three figures that showcase both what IRT-M can do and

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

FIGURE 1. Comparison of Distributions of Latent Dimension Posterior Means



Note: These figures are obtained by applying the traditional kernel density estimator with automatic bandwidth selection to the posterior means obtained by each methods for each the *N* respondents. Left: IRT-M output, with named dimensions. Right: unconstrained IRT output.

some differences between IRT-M and an unconstrained IRT model.

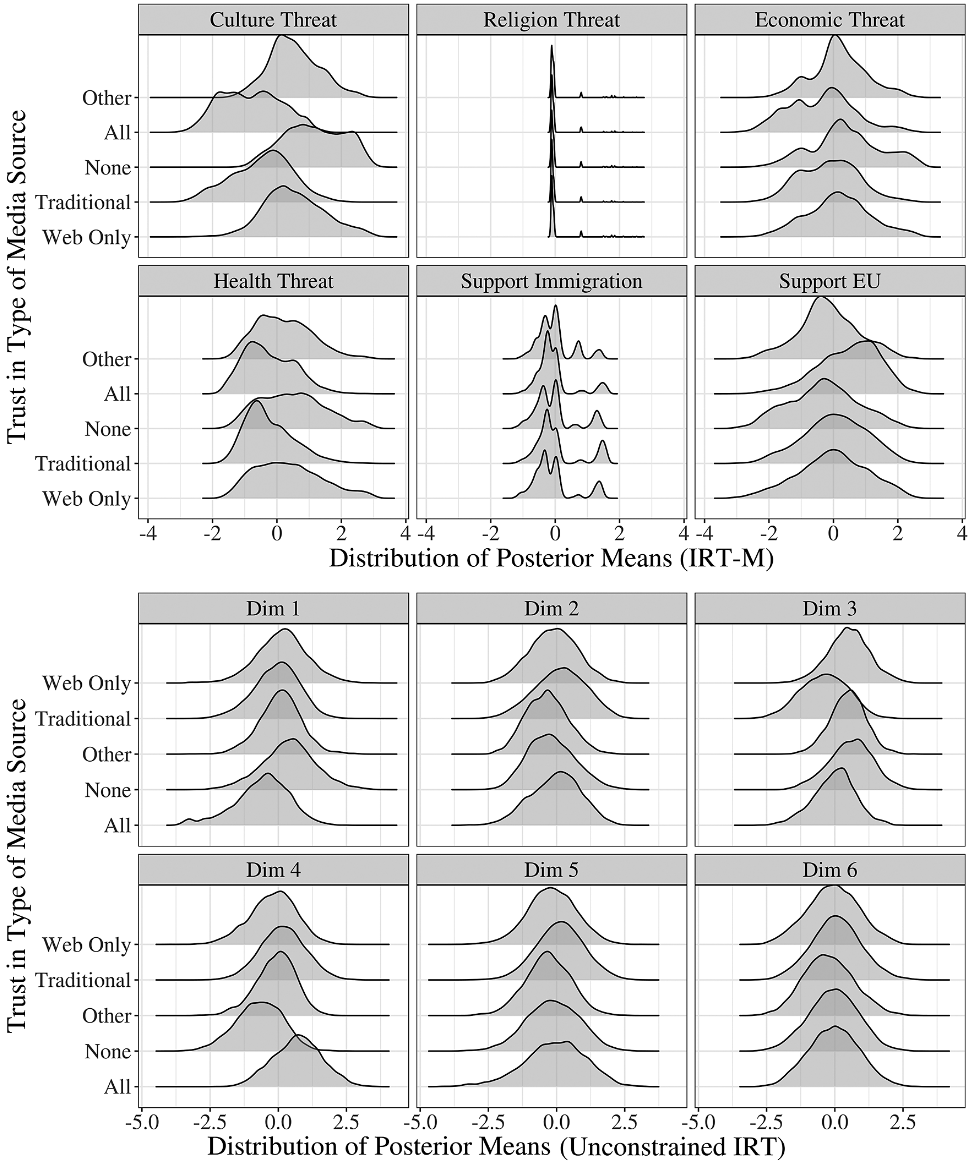
Figure 1 compares the six posterior distributions of units' positions along the latent dimensions arising from IRT-M (on the top) to those arising from an unconstrained six-dimensional IRT model (on the bottom). It is clear that the distributions of latent positions in the sample seem to vary significantly across theoretical concepts in IRT-M, but do not appreciably vary in the unconstrained IRT model. We will return to that point shortly, but first we highlight the distribution arising from IRT-M most different from the rest, that corresponding to religious threat. Unlike the other five dimensions, the distribution of religious threat is sharply peaked near zero with only a small rise away from that peak. Is it the case that, unlike each of the other perceptions of threat, almost no one in the sample perceives religious threat? To answer that, we can turn to the constraints and the coding rules from which they were derived. What we find is that the February–March 2021 wave of Eurobarometer is a poor data source for measuring religious threat. Only three questions, all regarding whether or not terrorism is the most important issue, are reasonably related to religious threat. Further, during a pandemic, very few people felt that terrorism was the most important issue, leading to little variation in response to those questions. Consequently, we should have little confidence in our measure of the religious threat concept. Instead that latent dimension is capturing a more narrow concept of issue importance. That highlights a point made in the previous section: IRT-M does not find latent dimensions that best predict the data. Rather, it computes positions along a latent dimension using only those items tied closely to the theoretical concept connected to that latent dimension. When there are few or no such items, IRT-M will not produce a good measure of that concept. That is a feature, not a bug, of the approach, and we left that dimension in the analysis to reiterate that important point.

What about the other measures of threat? One way to assess the level of confidence we should have in them is by exploring their construct validity. One way to do that is to disaggregate each distribution so as to enable comparisons to behavioral expectations. Figure 2 displays results for one possible disaggregation, in which the population is split according to trust in different media sources. We would expect, all else equal, that those who do not trust the media would be more likely to perceive threat, particularly cultural threat, whereas those who trust the traditional media would be less likely to perceive threat. That is what we see in the distributions derived from IRT-M, particularly for cultural and health threat, less strongly for economic threat. Further, those who trust all media sources generally perceive threat more similarly to those who trust traditional sources than to those who trust no sources. In contrast, only one of the dimensions derived from the unconstrained IRT model, that for Theta3, approaches that behavior. One can repeat that exercise for different subgroups. For example, we find (not shown) that upper class respondents generally perceive less threat than lower class respondents across the three meaningfully measured threat dimensions. That pattern is not observed in the unconstrained model application. Figure 2 thus reiterates another key point: where there are sufficient informative items in the data, IRT-M produces latent dimensions that possess substantive validity.

Figure 3 extends the point about validity to the attitude measures. For both IRT-M and an unconstrained IRT, it displays a correlation matrix linking the six latent dimensions, as well as four values of trust in the media, a social class variable, and one question directly asking about more border controls. The border

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

FIGURE 2. Comparison of Distributions of Posterior Means of Latent Dimensions by Trust in Media



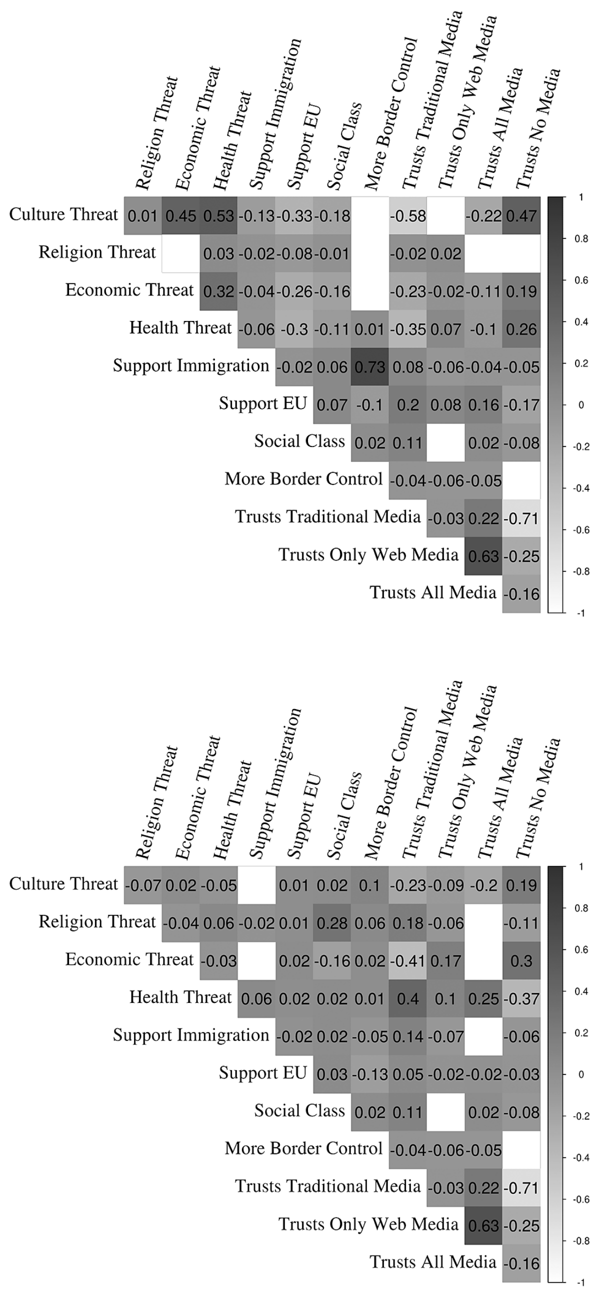
Note: These figures are obtained by applying the traditional kernel density estimator with automatic bandwidth selection to the posterior means obtained by each methods for each the *N* respondents. Top: IRT-M output, with named dimensions. Bottom: unconstrained IRT output.

control question is coded so that higher values imply less desire for more border controls. We would expect it to be strongly correlated with the Support for Immigration latent variable, as it is both substantively related and 1 of about 10 survey questions that are informative about that latent dimension and coded as such in the constraints. The figure bears out that expectation strongly, while the same is not true for any of the latent dimensions arising from the unconstrained IRT model. Thus, whatever is being captured by those unconstrained dimensions, it is not directly interpretable as support for immigration, one outcome variable of interest.

Figure 3 also speaks to another characteristic of IRT-M: it models correlated latent dimensions, in order to capture correlated theoretical concepts. IRT models typically assume independent priors over the latent dimensions; yet there are often theoretical reasons to believe that the underlying latent dimensions of interest are correlated. Explicitly modeling the correlation between dimensions, as does IRT-M, avoids a fundamental disconnect between theory, data, and model. That comes through in the figure: the three meaningful threats are all positively correlated with each other, in addition to a lack of trust in the media, while being negatively correlated with trust in traditional media.

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

FIGURE 3. Correlations: Latent Dimensions, Media Trust, and Support for Border Controls



	Religion Threat	Economic Threat	Health Threat	Support Immigration	Support EU	Social Class	More Border Control	Trusts Traditional Media	Trusts Only Web Media	Trusts All Media	Trusts No Media
Culture Threat	0.01	0.45	0.53	-0.13	-0.33	-0.18		-0.58		-0.22	0.47
Religion Threat			0.03	-0.02	-0.08	-0.01		-0.02	0.02		
Economic Threat			0.32	-0.04	-0.26	-0.16		-0.23	-0.02	-0.11	0.19
Health Threat				-0.06	-0.3	-0.11	0.01	-0.35	0.07	-0.1	0.26
Support Immigration					-0.02	0.06	0.73	0.08	-0.06	-0.04	-0.05
Support EU						0.07	-0.1	0.2	0.08	0.16	-0.17
Social Class							0.02	0.11		0.02	-0.08
More Border Control								-0.04	-0.06	-0.05	
Trusts Traditional Media									-0.03	0.22	-0.71
Trusts Only Web Media										0.63	-0.25
Trusts All Media											-0.16

	Religion Threat	Economic Threat	Health Threat	Support Immigration	Support EU	Social Class	More Border Control	Trusts Traditional Media	Trusts Only Web Media	Trusts All Media	Trusts No Media
Culture Threat	-0.07	0.02	-0.05		0.01	0.02	0.1	-0.23	-0.09	-0.2	0.19
Religion Threat		-0.04	0.06	-0.02	0.01	0.28	0.06	0.18	-0.06		-0.11
Economic Threat			-0.03		0.02	-0.16	0.02	-0.41	0.17		0.3
Health Threat				0.06	0.02	0.02	0.01	0.4	0.1	0.25	-0.37
Support Immigration					-0.02	0.02	-0.05	0.14	-0.07		-0.06
Support EU						0.03	-0.13	0.05	-0.02	-0.02	-0.03
Social Class							0.02	0.11		0.02	-0.08
More Border Control								-0.04	-0.06	-0.05	
Trusts Traditional Media									-0.03	0.22	-0.71
Trusts Only Web Media										0.63	-0.25
Trusts All Media											-0.16

Note: Top: IRT-M output, with named dimensions. Bottom: unconstrained IRT output.

Those correlations are substantively important, and a model without them would not be able to capture different aspects of threat. In contrast, in the unconstrained IRT, the latent dimensions are not substantially correlated with each other, which forces the concepts they represent to be largely unrelated.

Finally, Figure 3 speaks to the idea in Kentmen-Cin and Erisen (2017) that perceived threats are linked to attitudes about immigration and the EU. We find that the three meaningful threats are all negatively correlated with support for both immigration and the EU, though the correlations with the former are weak.

We present the technical specification of IRT-M in the next section, but before doing so, it is worth highlighting one more difference between IRT-M and unconstrained IRT models that is more technical in nature. In an unconstrained IRT model, identification and scaling of the latent dimensions occurs via the use of exogenous information relating to units in the data. For example, ideological ideal points might be taken as given for a small number of legislators. In IRT-M, the constraints both directly allow for model identification and indirectly scale the latent dimensions. The latter occurs automatically in the course of running the model via the creation of anchor points. Anchor points are artificial units that possess extreme positions along the latent dimensions. The constraints characterize behavior consistent with extreme values: if someone has a positive response to every item coded as 1 and a negative response to every item coded as –1 for some latent dimension, then it is impossible for a person to show up as more extreme along that latent dimension in the data at hand. Accordingly, the extreme anchor points we construct serve as reference points for the bounds of the space spanned by the latent dimensions. Other, real, units are measured relative to those anchor points, aiding in interpretation of their values along each latent dimension. No exogenous information relating to the units is needed: all required information arises from the pre-analysis step of coding the constraints, and so relates only to the items used by IRT-M to compute latent positions and is theoretically rather than empirically driven. Anchor points are automatically removed prior to output.

METHODOLOGY

We now describe our approach more formally. There are several key points of divergence from existing IRT-based models such as Zeileis, Kleiber, and Jackman (2008). The first is our pre-analysis coding of the constraints, described earlier. That coding entails conversion of the data to a series of binary (yes/no) items, followed by assignment of a 1, –1, 0, or NA for each item–latent dimension pair. Each assignment is based on whether or not the theoretically defined latent dimension positively, negatively, or does not predict at all the value of that item. The second point of divergence is that identification in our model is accomplished via our constraints, as we will describe. In contrast, typically multidimensional ideal-point-estimation models require for identification purposes some exogenous information relating to the units in the data, usually the ideal points of several units. The third point of divergence is that we explicitly model the covariance of the latent dimensions in our model. IRT models typically assume independent priors for the latent dimensions and do not model their covariance; yet researchers often have theoretical reasons to believe that the underlying latent dimensions of interest are correlated. By modeling covariance, we

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

allow for closer connections between theory, data, and model.

Formally, begin by considering a set of $i = 1, \ldots, N$ units each responding to $k = 1, \ldots, K$ binary (e.g., yes/no) items. Let $y_{ik} \in \{0,1\}$ denote unit i's response to item k, and $\mathbf{Y} \in \{0,1\}^{N \times K}$ be a binary matrix in which row i is unit i's set of responses to all items, and column k is the vector of all units' responses to item k, so that each entry corresponds to y_{ik} as just defined. We would like to model units' responses to items as a function of values along $j = 1, \ldots, d$ latent dimensions, denoted for unit i by the vector $\boldsymbol{\theta}_i$. We will also refer to each θ_{ji} as a factor j. We choose a two-parameter model (Rasch 1960) for this task, as it is commonly used in applied political science (e.g., Clinton, Jackman, and Rivers 2004; Tahk 2018):

$$\Pr(y_{ik} = 1) = g(\boldsymbol{\lambda}_k^T \boldsymbol{\theta}_i - b_k), \tag{1}$$

where $\boldsymbol{\lambda}$ is a d-length vector of loadings, b_k is a negative intercept, and g is a link function, mapping from $\mathcal{R}$ to $[0,1]$. The intercept is commonly understood as a difficulty parameter in the psychometric testing literature, and can be seen here as an average rate of "yes" responses to item k. Each entry in the vector $\boldsymbol{\lambda}_k$ is a real value, denoted by λ_{kj}. The sign of λ_{kj} represents whether a larger factor j (i.e., a larger θ_{ji}) will increase or reduce the likelihood of a "yes" response to item k, while the magnitude of λ_{kj} represents the overall influence of factor j on the likelihood of responding "yes" to item k. For example, a large positive λ_{kj} tells us that units with a large factor j are much more likely to respond "yes" to item k.

The main objective of this article, phrased formally, is to introduce a strategy to encode theoretical information linking latent dimensions and items in the model above, while at the same time resolving its identification issues. We do so by introducing a set of K matrices: $\mathbf{M}_1, \ldots, \mathbf{M}_K$. There is one matrix for each item, and each matrix is $d \times d$ and diagonal. Entries in the diagonal of each matrix are denoted by m_{kjj} and are allowed to either take values in $\mathbb{R}$ or to be missing, where missingness is denoted by $m_{kjj} = \mathcal{NA}$. We then retain the same model in Equation 1, but constrain the loadings $\boldsymbol{\lambda}$ as indicated in Table 1. As Table 1 shows, we constrain our model by (potentially) prespecifying whether the relationship between the response to an item and each latent dimension is positive, negative, zero, or not defined. That accomplishes two fundamental goals. First, it lets one introduce known theoretical connections between items and latent dimensions into the modeling framework. Second, it allows for identification of the model. Before expanding on how our framework accomplishes both of those goals, we give a full description of the model we employ. We choose to adopt a Bayesian specification for our model in Equation 1: we place priors on all of its parameters and express the constraints imposed on $\boldsymbol{\lambda}$ by $\mathbf{M}$ as prior items on the former.

The full hierarchical model is as follows:

$$Y_{ik} = \mathbb{I}_{[\mu_{ik} > 0]}, \tag{2}$$

$$\mu_{ik} \sim \mathcal{N}(\boldsymbol{\lambda}_k^T \boldsymbol{\theta}_i - b_k, 1), \tag{3}$$

$$\lambda_{kj} \sim \begin{cases} \mathcal{N}_{[0\infty]}\left(0, m_{kjj}^2\right), & \text{if } m_{kjj} > 0, \\ \mathcal{N}_{[-\infty 0]}\left(0, m_{kjj}^2\right), & \text{if } m_{kjj} < 0, \\ \delta_{\left[\lambda_{kj}=0\right]}, & \text{if } m_{kjj} = 0, \\ \mathcal{N}(0,5), & \text{if } m_{kjj} = \mathcal{NA}, \end{cases} \quad k = 1, \ldots, K, \tag{4}$$

$$b_k \sim \mathcal{N}(0,1), \tag{5}$$

$$\boldsymbol{\theta}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, N, \tag{6}$$

$$\Sigma \sim \mathcal{IW}_d(v_0, \mathbf{S}_0), \tag{7}$$

where $\mathbb{I}$ is the indicator function, $\mathcal{N}_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution of size d, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathcal{N}_{[a,b]}(\mu, \sigma^2)$ denotes the univariate normal density truncated between real valued constants a and b, and $\mathcal{IW}_d$ denotes the Inverse-Wishart distribution with v_0 degrees of freedom and positive, semi-definite matrix $\mathbf{S}_0$. Finally, the vector of length d that is 0 at all positions is denoted by $\mathbf{0}$.

We choose a probit specification for our link function g as this is standard practice in Bayesian IRT modeling (e.g., Zeileis, Kleiber, and Jackman 2008); however, the model above can also be extended to logistic IRT by changing the distribution of μ_{ij}. We express the constraints imposed on the loading vector, $\boldsymbol{\lambda}$, by the

TABLE 1. Constraints on Loadings λ from Theory

Entry in diagonal *j* of matrix *k*	Constraint on λ_{kj}	Prior belief
$m_{kjj} = 0$	$\lambda_{kjj} = 0$	Factor *j* does not affect answers to item *k*.
$m_{kjj} > 0$	$\lambda_{kjj} \in (0, \infty)$	Factor *j* *positively* affects likelihood of answering yes to item *k*.
$m_{kjj} < 0$	$\lambda_{kjj} \in (-\infty, 0)$	Factor *j* *negatively* affects likelihood of answering yes to item *k*.
$m_{kjj} = \text{NA}$	$\lambda_{kjj} \in (-\infty, \infty)$	No prior on whether and how factor *j* affects answers to item *k*.

M-matrix by varying which prior is chosen for this vector depending on $\mathbf{M}$ (Equation 4). Both ν_0 and $\mathbf{S}_0$ are hyperparameters fixed at standard weakly informative (see, e.g., Gelman et al. 2008) values in our model. That specification is motivated partly by our application, and partly by the fact that it allows us to solve the identification problems associated with the model in Equation 1 in conjunction with the $\mathbf{M}$ matrices. Additional discussion of our model specification is available in Appendix A of the Supplementary Material.

A major contribution of our model specification is that it models the covariance between latent factors as a free parameter and the covariance between loadings as the identity matrix. Model identification while allowing correlation between factors requires shifting the assumption of independence to the item loadings in the M matrices. This is, of course, a consequential modeling assumption. However, in the context of many applications of interest to social scientists, underlying factors—such as dimensions of ideology—are often interlinked. In contrast, items—such as survey question answers or elements of texts—are often functionally independent, frequently by design. By modeling independence at the item loading level, we leverage the data generating process's tendency to produce nearly independent loadings. Our resulting IRT model better fits with many applications of interest. We will expand more on the effect of covariance parameters on model identification and estimation shortly. First, we elaborate on the role of our constraint matrix, M.

Using M to Link Model and Theoretical Knowledge

As indicated in Table 1, in our approach, M-matrices are used to model the extent to which we expect a given item k to load on a factor j. Those loadings are encoded by changing the value of m_{kjj}. A value of 0 indicates that factor j does not influence the likelihood of a "yes" response to item k, and therefore item k should not load on that factor. An absolute value of $|m_{kjj}| = 1$ indicates that we believe that factor j affects item k, but have no strong belief about the magnitude of its effect. A larger absolute value of $|m_{kjj}| > 1$ indicates that factor j should have an effect stronger than the average factor on the likelihood of response to item k. Values of m_{kjj} that are greater (less) than 0 indicate that we believe increasing factor j leads to a greater (lesser) likelihood of a "yes" response to item k. All those items are visibly encoded in our model by the prior items in Equation 4. Those priors are important for both model identification and substantive interpretation of results. On the latter point, they are crucial in assigning a clearly defined meaning to each dimension: dimensions are theoretically prior in our approach, and they may be more or less relevant for understanding a unit's responses to particular items. On the former point, as we will discuss further, constraints embedded in the matrices allow for model identification in the absence of exogenous information about the units.

That does not mean, however, that a user needs to specify all m_{kjj} in our model. Only a small number ($d(1-d)$, as we will see) of m_{kjj} must be set to zero for identification. Further, even that requirement could be weakened if one desired to use exogenous information on units. Whenever a loading is not specified, m_{kjj} is set to missing, and the relevant parameter, λ_{kj}, is drawn from a standard normal distribution, as is normally done in IRT modeling. The most likely scenario in practice is one in which analysts have good knowledge of the loadings of only *some* items, and this model specification allows them to encode their knowledge without having to put strong priors on items about which they do not have strong knowledge.

Using M for Model Identification

Identification issues arise when dealing with models such as that in Equation 1. Here, we show how diagonal matrices such as our constraint matrices can be used to solve those issues, without the need for additional prior distributions or more complex model specifications.

Location and Scale Invariance

Models such as that in Equation 1 are not identified because, for any value of the parameters $\mu_{ik} = \boldsymbol{\lambda}_k^T \boldsymbol{\theta}_i$ and b_k, a different set of parameter values that gives rise to the same likelihood could be constructed with $\mu'_{ik} = \boldsymbol{\lambda}_k^T \boldsymbol{\theta}_i + a$, and $b'_k = b_k - a$, for some real value a. Similarly, for any value of $\boldsymbol{\lambda}_k^T \boldsymbol{\theta}_i$, we could construct a different set of parameters that gives rise to the same likelihood with $\boldsymbol{\lambda}'_k = \boldsymbol{\lambda}_k \cdot a$ and $\boldsymbol{\theta}'_i = \boldsymbol{\theta}_i \cdot \frac{1}{a}$. Those issues are known as location and scale invariance, respectively, and can be prevented by fixing the mean of $\boldsymbol{\theta}_i$ and b_k, and the variance of $\boldsymbol{\lambda}_k$. We implement this with our model prior choices in Equations 4–7.

Rotation Invariance

One other identification issue that arises is that, for any $\boldsymbol{\lambda}_k$ and $\boldsymbol{\theta}_i$, the same exact value of the product $\boldsymbol{\lambda}_k^T \boldsymbol{\theta}_i$ can be obtained with the product of different parameter values $\boldsymbol{\lambda}'_k = \boldsymbol{\lambda}_k \mathbf{A}$ and $\boldsymbol{\theta}'_i = \mathbf{A} \boldsymbol{\theta}_i$, where $\mathbf{A}$ is an orthogonal rotation matrix. This implies that $\boldsymbol{\lambda}_k^T \boldsymbol{\theta}_i = \boldsymbol{\lambda}_k'^T \boldsymbol{\theta}'_i$. There are several possible ways to prevent this issue, commonly known as rotation invariance. In general, $d(d-1)$ constraints must be imposed on the model to prevent this issue (Howe 1955). A researcher may want to impose more than $d(d-1)$ constraints if perceived theoretical clarity allows for this; we discuss this setting in Appendix A of the Supplementary Material. Applications of unconstrained IRT generally achieve identification by specifying, before analysis, values along the latent dimensions for a set of units in the data (see, e.g., Zeileis, Kleiber, and Jackman 2008). Though that approach does encode links between latent dimensions and theoretical concepts, it also requires that the analyst knows the location of some units in the latent space ahead of time, which is not always possible.

Instead, IRT-M imposes the necessary $d(d-1)$ constraints to prevent rotation invariance via the $\mathbf{M}$

matrices. There are two ways to do so. First, one can set at least $d(d–1)$ zeros in the diagonals of all the M-matrices, cumulatively. That is the method we describe above. Formally, this requirement can be stated as

$$\sum_{k=1}^{K}\sum_{j=1}^{d}\mathbb{I}_{[m_{kjj}=0]} \geq d(1–d).$$

Setting $m_{kjj} = 0$ is equivalent to assuming that item k does not load on factor j, or that factor j does not influence the likelihood of responding "yes" to item k. Substantively, that means that one must have a certain amount of prior knowledge regarding the connection between items in the data and the latent dimensions. While one does not need any of the items to load on only one of the latent dimensions, a certain number of items must not load on all the latent dimensions.

Second, one can fix at least $d(1–d)$ of the $\boldsymbol{\theta}_i$. Rather than make assumptions about the locations of units in latent space as is commonly done, however, IRT-M instead creates anchor points from the M matrices. Those anchor points rely on assumptions about links between items and latent dimensions, rather than about units, and so their use allows analysts to solve rotation invariance issues without having to make strong assumptions about the units themselves. Anchor points also set the scale for the latent space, as described previously. Relying solely on anchor points for identification requires the **M** matrices to generate two anchors per dimension, according to the procedure described below. However, any combination of **M** matrix constraints and anchor points allows for identification, as long as the total number of constraints is at least $d(d–1)$.

The intuition behind anchor points is that a hypothetical data unit that responds "yes" to all items that are positively influenced by one factor j, and "no" to all the items that are not will have a large value of factor j. The formal procedure to create a positive anchor point for one factor j, starting from a set of M-matrices, is as follows:

1. Create a new data point, $\mathbf{y}^{new}$.
2. For all items $\ell \in \{m_{kjj} > 0 : k = 1, \ldots, K\}$, that is, all items such that the direction of their loading on factor j is known and positive, set $y_\ell^{new} = 1$.
3. For all items $\ell \in \{m_{kjj} < 0 : k = 1, \ldots, K\}$, that is, all items such that the direction of their loading on factor j is known and negative, set $y_\ell^{new} = 0$.
4. For all the remaining items, r, set y_r^{new} to missing.
5. Set $\theta_j^{new} = D$, where D is some positive constant that is in an extreme of the latent space.

To create a negative anchor for the same factor j, the procedure can be followed exactly, except by setting $y_\ell^{new} = 0$ at step 2, $y_\ell^{new} = 1$ at step 3, and by making D a negative constant at step 5. In the rare occurrence that an anchor point has exactly the same answers as an actual respondent, then the $\boldsymbol{\theta}$ for that respondent should also be set to that of the anchor point just created.

Model Extensions

In order to simplify exposition, we have presented our model in the context of a probit-based IRT model; however, our proposed approach can be easily extended to logistic IRT and multinomial IRT and, while we will not discuss it in detail, to standard factor analysis for a real-valued outcome as well. The hierarchical formulation of our model also allows for simple handling of missing outcome data. Finally, our model can be modified so that the covariance matrix of the loadings, rather than the factors, is learned from the data. We discuss that last modification in Appendix A of the Supplementary Material. All of those extended models allow Gibbs sampling formulations for their MCMC posterior sampling procedures, greatly improving convergence speed. Importantly, none of those extensions affect how analysts specify and use M-matrices in our approach, further supporting the flexibility and wide applicability of our proposed method.

Logistic and Multinomial IRT

Our model can be extended to logistic IRT and multinomial IRT by adopting the Polya–Gamma formulation of logistic regression of Polson, Scott, and Windle (2013). In a binary logistic IRT, we model our responses as $\Pr(Y_{ik} = 1) = \frac{\exp(\psi_{ik})}{1+\exp(\psi_{ik})}$, where $\psi_{ik} = \boldsymbol{\lambda}_k^T\boldsymbol{\theta}_i – b_k$. Polson, Scott, and Windle (2013) show that the likelihood for one item in this model is proportional to

$$L_{ik}(\psi_{ik}) \propto \exp(\gamma_{ik}\psi_{ik}) \int_0^\infty \exp(-\omega_{ik}\psi_{ik}^2/2)p(\omega_{ik})d\omega_{ik},$$

where $\gamma_{ik} = Y_{ik} – 1/2$ and ω_{ik} follows a Polya–Gamma distribution with parameters $(1, 0)$. If all the priors on the IRT parameters $\boldsymbol{\theta}_i, \boldsymbol{\lambda}_k, b_k$ are left as they are in Equations 4–6, then that formulation for the likelihood leads to simple conditionally conjugate updates for all three parameters depending on ω_{ik}, which also has a conditionally conjugate update. That allows for simple extension of the Gibbs sampler for our existing model to the logistic setting without loss of performance in terms of computation time.

Additionally, that model is easily extended to multinomial IRT settings in which one item can have more options than a yes/no response. Suppose that Y_{ik} can be one of $\ell = 1, \ldots, L$ options. In that case, the outcome model becomes: $\Pr(Y_{ik} = \ell) = \frac{\exp(\psi_{ik\ell})}{\sum_{r=1}^{L}\exp(\psi_{ikr})}$, where: $\psi_{ik\ell} = \boldsymbol{\lambda}_{k\ell}^T\boldsymbol{\theta}_i – b_{k\ell}$. As Polson, Scott, and Windle (2013) note, that model can also be expressed as a special case of the binary logistic model just discussed, leading to similar conditional updates for all the IRT parameters. Notably, the number of item loadings is now different for each item, as one loading vector $\boldsymbol{\lambda}_{k\ell}$ has to be estimated for each possible response. That implies that it is possible to further constrain such response-level loadings with M-matrices: in that case, the analyst may specify a different $M_{k\ell}$ for each possible answer to a multiple-choice question. The prior on $\lambda_{k\ell}$ is then determined based on such a matrix in the same way that it is in the current model, for binary responses.

Missing Responses

Finally, our current formulation allows for simple inclusion of missing outcomes (responses) by modifying the Gibbs sampling update for μ_{ik}: if a response is missing, then μ_{ik} can be drawn from an untruncated normal distribution centered at the conditionally updated mean and variance. That is a common tool in Bayesian inference for probit models, and it is explained in detail in, e.g., Gelman et al. (1995). That approach has the clear advantage of allowing analysts to directly deal with missing data without the need to specify other parameters; however, we note that there is evidence that values imputed this way are biased toward the conditional mean of their imputation distribution (Näf et al. 2023). As we already have incorporated this model extension, we caution analysts who believe that imputation-induced bias might affect their estimates too strongly to make use of other imputation methods before employing IRT-M.

MODEL VALIDATION

In this section, we present an in-depth empirical evaluation of our proposed methodology. We first study the performance of IRT-M on a set of simulated data, where ground truth is known and so can be coded directly into the $\mathbf{M}$ matrices. Then, we apply our model to roll call data in the U.S. Congress in order to compare our results to those arising from a common IRT application, in a setting in which information on the positions of some legislators in latent ideological space is thought to be known.

Simulated Data

We generate one hundred artificial datasets from the model specified in Equations 2–7. Parameter values for the main simulation parameters are held constant for each simulated dataset at the following values: $N = 500$, $K = 50$, $\mathbf{S}_0 = \mathbf{I}_{3\times 3}$, $\nu_0 = 3$. We consider four different numbers of latent dimensions—$d = 2, 3, 5,$ and 8—in order to cover a range of substantively useful values.

Comparison with Other Methods

We compare performance of our model (IRT-M) and a correctly specified set of M-matrices to two other methodologies: principal component analysis (PCA) where the principal component associated with the largest eigenvalue is taken as the latent dimension of interest, and a model similar to the one in Equation 1, but without any M-matrices or any constraints on the respondents, as well as uncorrelated factors and loadings. That second comparison model (IRT) is an implementation often used in applications in political science. We compare IRT-M both with and without correlated factors to those two models. For the IRT-M model with uncorrelated factors, we remove the prior in Equation 7 from the model, and fix $\Sigma = \mathbf{I}_{d\times d}$.

We first compare the mean squared error (MSE) of estimates obtained across the three methodologies. The MSE captures the average distance between the location of the true value of the theta (position along each latent dimension) for each observation in the data (each data unit) and the model's estimate of that observation's theta. As in other contexts, an MSE close to zero indicates that the model tends to estimate thetas close to their true values. The results, averaged across dimensions, are shown in Table 2. We note that assessing performance of latent factor models is a complex problem, and MSE can have problems as a metric (Näf et al. 2023). Due to that, we randomly selected several points from each simulated dataset and inspected them to make sure that average MSE did indeed correspond to estimates closer to their true value: we found that to be the case.

We see that our proposed methodology performs uniformly better than both PCA and the same model without M-matrices. That is evidence that including M-matrices in the model does indeed lead to superior estimation performance. Allowing for correlated factors also seems to lead to improvements in estimation error, demonstrating the usefulness of that addition to our modeling framework. We report MSE and 95% Credible Interval coverage results for both the latent factors and loadings in Appendix B.

We next compare our model (IRT-M) with the other Bayesian method (IRT), in terms of speed of convergence of the MCMC sampler for estimating the latent factors. To perform that comparison, we use the version of the effective sample size (ESS) statistic proposed in Vehtari et al. (2021). Results are shown in Table 3; larger values denote more information being captured in the Markov chain. We see that IRT-M displays similar or better convergence to simple IRT after the same number of iterations. This is generally true for both the version of our model with correlated factors and the version without; however, the uncorrelated model displays overall best convergence. This is expected, as exploring the posterior of correlated factors is generally harder than exploring that of uncorrelated factors. In addition, Tables 7 and 8 in Appendix B of the Supplementary Material show that our models achieve similar convergence performance according to other widely used MCMC convergence statistics.

Those results indicate that our proposed approach can achieve performance superior to the standard methodology in terms of MSE without requiring the same strong assumptions and while displaying similar convergence properties. The last is notable because our model requires sampling from a truncated multivariate normal posterior when loadings are direction-constrained by M: that sampling step is nontrivial and requires MCMC tools like rejection or Gibbs sampling in itself (Wilhelm and Manjunath 2010). Our

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

TABLE 2. RMSE for θ—Averaged over Dimensions, Respondents, and Simulations

N	*K*	*d* = 2				*d* = 3				*d* = 5				*d* = 8			
		PCA	IRT	IRT-M	IRT-M	PCA	IRT	IRT-M	IRT-M	PCA	IRT	IRT-M	IRT-M	PCA	IRT	IRT-M	IRT-M
Correlated θ?		No	No	No	Yes	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
10	10	1.953	4.145	0.024	**0.023**	2.431	3.849	0.028	**0.027**	2.576	2.625	0.031	**0.03**	—	—	—	—
10	50	1.781	4.699	**0.011**	0.011	2.059	3.565	**0.013**	0.013	2.372	3.509	0.019	**0.017**	—	—	—	—
10	100	2.284	10.131	**0.008**	0.008	2.092	5.828	**0.009**	0.009	2.309	4.298	0.013	**0.012**	—	—	—	—
10	250	2.327	13.560	0.006	**0.005**	2.099	8.927	**0.006**	0.006	2.357	4.808	0.009	**0.008**	—	—	—	—
10	500	2.082	58.273	**0.004**	0.004	2.141	17.066	**0.004**	0.004	2.183	4.144	0.007	**0.006**	—	—	—	—
50	10	2.034	3.167	0.1	**0.099**	2.103	3.105	0.106	**0.104**	2.332	2.630	0.128	**0.126**	2.521	2.370	0.148	**0.146**
50	50	2.256	3.348	0.028	**0.027**	2.254	3.250	0.034	**0.033**	2.486	2.625	0.044	**0.042**	2.553	2.489	0.060	**0.058**
50	100	2.979	15.258	0.019	**0.018**	2.331	6.120	**0.021**	0.021	2.572	3.435	0.028	**0.027**	2.652	2.752	0.042	**0.04**
50	250	2.561	34.504	**0.011**	0.011	2.721	18.009	**0.011**	0.011	2.841	9.653	0.016	**0.015**	2.615	4.457	0.029	**0.028**
50	500	2.529	69.705	**0.007**	0.007	2.769	80.276	**0.007**	0.007	2.753	16.239	**0.01**	0.01	2.897	7.970	0.023	**0.022**
100	10	2.222	3.213	0.153	**0.15**	2.199	3.196	0.173	**0.167**	2.204	2.643	0.213	**0.208**	2.421	2.555	0.246	**0.243**
100	50	2.640	3.296	0.045	**0.043**	2.667	2.643	0.052	**0.048**	2.682	2.635	0.065	**0.059**	2.621	2.378	0.087	**0.08**
100	100	2.461	8.781	0.025	**0.024**	2.710	3.376	0.031	**0.029**	2.948	2.860	0.038	**0.035**	2.812	2.470	0.050	**0.046**
100	250	2.589	13.037	0.013	**0.012**	2.666	10.391	0.016	**0.015**	3.201	6.670	0.019	**0.018**	3.001	4.523	0.030	**0.026**
100	500	2.395	59.279	0.009	**0.008**	2.717	28.505	0.01	**0.009**	3.011	17.176	0.012	**0.011**	3.236	7.282	0.020	**0.018**
250	10	2.076	3.761	0.267	**0.253**	2.298	2.730	0.305	**0.283**	2.278	2.671	0.362	**0.348**	2.440	2.587	0.451	**0.437**
250	50	2.760	3.951	0.07	**0.064**	2.728	2.560	0.086	**0.075**	2.672	2.537	0.107	**0.09**	2.832	2.336	0.138	**0.115**
250	100	2.099	3.995	0.038	**0.036**	2.896	3.425	0.045	**0.042**	3.291	2.758	0.059	**0.05**	2.971	2.167	0.072	**0.06**
250	250	2.418	5.269	0.017	**0.016**	3.117	3.603	0.021	**0.02**	3.358	3.580	0.025	**0.024**	3.367	2.692	0.032	**0.028**
250	500	3.167	34.596	0.012	**0.011**	3.184	8.250	0.012	**0.011**	3.475	9.579	0.015	**0.014**	3.617	5.845	0.020	**0.017**
500	10	2.122	3.507	0.317	**0.298**	2.117	2.742	0.389	**0.356**	2.346	2.476	0.5	**0.471**	2.393	2.335	0.607	**0.591**
500	50	2.768	4.480	0.094	**0.078**	3.108	2.676	0.11	**0.087**	2.933	2.210	0.144	**0.107**	2.783	2.322	0.184	**0.139**
500	100	2.889	2.642	0.054	**0.044**	2.772	2.576	0.061	**0.05**	3.029	2.403	0.075	**0.058**	3.083	2.220	0.093	**0.07**
500	250	2.781	15.625	0.029	**0.023**	2.970	4.052	0.03	**0.025**	3.262	2.361	0.036	**0.03**	3.649	2.237	0.043	**0.035**
500	500	2.359	5.530	0.017	**0.013**	3.333	6.459	0.021	**0.018**	3.675	3.509	0.022	**0.019**	3.823	2.741	0.024	**0.02**
1000	10	2.256	2.942	0.404	**0.359**	2.246	3.266	0.473	**0.421**	2.458	2.502	0.603	**0.554**	2.399	2.446	0.765	**0.731**
1000	50	2.339	4.734	0.107	**0.088**	2.860	2.932	0.14	**0.094**	2.866	2.389	0.165	**0.114**	2.841	2.397	0.210	**0.15**
1000	100	2.788	2.947	0.069	**0.048**	3.375	2.729	0.08	**0.055**	3.239	2.353	0.087	**0.062**	3.306	2.167	0.112	**0.076**
1000	250	2.703	3.391	0.043	**0.023**	3.445	2.951	0.055	**0.031**	3.561	2.406	0.058	**0.034**	3.813	2.188	0.055	**0.036**
1000	500	2.995	3.471	0.042	**0.015**	2.956	2.814	0.041	**0.023**	3.676	2.413	0.048	**0.025**	3.804	2.242	0.043	**0.025**
2500	10	2.240	4.955	0.479	**0.399**	2.086	3.721	0.597	**0.5**	2.355	2.465	0.736	**0.66**	2.415	2.373	0.939	**0.88**
2500	50	2.561	3.107	0.123	**0.093**	3.073	2.698	0.156	**0.097**	2.992	2.712	0.201	**0.121**	3.043	2.276	0.247	**0.159**
2500	100	2.686	4.588	0.086	**0.048**	3.061	2.720	0.091	**0.057**	3.221	2.383	0.108	**0.065**	3.204	2.267	0.129	**0.079**
2500	250	2.631	8.042	0.07	**0.022**	3.459	2.797	0.098	**0.032**	3.589	2.432	0.087	**0.037**	3.986	2.280	0.085	**0.038**
2500	500	3.101	8.044	0.096	**0.013**	3.468	3.191	0.11	**0.029**	3.527	2.412	0.092	**0.026**	4.099	2.259	0.088	**0.028**

Note: Lower is better; best method for each *N*,*K*,*d* in bold. Values are root-mean-square error for estimated versus true latent factors, averaged over *d* dimensions, *N* units, and 50 simulations. For Bayesian models, estimates are posterior means computed by averaging over 10,000 posterior samples. All results from Bayesian models are computed from four thousand posterior samples obtained from four parallel MCMC chains after two thousand burn-in iterations.

TABLE 3. ESS Convergence for θ

N	K	$d = 2$			$d = 3$			$d = 5$			$d = 8$		
		IRT	IRT-M	IRT-M	IRT	IRT-M	IRT-M	IRT	IRT-M	IRT-M	IRT	IRT-M	IRT-M
Correlated θ?		No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes
10	10	39	74	**77**	38	65	**68**	48	53	**58**	—	—	—
10	50	21	30	**32**	16	29	**30**	13	22	**24**	—	—	—
10	100	18	23	**24**	14	22	**23**	10	18	**19**	—	—	—
10	250	15	18	**19**	12	18	**19**	8	16	**16**	—	—	—
10	500	14	16	**17**	11	**17**	17	7	**15**	15	—	—	—
50	10	68	**149**	149	49	133	**134**	43	121	**123**	51	108	**112**
50	50	49	**83**	83	35	78	**78**	25	65	**65**	20	49	**49**
50	100	42	**63**	61	32	62	**62**	23	55	**55**	17	43	**43**
50	250	30	**44**	44	26	49	**49**	21	47	**47**	16	**38**	38
50	500	27	**38**	37	23	43	**44**	19	**43**	43	16	**36**	36
100	10	97	**180**	179	67	**168**	167	54	**148**	148	57	**140**	139
100	50	72	**117**	117	55	**110**	110	37	**93**	92	28	**71**	70
100	100	61	**93**	92	46	91	**92**	32	**80**	79	24	**63**	62
100	250	45	67	**68**	38	72	**72**	30	**68**	68	23	**57**	56
100	500	39	**58**	58	34	61	**62**	27	**63**	63	22	**54**	54
250	10	151	**230**	229	112	**213**	207	85	**193**	185	79	**179**	173
250	50	122	165	**165**	95	**154**	153	67	**130**	127	50	**104**	100
250	100	107	**143**	143	84	135	**136**	60	**117**	116	44	**95**	93
250	250	82	111	**114**	68	109	**112**	48	99	**101**	37	**85**	84
250	500	67	90	**95**	56	95	**97**	43	92	**93**	34	**80**	80
500	10	199	**253**	248	157	**242**	237	121	**229**	216	108	**206**	194
500	50	162	**194**	193	133	**178**	176	102	**157**	152	76	**133**	126
500	100	142	174	**176**	120	**160**	159	92	**143**	141	69	**115**	111
500	250	122	148	**152**	102	139	**145**	80	125	**126**	59	**105**	104
500	500	102	126	**133**	86	125	**130**	65	113	**117**	50	98	**99**
1000	10	248	**277**	269	203	**265**	255	165	**240**	232	142	**232**	216
1000	50	188	**201**	199	174	**202**	198	139	**177**	167	108	**150**	138
1000	100	180	**194**	192	156	**181**	176	126	**159**	151	100	**134**	127
1000	250	157	171	**174**	142	**164**	163	111	**143**	140	89	**120**	116
1000	500	148	159	**165**	123	151	**151**	100	**135**	133	77	**112**	111
2500	10	**289**	286	275	275	**294**	273	228	**274**	251	194	**248**	243
2500	50	216	**217**	211	**214**	206	198	188	**193**	179	156	**169**	148
2500	100	204	**206**	203	**196**	196	188	173	**176**	164	142	**150**	135
2500	250	189	**193**	187	176	**183**	175	153	**161**	149	130	**139**	125
2500	500	184	**187**	186	167	**173**	164	141	**151**	143	116	**128**	120

Note: Higher is better. Best method for each N,K,d in bold. Values are ESS averaged over N units, d dimensions, and 50 simulations. ESS is a statistic that outputs the number of fully i.i.d. samples that have the same estimation power as the autocorrelated MCMC samples. Here, ESS is computed over 10,000 posterior samples. All results from Bayesian models are computed from four thousand posterior samples obtained from four parallel MCMC chains after two thousand burn-in iterations.

convergence comparison shows that this harder sampling step does not affect speed and MCMC mixing performance. Note here that including the prior on variance hurts us in terms of relative convergence, but we still do well despite that.

Performance under Misspecification of M-*Matrices*

We have seen that our model performs well given a correctly specified set of M-matrices, but in real applications coding of the M-matrices is unlikely to be perfect. Thus, we also assess the robustness of our methodology to misspecification of the M-matrices. We do so in the same simulated-data setting as before, but this time our model is given progressively more misspecified M-matrices. Those misspecified M matrices have two effects, and so misspecification affects the model in two ways. One, it alters elements in the diagonals of the matrices, which alters whether or not a certain item loads on a certain factor. Two, those altered loadings are used to generate anchor points, using the procedure described previously.

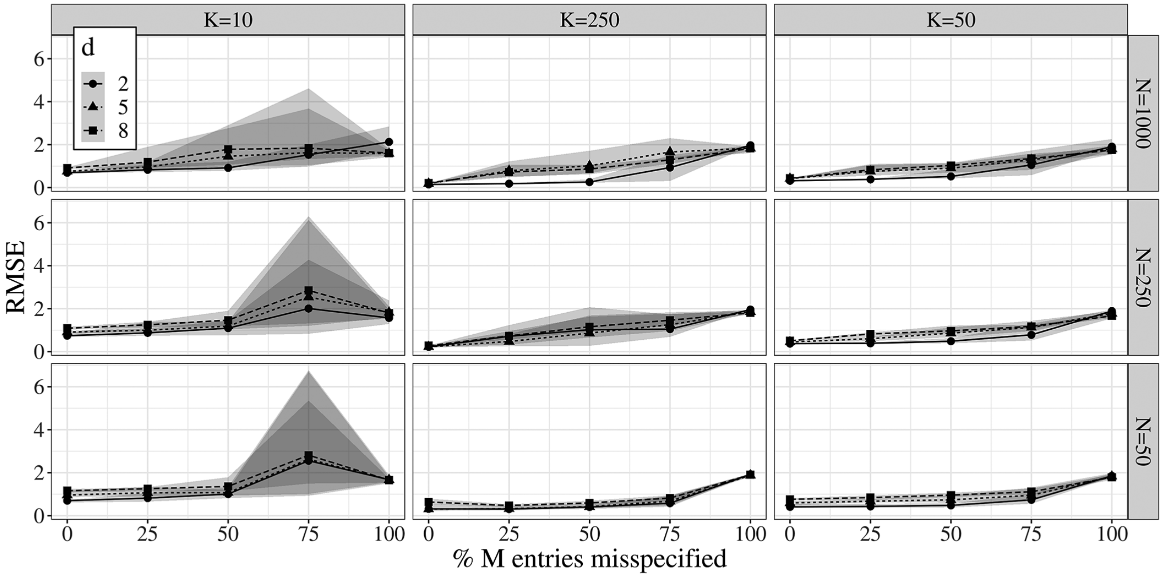
Results are shown in Figure 4. The figure shows an expected trend: when the M-matrices are in large part misspecified, model performance is poor, while a correctly specified model displays very low error. The key point, however, is that, in general, even with only half of the correct specification for the M-matrices, our model still performs very well.

Application to Roll Call Data

Our model performs quite well according to reasonable benchmarks, as long as there exists ground truth to capture. Our motivating example shows that IRT-M also produces reasonable results in the absence of ground truth. However, in that example, we had no existing measures with which to compare ours.

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

FIGURE 4. MSE of the Three-Dimensional IRT Model at Progressively More Misspecified *M*-Matrices



Note: Values on the horizontal axis represent percentages of misspecification of *M*-matrix diagonals. The leftmost point corresponds to a model in which the *M*-matrices are completely correct, while the rightmost point corresponds to a model in which the *M*-matrices are completely misspecified.

Therefore, for our final exercise in model validation, we apply the same procedure detailed in our motivating example to congressional roll call data from four sources: the 85th and 109th U.S. House and Senate. Roll call data are commonly used as inputs to ideal-point-estimation techniques (e.g., Aldrich, Montgomery, and Sparks 2014; Clinton, Jackman, and Rivers 2004; Poole and Rosenthal 1985; 1991; Tahk 2018; Treier 2011). Further, those techniques use exogenous information about legislators' latent ideological positions to identify the models, providing a level of ground truth. In contrast, to apply IRT-M, we will code bills—the items in this context—and make no assumptions regarding legislator positions. We can then compare latent positions derived from applying IRT-M with a two-dimensional latent space to existing ideal point measures such as DW-NOMINATE (DWN) scores, obtainable from voteview.com. That comparison not only reinforces the validity of IRT-M, it also helps us understand better the substantive consequences of employing correlated latent dimensions and varying what gets coded as part of each latent dimension.

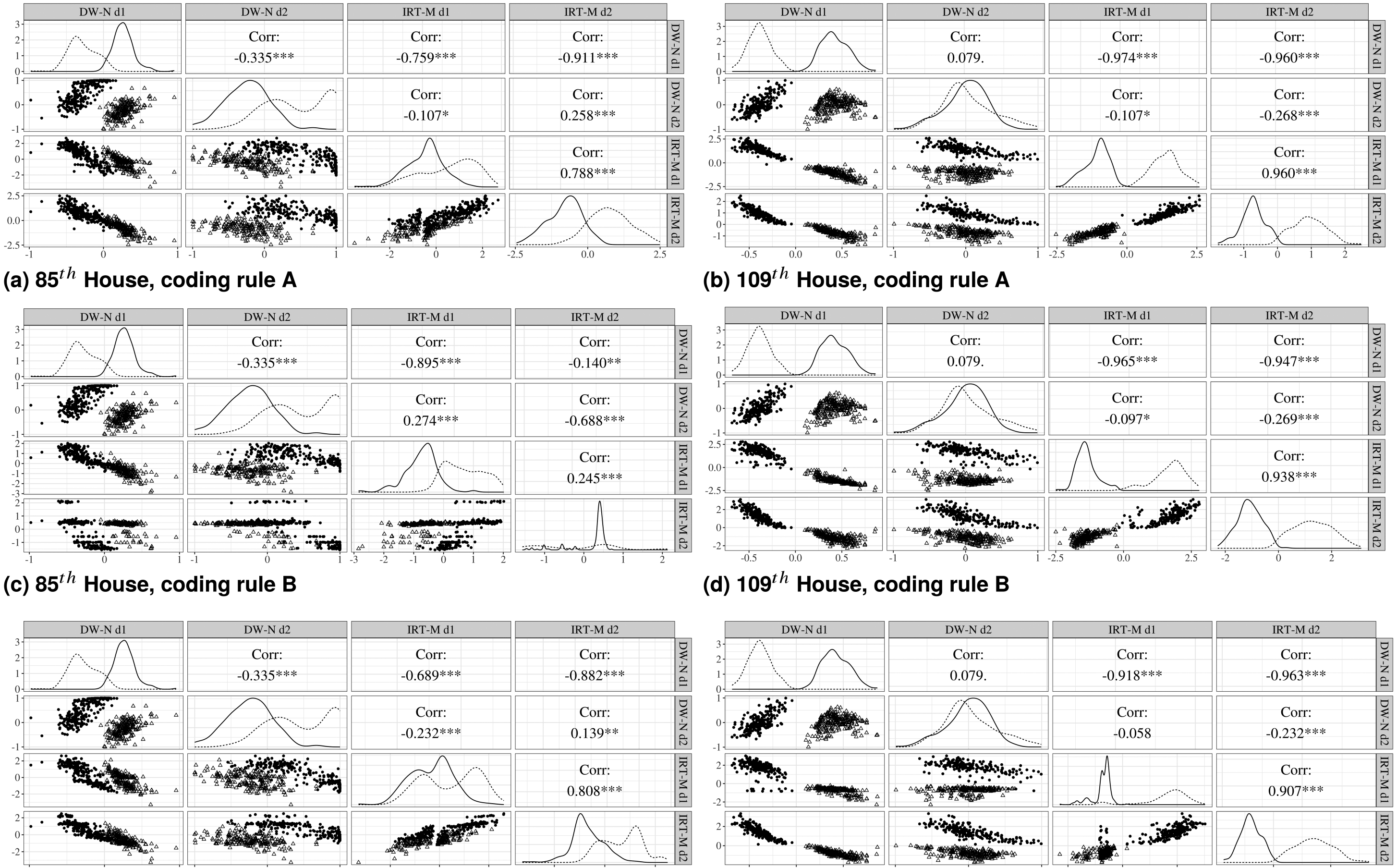
The first step in applying IRT-M to roll call data is to specify a set of theoretically informed latent dimensions. Our approach does not fix the number of latent dimensions; however, to compare to DWN, we need to limit the number of latent dimensions to 2. As there is no theoretically unique set of two latent dimensions given the range of substantive topics addressed in Congress, we opted for three different sets of coding rules. Coding rule A uses dimensions corresponding to (1) Economic/Redistribution and (2) Social/Cultural/Civil Rights/Equality. Coding rule B uses dimensions corresponding to (1) Economy/Distribution/Power and (2) Civil Rights. Coding rule C uses dimensions corresponding to (1) Economy/Public Distribution/Power and (2) Civil Rights/Redistribution. Bills not corresponding to any of those topics are coded 0 for all latent dimensions. The major differences between coding rules are where Redistribution falls and what else gets lumped together with Civil Rights. Appendix C of the Supplementary Material contains a full description of all three coding rules, and we make all coded bills and files to replicate figures available as well.

Figure 5 expresses correlations between different latent dimensions for the 85th and 109th House of Representatives. A similar figure for the Senate is in Appendix D of the Supplementary Material. There are six plots in the figure, one for each House and coding rule combination. Within each plot, the four rows and columns correspond to the first and second latent dimensions derived from DWN and IRT-M. Within each plot are 16 subplots, one for each pair of those four dimensions. The subplots along the diagonal illustrate the distribution of that latent dimension for each of Democrats and Republicans. Above the diagonal, which will be our focus, are correlations between the latent dimension in the column and that in the row, both in the aggregate and broken down by political party. We will focus on those correlations in our brief discussion of model validity, though there is much more one could do with our analysis than we have space for here. High correlations between any two dimensions indicate that those dimensions predict voting behavior similarly.

We first highlight two features of Figure 5 that speak to model validity. One, in all but coding rule B of the

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

FIGURE 5. Correlations between IRT-M and DW-NOMINATE Ideal Points in the House



(a) 85th House, coding rule A

(b) 109th House, coding rule A

(c) 85th House, coding rule B

(d) 109th House, coding rule B

(e) 85th House, coding rule C

(f) 109th House, coding rule C

Note: Each row/column within each subfigure is one of the latent dimensions estimated either by Nominate or IRT-M. The bottom triangle of each subfigure displays scatterplots with each pair of dimensions on each axis. The diagonal contains density plots for each pair of dimensions. The top triangle contains Spearman correlation coefficients for each pair of dimensions. Solid line, triangle = Democratic. Dashed line, filled circle = Republican.

85th House, our two dimensions are highly correlated. In contrast, DWN's two dimensions are mildly correlated in the 85th House and nearly uncorrelated in the 109th. Given the breadth of bill topics covered in all dimensions we have coded, save for the second dimension in coding rule B which uniquely specifies civil rights, that suggests there may be an underlying factor, such as partisanship, driving voting behavior across a range of bills that may obscure more complex ideological preferences (Aldrich, Montgomery, and Sparks 2014). The increasing correlation of our two dimensions over time is consistent with that point. The difference observed in coding rule B of the 85th House is also consistent with that point, given the known schism in the Democratic party at that time over civil rights.

Two, our first latent dimension is consistently highly correlated with DWN's first dimension. That suggests that we are capturing a similar theoretical concept to that captured in DWN's first, economic, dimension. For all but coding rule B of the 85th House, which solely captures Civil Rights, our second dimension is also strongly correlated with DWN's first, for reasons we have noted. However, for coding rule B of the 85th House, our second dimension is much more highly correlated with DWN's second dimension. That provides further confidence in the validity of our measures, since that dimension in DWN is typically interpreted as being related to civil rights as well, unlike DWN's second dimension in the 109th House, which does not share that interpretation.

Together, those two features of Figure 5 support our claim to the validity of our approach, and lend us more confidence that we are able to capture substantive meanings without making assumptions about legislators' preferences. They also illustrate the substantive effects of two aspects of IRT-M: correlated latent dimensions and coding rules. Modeling correlated latent dimensions lets us capture a scenario in which at least one plausible underlying factor drives more than one latent dimension. The underlying factor does not eliminate the meaning of the two latent dimensions—they still have the substantive meanings our coding rules constrained them to have—but it does suggest a richer causal story, of the type described in Aldrich, Montgomery, and Sparks (2014). That richer story also suggests that, absent correlated dimensions and our coding rules, one dimension found by an unconstrained IRT model may capture a complex combination of theoretical concepts driven by an underlying factor, while other dimensions capture concepts not present in the theory, and potentially of less importance. It is possible that DWN applied to the 109th House is doing just that. Its first dimension, termed "Economic/Redistributive" on voteview.com, may be capturing the influence of partisanship, while its second dimension, left without a meaning and called "Other Votes" on voteview.com, may be capturing behavior that is not part of the underlying theory of two-dimensional ideological voting.

Using coding rules to capture the link between theory and measure has several implications. One is that, as long as the coding rules are adjusted for variation in meaning across time or space, the substantive meaning of the latent dimensions will remain constant. As an example, in coding rule B, the second dimension captures civil rights in both time periods. We can see that the split in the Democratic party during the 85th House is no longer appreciably present by the time of the 109th House as both dimensions are highly correlated in the 109th House. A second implication is that the items that go into what theoretical concepts each dimension captures fundamentally change what each dimension is measuring. As we saw, isolating civil rights as its own dimension leads to a very different measure than what we obtained under either set of coding rules in which it was combined with other related, but theoretically distinct, concepts.

CONCLUSION

Measurement is the necessary bridge between theory and empirical test. In the social sciences, it is also the weak link: we traffic in complex concepts such as ideology, identity, and legitimacy, and deriving appropriate measures of those concepts is not trivial. Yet, without measurement that matches our theoretical constructs, our careful empirical studies may not truly be testing what they were intended to test. Further, in the absence of measurement that holds its meaning across time and place, it is difficult to build on each other's work. What captures an aspect of ideology or legitimacy in one context, for instance, may not carry into another.

Dimensional-reduction techniques, such as IRT models, produce improvable measures of latent variables that are thought to underlie behavior, and their construction is transparent given knowledge of the data from which they draw. However, such unsupervised methods do not provide the latent dimensions they discover with intrinsic substantive interpretations. Prior approaches to assigning substantive meaning to latent dimensions either require additional information about units in the data—for example, the ideological positions of certain well-known legislators—or limit the number of latent dimensions to one. As a result, prior approaches cannot solve the problem of substantive interpretation when additional information is not available—as it is not, for example, in anonymous surveys—and the theory under investigation includes multiple, potentially correlated, latent dimensions.

We offer a novel solution to the problem of substantive interpretation: the IRT-M model. Applying the IRT-M model requires coding all responses by a set of units to a set of items—for example, responses by people to survey questions or votes by legislators on bills—according to whether and how that response could be predicted by each latent dimension in one's theory. That coding, used in conjunction with Bayesian IRT, allows IRT-M to produce posterior distributions over each unit's position on each of the latent dimensions, with each latent dimension having the substantive meaning specified in one's theory. We provide two worked examples of IRT-M's use: a motivating

https://doi.org/10.1017/S000305542400039X Published online by Cambridge University Press

example applying it to survey data and latent dimensions of threats and attitudes, and an example applying it to roll call data in the U.S. Congress used as model validation. We also provide an R package that will allow analysts to apply the IRT-M model to their own data and theories.

In applying IRT-M, it is important to keep in mind the importance of theory to the approach. IRT-M is not designed to provide latent dimensions that would best predict responses to items by units in the data. Rather, it is designed to best measure theoretical concepts that may have driven behavior by data units. Thus, a theory as to how the latent dimensions are affecting the responses of the units is essential to the approach. That theory drives the coding of each item–dimension pair. As long as the theory is applied consistently in coding, one can apply IRT-M to disparate data sources across time and place. In all cases, IRT-M will return substantively meaningful latent dimensions, allowing one to make comparisons that previously required exogenous assumptions on individual units. For example, in the roll call application, ideal points located on a civil rights dimension maintain more or less the same meaning across time, as long as sufficient votes related to civil rights continue to be taken.

In the future, we aim to extend our model to additional forms of data, but even limited to dichotomous item data, our approach makes possible consistent, theoretically meaningful measurements that can combine data from numerous sources. We expect such measurements to be substantially more precise than simple indices, let alone single proxies, of the types of theoretical concepts social scientists consider. For instance, our motivating example—employing IRT-M to draw out latent concepts from survey data—illustrates how one can use existing survey data to test social-scientific theories in a straightforward fashion.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S000305542400039X.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/FH74D9.

ACKNOWLEDGEMENTS

The authors thank Miranda Ding, Cat Jeon, Yangfan Ren, and Priya Subramanian for excellent research assistance.

FUNDING STATEMENT

This research was funded by the National Science Foundation under Grant No. SES-1727249.

CONFLICT OF INTEREST

The author declares no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human participants.

REFERENCES

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–46.

Albert, James H. 1992. "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling." *Journal of Educational Statistics* 17 (3): 251–69.

Aldrich, John H., Jacob M. Montgomery, and David B. Sparks. 2014. "Polarization and Ideology: Partisan Sources of Low Dimensionality in Scaled Roll Call Analyses." *Political Analysis* 22 (4): 435–56.

Bailey, Michael A., and Erik Voeten. 2018. "A Two-Dimensional Analysis of Seventy Years of United Nations Voting." *Public Choice* 176 (1): 33–55.

Barber, Michael. 2022. "Comparing Campaign Finance and Vote-Based Measures of Ideology." *Journal of Politics* 84 (1): 613–9.

Bestvater, Samuel E., and Burt L. Monroe. 2023. "Sentiment Is Not Stance: Target-Aware Opinion Classification for Political Text Analysis." *Political Analysis* 31 (2): 235–56.

Carter, Jeff, and Charles E. Smith. 2020. "A Framework for Measuring Leaders' Willingness to Use Force." *American Political Science Review* 114 (4): 1352–8.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (2): 355–70.

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26 (2): 168–89.

European Commission. 2021. Eurobarometer 94.3: Standard Eurobarometer and COVID-19 Pandemic (Study #SI395). Data File Version 1.0. Kantar Public and GESIS—Leibniz Institute for the Social Sciences [producers]. Milan: UniData—Bicocca Data Archive.

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. 2019. "Yes, Human Rights Practices Are Improving over Time." *American Political Science Review* 113 (3): 868–81.

Foster, Margaret J. 2023. "Subject to Change: Quantifying Transformation in Armed Conflict Actors." Working Paper.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *Annals of Applied Statistics* 2 (4): 1360–83.

Gerrish, Sean, and David Blei. 2012. "How They Vote: Issue-Adjusted Models of Legislative Behavior." In *Advances in Neural Information Processing Systems*, Vol. 25, eds. F. Pereira, C. J. Burges, L. Bottou, and K.Q. Weinberger, 2753–61. New York: Curran Associates.

Gerrish, Sean M., and David M. Blei. 2011. "Predicting Legislative Roll Calls from Text." In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, 489–96. Madison, WI: Omnipress.

Geweke, John. 1992. “Evaluating the Accurating of Sampling-Based Approaches to the Calculation of Posterior Moments.” In *Bayesian Statistics* 4, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 169–93. Oxford: Oxford University Press.

Girard, Tyler. 2021. “Reconciling the Theoretical and Empirical Study of International Norms: A New Approach to Measurement.” *American Political Science Review* 115 (1): 331–8.

Hanson, Jonathan K., and Rachel Sigman. 2021. “Leviathan’s Latent Dimensions: Measuring State Capacity for Comparative Political Research.” *Journal of Politics* 83 (4): 1495–510.

Hill Jr., Daniel W. 2016. “Avoiding Obligation: Reservations to Human Rights Treaties.” *Journal of Conflict Resolution* 60 (6): 1129–58.

Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods*, Vol. 580 of Springer Texts in Statistics. New York: Springer.

Howe, William Gerow. 1955. “Some Contributions to Factor Analysis.” Technical Report, Oak Ridge National Laboratory.

Karim, Sabrina, and Daniel Hill Jr. 2018. “The Study of Gender and Women in Cross-National Political Science Research: Rethinking Concepts and Measurement.” Working Paper.

Kentmen-Cin, Cigdem, and Cengiz Erisen. 2017. “Anti-Immigration Attitudes and the Opposition to European Integration: A Critical Assessment.” *European Union Politics* 18 (1): 3–25.

Krüger, Jule, and Ragnhild Nordås. 2020. “A Latent Variable Approach to Measuring Wartime Sexual Violence.” *Journal of Peace Research* 57 (6): 728–39.

Lauderdale, Benjamin E., and Tom S. Clark. 2014. “Scaling Politically Meaningful Dimensions Using Texts and Votes.” *American Journal of Political Science* 58 (3): 754–71.

Lee, Jieun, and Iain Osgood. 2019. “Exports, Jobs, Growth! Congressional Hearings on US Trade Agreements.” *Economics & Politics* 31 (1): 1–26.

Lord, Frederic M. 1953. “An Application of Confidence Intervals and of Maximum Likelihood to the Estimation of an Examinee’s Ability.” *Psychometrika* 18 (1): 57–76.

Marquardt, Kyle L., Daniel Pemstein, Brigitte Seim, and Yi-ting Wang. 2019. “What Makes Experts Reliable? Expert Reliability and the Estimation of Latent Traits.” *Research & Politics* 6 (4). https://doi.org/10.1177/2053168019879561.

Martin, Andrew D., and Kevin M. Quinn. 2002. “Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the US Supreme Court, 1953–1999” *Political Analysis* 10 (2): 134–53.

Montal, Florencia, Carly Potz-Nielsen, and Jane Lawrence Sumner. 2020. “What States Want: Estimating Ideal Points from International Investment Treaty Content.” *Journal of Peace Research* 57 (6): 679–91.

Morucci, Marco, Margaret J. Foster, Kaitlyn Webster, So Jin Lee, and David A. Siegel. 2024. “Replication Data for: Measurement That Matches Theory: Theory-Driven Identification in Item Response Theory Models.” Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/FH74D9.

Näf, Jeffrey, Meta-Lina Spohn, Loris Michel, and Nicolai Meinshausen. 2023. “Imputation Scores.” *Annals of Applied Statistics* 17 (3): 2452–72.

Pietryka, Matthew T., and Randall C. MacIntosh. 2022. “ANES Scales Often Do Not Measure What You Think They Measure.” *Journal of Politics* 84 (2): 1074–90.

Polson, Nicholas G., James G. Scott, and Jesse Windle. 2013. “Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables.” *Journal of the American Statistical Association* 108 (504): 1339–49.

Poole, Keith T., and Howard Rosenthal. 1985. “A Spatial Model for Legislative Roll Call Analysis.” *American Journal of Political Science* 29 (2): 357–84.

Poole, Keith T., and Howard Rosenthal. 1991. “Patterns of Congressional Voting.” *American Journal of Political Science* 35 (1): 228–78.

Rasch, Georg. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Reuning, Kevin, Michael R. Kenwick, and Christopher J. Fariss. 2019. “Exploring the Dynamics of Latent Variable Models.” *Political Analysis* 27 (4): 503–17.

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

Schofield, Norman, and Itai Sened. 2005. “Multiparty Competition in Israel, 1988–96.” *British Journal of Political Science* 35 (4): 635–63.

Solis, Jonathan A., and Philip D. Waggoner. 2020. “Measuring Media Freedom: An Item Response Theory Analysis of Existing Indicators.” *British Journal of Political Science* 51 (4): 1–20.

Tahk, Alexander. 2018. “Nonparametric Ideal-Point Estimation and Inference.” *Political Analysis* 26 (2): 131–46.

Terechshenko, Zhanna. 2020. “Hot under the Collar: A Latent Measure of Interstate Hostility.” *Journal of Peace Research* 57 (6): 764–76.

Treier, Shawn. 2011. “Comparing Ideal Points across Institutions and Time.” *American Politics Research* 39 (5): 804–31.

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. “Rank-Normalization, Folding, and Localization: An Improved R for Assessing Convergence of MCMC (with Discussion).” *Bayesian Analysis* 16 (2): 667–718.

Wilhelm, Stefan, and B. G. Manjunath. 2010. “tmvtnorm: A Package for the Truncated Multivariate Normal Distribution.” *Sigma* 2 (2): 25–9.

Williams, Rob, Daniel J. Gustafson, Stephen E. Gent, and Mark J. C. Crescenzi. 2021. “A Latent Variable Approach to Measuring and Explaining Peace Agreement Strength.” *Political Science Research and Methods* 9 (1): 89–105.

Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. “Regression Models for Count Data in R.” *Journal of Statistical Software* 27 (8): 1–25.