# ORDINAL OUTCOME STATE-SPACE MODELS FOR INTENSIVE LONGITUDINAL DATA

TEAGUE R. HENRY

UNIVERSITY OF VIRGINIA

LINDLEY R. SLIPETZ AND AMI FALK

UNIVERSITY OF VIRGINIA

JIAXING QIU

UNIVERSITY OF VIRGINIA

MENG CHEN

UNIVERSITY OF OKLAHOMA

Intensive longitudinal (IL) data are increasingly prevalent in psychological science, coinciding with technological advancements that make it simple to deploy study designs such as daily diary and ecological momentary assessments. IL data are characterized by a rapid rate of data collection (1+ collections per day), over a period of time, allowing for the capture of the dynamics that underlie psychological and behavioral processes. One powerful framework for analyzing IL data is state-space modeling, where observed variables are considered measurements for underlying states (i.e., latent variables) that change together over time. However, state-space modeling has typically relied on continuous measurements, whereas psychological data often come in the form of ordinal measurements such as Likert scale items. In this manuscript, we develop a general estimation approach for state-space models with ordinal measurements, specifically focusing on a graded response model for Likert scale items. We evaluate the performance of our model and estimator against that of the commonly used "linear approximation" model, which treats ordinal measurements as though they are continuous. We find that our model resulted in unbiased estimates of the state dynamics, while the linear approximation resulted in strongly biased estimates of the state dynamics. Finally, we develop an approximate standard error, termed slice standard errors and show that these approximate standard errors are more liberal than true standard errors (i.e., smaller) at a consistent bias.

Key words: state-space modeling, intensive longitudinal data, ecological momentary assessment, ordinal measurements, item response theory, particle filtering.

Intensive longitudinal data in psychological and behavioral sciences typically consist of self-report, behavioral and/or psychophysiological data collected multiple times a day for multiple participants. There are a number of study designs that produce intensive longitudinal data, such as daily diary studies, experience sampling, and ecological momentary assessment. In this manuscript, we develop the estimation approach for modeling intensive longitudinal data in the state-space framework while accounting for the presence of ordinal measurements such as Likert scale-type items. Our motivating example of a study design that produces intensive longitudinal data is that of ecological momentary assessment, a design that is increasingly used in the study of

1203

psychopathology; however, we note that the state-space approach is appropriate for the modeling of any intensive longitudinal data.

Ecological momentary assessment (EMA) is a type of intensive time series in which data are gathered in the participants' natural environments, typically multiple times a day over several days. This results in strong ecological validity of measurement as, in theory, the data reflects the participants' lived experience rather than the effects of a lab. In addition, a benefit of EMA is that it has the ability to capture the near real-time dynamics of the phenomenon under study, rather than offering mere snapshots, as is the case with laboratory-based designs or more traditional longitudinal designs (Wright et al., 2021). The dynamic relationships of psychological variables and behavior can be tracked in near real time in EMA. For instance, consider the relationship between stress and cigarette smoking. In a laboratory study, the participant would be required to recall when they were stressed, when they smoked, and if there was a link between the two (Mote & Fulford, 2020). For EMA, the need for retrospective recall is eliminated as stress and cigarette smoking can be tracked in near real time via smartphone-based EMA data collection. EMA studies of psychopathology have been increasing in the last few years (Stumpp et al., 2023; Saulnier et al., 2022; Seidman et al., 2022) due to both the increased ubiquity of smartphones and the theoretical turn toward complex systems depictions of psychopathology. Of course there are a number of disadvantages to an EMA approach relative to other protocols. Notably, EMA studies are high in participant burden, making recruitment and, in particular, retention difficult. EMA studies, by dint of having more naturalistic data collections, do not allow for the same level of experimental control as laboratory studies. Finally, due to the high participant burden, there is a limit on the number of items that can be administered in any given "ping," making the careful selection of psychometrically valid items an extremely important aspect of study design.

EMA data can be modeled in a variety of ways from observed variable time series models to mixed effects models. Here, we use a general framework that represents observed time-varying variables as indicators of underlying, latent state variables: *the state-space model* (Durbin & Koopman, 2012). Using this framework/model, the temporal dynamics of the psychological phenomenon under study are wholly determined by the temporal relations between the states, which are then "measured" by the observed variables. As a general framework, state-space models can allow for nonlinear relations and unequal time intervals between data collections, but the key benefit of the state-space approach relevant to the work presented here is the ability to account for different measurement types. State-space modeling has been used to study psychological phenomena previously, and we refer readers to the works of Oud et al. (1999), Browne and Nesselroade (2005), Chow et al. (2009), Song and Ferrer (2009) as representative methodological work in this space.

EMA data can be collected in a variety of modalities from psychological batteries (Levinson et al., 2023) to physiological/biological assay data (Ditzen et al., 2008). One common type of data collected in EMA studies (and more broadly in psychological or behavioral studies) are ordinal responses. Generally speaking, an ordinal variable is one where the values of the variable can be rank-ordered, but the numerical distance between the values is not defined. The best way of illustrating this is via example. One common example of an ordinal variable is a Likert scale item, which, with 5 responses might have the following options: 1—strongly disagree, 2—disagree, 3—no opinion, 4—agree, 5—strongly agree. While the numerical values can be ordered based on the strength and direction of the agreement, the theoretical distance between 1—strongly disagree and 2—disagree is not necessarily the same distance between 2—disagree and 3—no opinion. Similarly, another ordinal type item would be a binned count of substance use: 0—no use, 1—1–2 drinks, 2—3–5 drinks, 3—5+ drinks. Here, the distances between the response categories are prima facie unequal and ill-defined. The issue with ordinal items is, of course, that they are not continuous. Most analytic methods commonly used in psychological science treat ordinal variables as continuous (indeed, the above references for state-space modeling applied to psychological data

all make use of this approximation), and previous work on the quality of this approximation has shown that this results in decreased true positives, increased false positives, and biased effect size estimates in a simulation study where parameter values were known (Liddell & Krushke, 2017) . In this manuscript, we develop a general estimating approach for state-space models with non-continuous outcomes, and show how this estimating approach can be used to fit state-space models with ordinal outcomes using graded response model measurements. We evaluate the performance of our approach in a simulation study against that of the linear approximation approach that is commonly used in analyzing psychological data.

The remainder of this paper is structured as follows: We begin with an overview of the state-space modeling framework and discuss specific technical issues that arise during model estimation. We then describe the ordinal outcome state-space model that is the core contribution of this manuscript. Following that, we describe the four central issues and solutions that make estimating these models feasible: model identification, state filtering, parameter estimation and standard error approximation.

Following the technical description of the model and its estimation methods, we present a simulation study that evaluates the performance of the proposed estimation method and how well a linear approximation approach (i.e., what is most typically used when analyzing ordinal psychological data) can recover state dynamics. Next, we demonstrate the use of this method on an empirical dataset, and compare with the performance of the linear approximation approach. We close with a discussion of the modeling approach and its current limitations, as well as a discussion of future directions in this methodological space.

## 1. State-Space Models

State-space modeling is a general framework for representing a dynamical process of latent variables (states) which have observed variables as their measurements (Durbin & Koopman, 2012; Kalman, 1960) . While the state-space modeling framework can be seen as related to dynamic structural equation modeling, particularly when both the latent states and the measured variables are normally distributed with linear dynamics (Asparouhov et al., 2018) , the state-space approach is a broader framework than dynamic structural equation modeling. State-space models allow for, in theory, nonlinear dynamics and non-normally distributed states/measurements (Durbin & Koopman, 2012) , though most current implementations of state-space models lack the same capability as DSEM for simultaneous multi-participant models.

A simple example of a state-space model is that of a discrete-time, time-invariant, normal-state-normal-measurement with linear state dynamics, which can be described with the following equations:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \varepsilon_t \tag{1}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \varsigma_{\mathbf{t}} \tag{2}$$

$$\varepsilon_t \sim N_p(0, \mathbf{\Sigma}) \tag{3}$$

$$\varsigma_t \sim N_q(0, \mathbf{\Psi}) \tag{4}$$

where $\mathbf{x}_t$ is a $p$-dimensional vector of state values at time $t = 1, 2, \ldots$, $\mathbf{y}_t$ is a $q$-dimensional vector of observed measurements at time $t = 1, 2, \ldots$, $\mathbf{A}$ is a $p \times p$ matrix that governs the dynamics of the states over time, $\mathbf{C}$ is a $q \times p$ matrix of state-measurement loadings, $\boldsymbol{\varepsilon}$ is the multivariate normally distributed *disturbance* term with mean 0 and covariance matrix $\mathbf{\Sigma}$, and $\boldsymbol{\varsigma}$ is the multivariate normally distributed *measurement error* term with mean 0 and covariance matrix $\mathbf{\Psi}$.

To unpack the description of this model as "discrete time, time invariant, normal state-normal measurement with linear state dynamics": Discrete time refers to the equally spaced intervals with respect to $t$. In a discrete time state-space model, the time interval between measurements is assumed to be constant (and therefore ignorable). Time invariant refers to the various parameter matrices ($\mathbf{A}$, $\mathbf{C}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Theta}$) not varying as a function of $t$. Normal state-normal measurement corresponds to the use of multivariate normal disturbance and measurement error distributions, while the linear state dynamics refer to the use of $\mathbf{A}\mathbf{x}_t$ as the state dynamics.

It is important to note that each one of these assumptions can, in theory if not in practice, be relaxed. State-space models can be proposed in continuous time as differential equations, allowing for unequally spaced intervals  (Oud & Jansen, 2000) . All parameters can be made to be time-varying  (Fisher et al., 2022) , while states and measurements can have non-normal marginal distributions  (Kitagawa, 1987) . Finally, nonlinear state dynamics are also possible to represent in the state-space modeling framework (e.g., instead of linear multiplication $\mathbf{A}x_t$, the state dynamics can be a nonlinear function $f(\mathbf{A}, x_t)$)  (Kitagawa, 1996) .

However, as with any complex modeling approach, the devil is in the (estimation) details. State-space modeling originated in an engineering context  (Kalman, 1960) , with some of the first public use cases being for the navigation and control of the Apollo space mission  (McGee & Schmidt, 1985) . In an engineering context, the model parameters are often considered fixed and known, and the interest is in online estimating the trajectories of the states in real time. Additionally, within the engineering context the measured variables, while not necessarily normally distributed, tend to be continuously distributed, which simplifies estimation considerably. The estimation of the state values is known as prediction, filtering or smoothing, depending on if the states are estimated using only previous state estimates, previous state estimates and current measurements, or all state estimates and measurements past and future, respectively. In the context of parameter estimation, filtering is the most relevant, and we will focus our discussion on filtering in this manuscript.

The first filtering approach, the Kalman filter  (Kalman, 1960) , was developed for models of the form described in Eqs. 1–4 and provides an analytic solution to estimating the expected value of states at each timepoint. Since its development, the Kalman filter has been extended to continuous time models (Kalman-Bucy; Kalman & Bucy, 1961), nonlinear state dynamics (Extended Kalman; Sorenson, 1985), nonlinear measurement functions (Unscented Kalman; Wan & Van Der Merwe, 2000), and regime-switching state-space models (Kim-Nelson Filter Kim & Nelson, 2017), to name just a few of the variants. It is important to note here that all of these analytic filters assume that the measurements are continuous. Filtering when measurements are discrete/categorical requires the use of a different class of methods, which will be discussed later in this manuscript.

Parameter estimation (or system identification as it can be known) is a rarer need in most common applications of state-space models, but is extremely important when applying state-space models to psychological or behavioral data. In the case of the model described in Eqs. 1–4, the Kalman filter permits a prediction error decomposition approach to calculating the likelihood directly. This in turn allows for gradient-based optimization; however, relaxing the various assumptions of that model quickly results in intractable likelihood expressions  (Durbin & Koopman, 2012) . Additionally, state-space models tend to have ill-behaved likelihood surfaces due to weakly identified parameters  (Ionides et al., 2015) , making gradient-based approaches to optimization difficult, if not impossible, outside of a restricted subset of model types.

## 2. Ordinal Measurement State-Space Models

As stated previously, the issue with ordinal measurements in a state-space modeling framework is that they are not continuous. This in turn makes most analytic filtering approaches, such as traditional Kalman filtering and the aforementioned variants, inappropriate. Applying standard approaches for fitting state-space models to ordinal outcomes is analogous to the use of fitting linear regression models to dichotomous data (and for that matter, fitting linear regression models to Likert type data). In the current manuscript, we refer to the use of continuous outcome state-space methods on ordinal data as the *linear approximation* approach, and note that this linear approximation approach is used in many applications of state-space modeling to psychological data.

Before developing our framework for ordinal measurement state-space models, there are several previous approaches to estimating state-space models for ordinal data that need to be discussed. Broadly, previous work in this area falls into two categories: Bayesian-based methods and the work of van Rijn (2008) which extends the work of Fahrmeir and Tutz (2001).

State-space models with ordinal/categorical outcomes have been proposed under a Bayesian estimation framework by a number of authors (Wang et al., 2013; Chaubert et al., 2008) . The Bayesian approach permits flexible measurement distributions as well as arbitrary state dynamics (up to the limits of identification) using MCMC-based estimating techniques such as Metropolis-Hastings and/or conditional Gibbs sampling. However, the cost of this flexibility is in computation time and the need to specify priors for each parameter. This combination of computational needs with the need for expertise in Bayesian model specification makes the application of state-space modeling with ordinal outcomes difficult for the applied scientist. That being said, Bayesian estimation allows for bespoke models to be constructed and estimated in situations where frequentist estimation has difficulties, and the computational aspects of Bayesian estimation might be relieved with more efficient sampling methods such as Hamiltonian MCMC (Neal, 1996) and/or variational methods (Blei et al., 2017) . As the goal of the current manuscript is to develop a standard frequentist estimation framework for these models, we will put the prior work on Bayesian estimation aside while noting its usefulness and applicability for this general class of models.

Related to Bayesian estimation of state-space models is the work of Asparouhov et al. (2018) who have developed an approach to fitting dynamic structural equation models (DSEM). DSEM can account for ordinal measurements by using probit measurement models. Probit measurement models assume that the underlying latent state is normally distributed, and is mapped onto the ordinal response categories via hard thresholds (e.g., if the state is below .5, $y = 0$, above .5 $y = 1$). The approach that Asparouhov and colleagues describes makes use of Bayesian estimation methods in Mplus (Muthén & Muthén, 1998) and additionally allows for the analysis of multiple participant's data via random effects. However, the probit model has hard thresholds, which in turn assumes that there is no measurement error in the response process (i.e., once the participant's latent state is in a certain region, they will always respond with the same category.) Accounting for measurement error in ordinal responses is better accomplished adopting methodology from item response theory, which both the current work, and the work of van Rijn (2008) do.

The work of van Rijn (2008) comes closest to the goals of the current manuscript. In that work, van Rijn applied the general estimation techniques from Fahrmeir and Tutz (2001), which allows for dynamical systems with exponential family measurement distributions, and is estimated using an iterative Kalman filtering strategy. van Rijn, like the current paper, focused on graded response type outcomes, which can be put in exponential family form. While this approach does provide a consistent estimation framework, it has a number of disadvantages: First, it requires that certain parameters, including the state dynamics (here, **A**) to be fixed or known. This is limiting in an idiographic modeling context, as the interest is typically in the individual differences in the

latent state dynamics. Second, the requirement of exponential family measurement distributions, while broad enough to cover many different ordinal item measurement distributions, removes from consideration a number of useful measurement distributions, most notably zero-inflated distributions such as the zero-inflated negative binomial or zero-inflated Poisson. Finally, the implementation of this technique requires analytic derivations for the measurement models, and there are no open source implementations of the approach available (it must be noted that the code from van Rijn (2008) is available from the author on request). On the other hand, our implementation of the ordinal state-space model uses a measurement family-agnostic estimation method, and does not require model parameters to be fixed and known (after model identification has been achieved), and we offer an open-source implementation of the modeling framework (Falk et al., 2023) .

Given the state of research on state-space models with ordinal measurements, our goal with this project was to develop a general frequentist approach to fitting state-space models with arbitrary measurement models. In this manuscript, we specifically develop our state-space models for graded response (GR; Samejima 1997, ) measurements, a model that links an ordinal response to an underlying continuous state via cumulative logistic functions. GR models were originally developed (as most item response theory models are) in an educational assessment framework, and the idea behind a GR model is to capture how differing levels of underlying ability correspond to different proportions of the item being correct, with higher ability corresponding to more of the item being correct. However, we note here that the estimation methods and methods for producing standard errors that we describe here are applicable to any measurement model, (for example, zero-inflated distributions), after identification considerations have been made.

The GR outcome state-space model can be expressed in the following form:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t \tag{5}$$

$$\boldsymbol{\varepsilon}_t \sim N_p(0, \Sigma) \tag{6}$$

$$p(y_{it} > j|\mathbf{x}_t) = \frac{1}{1 + \exp[-\alpha_i(\delta_i(\mathbf{x}_t) - \beta_{ij})]} \tag{7}$$

where $\mathbf{x}_t$, $\mathbf{A}$, and $\varepsilon_t$ are as previously described; $y_{it}$ is the $i$th ordinal measurement item, $p(y_{it} > j|\mathbf{x}_t)$ is the probability that the response of $y_{it}$ being greater than the $j$th response category given the value of the states at $t$; $\alpha_i$ is the discrimination parameter for item $i$, $\beta_{ij}$ is the $j$th threshold parameter for item $i$, and $\delta_i(\mathbf{x}_t)$ is a $p \to 1$ selector function that assigns each item to a single state. The use of a selector function here is for notational clarity, and corresponds to the assumption that items do not load onto multiple states. However, there is no explicit reason (beyond complexity) for items to not be indicators of multiple states, we simply make that assumption here to reduce computational complexity.

The probability of a response option $j$ for item $\mathbf{y}_{it}$ is then:

$$p(y_{it} = j|\mathbf{x}_t) = p(y_{it} > j|\mathbf{x}_t) - p(y_{it} > j - 1|\mathbf{x}_t) \tag{8}$$

with $p(y_{it} > 0|\mathbf{x}_t) = 1$. Figure 1 shows probability curves across values of $\mathbf{x}_t$ for two five-category graded response items. These items were taken from the PROMIS depression battery  (Pilkonis et al., 2011) , which is an extensively validated and calibrated scale designed to collect high quality data from patients in a medical setting. The top panel shows the responses for the item "In the past 7 days, I felt hopeless", with item parameters $\alpha = 4.46$, and $\beta = [.49, 1.00, 1.71, 2.46]$, while the top panel shows the responses for the item "In the past 7 days, I felt pessimistic", with item parameters $\alpha = 2.38$ and $\beta = [-.53, .41, 1.47, 2.56]$.
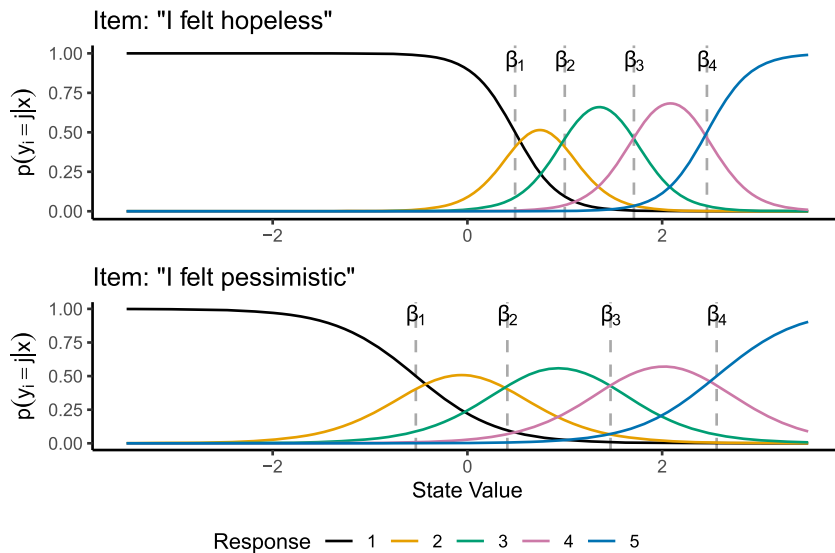
FIGURE 1.
Item response probabilities for two PROMIS depression items.

One key aspect of graded response items is that different items are more or less better at measuring specific areas of the underlying scale. For example, the large $\alpha$ for the item "I felt hopeless", combined with the locations of that item's $\boldsymbol{\beta}$ values shows that "I felt hopeless" has the most optimal measurement when the underlying state is above around .5. The second item, "I felt pessimistic", has less discriminant measurement but captures a broader range of state values.

### 2.1. Identification and Estimation

In developing the estimation strategy for the GR outcome state-space models described in Eqs. 5, 6, and 7, the first consideration we faced was how to identify the latent states. The *identification problem* is a well-known issue in latent variable modeling from both a structural equation modeling and item response theory perspective and refers to the fact that the location and scale of a latent variable is, by definition, arbitrary, whereas with observed data, the location and scale can be directly estimated (i.e., scale indeterminacy; Baker & Kim 2004).

In cross-sectional models using latent variables, the identification problem is solved by fixing a factor loading to a specific value (usually 1) or by fixing a threshold in the probit approach to a specific value, which corresponds with the latent factor taking on the same scale as the indicator, or by fixing the variance of the latent variables to a constant (again, usually 1). Location is easily identified by setting the expected value of the factor to a constant (usually 0). Additionally, for Rasch type measurement models, where the item discrimination parameters are fixed, the scale/location of the latent variables can be identified by fixing a single items threshold parameters, or by constraining the average threshold parameters across items (usually to 0), as an alternative to fixing the scale of the latent variable   (Baker & Kim, 2004) .

While fixing item parameters can identify the GR state-space model, unfortunately, specifying the scale of the latent states is more difficult, for the simple reason that the marginal distribution of the states is not directly parameterized by Eqs. 5 and 6. The marginal distribution of $\mathbf{x}_t$ for all values of $t$ is a function of the model dynamics $\mathbf{A}$ and the innovation covariance matrix $\boldsymbol{\Sigma}$. Specifically, $\mathbf{x}_t \sim N(0, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}$ can be calculated as follows:

$$\text{vec}(\mathbf{\Gamma}) = (\mathbf{I} - \mathbf{A} \otimes \mathbf{A})^{-1}\text{vec}(\mathbf{\Sigma}) \tag{9}$$

where $\text{vec}(\cdot)$ is the vectorizing operator and $\otimes$ is the Kronecker product[1].

It is clear that the marginal variance of the states is dependent on the values of the $\mathbf{A}$ and $\mathbf{\Sigma}$, which in turn complicates estimation as the marginal variance of the states must be constant throughout estimation else identification issues will arise. Unlike with cross-sectional modeling, the marginal variance of the states cannot be fixed by setting values in one of $\mathbf{\Sigma}$ or $\mathbf{A}$ to a constant, as the marginal variance of the states will still depend on the values of the other matrix. To account for this, we use the following set of dynamic constraints. First, we assume that $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ have the following forms.

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1N} \\ \gamma_{12} & 1 & \gamma_{23} & \cdots & \gamma_{2N} \\ \gamma_{13} & \gamma_{23} & 1 & \cdots & \gamma_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{1N} & \gamma_{2N} & \gamma_{3N} & \cdots & 1 \end{bmatrix} \tag{10}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_N \end{bmatrix} \tag{11}$$

We assume that $\mathbf{A}$ results in a stationary process, which ensures that the marginal variance of $\mathbf{x}_t$ exists. These assumptions correspond to setting the scale of the states to 1, while allowing for arbitrary between state marginal correlations. The assumption regarding $\mathbf{\Sigma}$ corresponds to the state innovation processes being independent of one another, above and beyond the state dynamics described by $\mathbf{A}$.

From here, given permissible values for $\mathbf{A}$, one can solve first for the off-diagonal elements of $\mathbf{\Gamma}$ and then by solving for the diagonal elements of $\mathbf{\Sigma}$. This can be easily accomplished using selection matrices.

Let $\mathbf{S}$ be a selection matrix formed by the rows of a $p^2 \times p^2$ identity matrix that correspond to the location of the off-diagonal entries of $\mathbf{\Gamma}$ in $\text{vec}(\mathbf{\Gamma})$, where $p$ is the number of states. Let vector $c$ be the same length as $\text{vec}(\mathbf{\Gamma})$ and consist of 1s in the locations of the diagonal elements of $\Gamma$ and 0s otherwise. Let $\text{vec}(\mathbf{\Gamma})^*$ be a vector formed by only the off-diagonal elements of $\mathbf{\Gamma}$. This vector can be expressed purely as a function of $\mathbf{A}$:

$$\text{vec}(\mathbf{\Gamma})^* = -(\mathbf{S}(\mathbf{I} - \mathbf{A} \otimes \mathbf{A})\mathbf{S}^T)^{-1}\mathbf{S}(\mathbf{I} - \mathbf{A} \otimes \mathbf{A})c. \tag{12}$$

One can then reconstruct the full $\text{vec}(\mathbf{\Gamma})$ using $\text{vec}(\mathbf{\Gamma})$ and use Eq 9 to solve for the diagonal elements of $\mathbf{\Sigma}$.

While this identification technique is computationally expensive, as it requires solving systems of equations whenever $\mathbf{A}$ is changed, it has a significant advantage: because the model is

---

[1]This function is slightly different when dealing with continuous time models where the state dynamics are in differential equation form: $\dot{\mathbf{x}}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$. Here, the function is $\text{vec}(\mathbf{\Gamma}) = (\mathbf{A} \oplus \mathbf{A})^{-1}\text{vec}(\mathbf{\Sigma})$, where $\oplus$ is the Kronecker sum (Vatiwutipong & Phewchean, 2019) .

identified without fixing values in the measurement model, estimated values of **A** are directly comparable for any measurement model. For example, we use this property later in this manuscript to directly compare **A** estimates when the measurement model is the correct GR model vs. a linear approximation.

*2.1.1. Estimation Issues and a 2-Step Solution*     The above procedure for identifying the GR state-space model by constraining the marginal variance of the states to 1 allows for the item parameters ($\boldsymbol{\beta}$, and in theory, $\boldsymbol{\alpha}$) to be estimated along with the parameters for state dynamics (**A**). The simulation study below uses this identification constraint to estimate both the state dynamics and the measurement threshold parameters ($\boldsymbol{\beta}$) successfully. However, as we found in our empirical analysis presented below in Section 7, real data can pose a number of difficulties for estimating both the state dynamics and measurement parameters simultaneously.

The first issue to discuss is that of how empirical under-identification intersects with the specifics of the estimation approach we use (MIF2; Ionides et al., 2015, described in detail in the next sections,) to lead to convergence issues. Here, empirical under-identification refers to the case where there is a marked imbalance in the response categories (e.g., only 1 or 2 instances of response category 7 on a 7 point Likert scale). This in turn makes estimating the threshold parameter associated with that response category difficult in the sense that there is not much information to inform the estimation. In a gradient based estimation approach, the MLE can still be consistently found, but the standard error of that parameter will be high, reflecting the lack of information. In the case of the algorithm we use, this situation leads to convergence issues, as the likelihood surface can be thought of as nearly "flat" in the neighborhood of the MLE. While Ionides et al. (2015) does prove that MIF2 converges asymptotically to the MLE, the rate of convergence is unknown and highly dependent on the specific tuning of the estimation. As such, the empirical under-identification can lead to issues with convergence in all parameters that in theory can be solved by carefully tuning the estimation procedure; in practice, this tuning is difficult and time consuming for an applied researcher to do.

The second issue with estimating the measurement parameters simultaneously with the state dynamics is conceptual: The GR state-space model as presented here is an idiographic model suitable for analyzing participant's timeseries data separately. This means that the item measurement parameters estimated will be idiographic in nature, and that inter-individual differences in the state trajectories will be reflected in those measurement parameters. To put it simply, two participants could have the same estimated state trajectories, but the meaning of the state values will differ due to the idiographicly estimated measurement parameters. Importantly, the state dynamic parameters will still be comparable between subjects, but the specific state values will not. This is a limitation of the idiographic nature of these models, and could be solved by extending this model to multiple subjects or by applying meta-analytic techniques (see Future Directions for more details).

Fortunately, one solution to both of these issues is to use a scale for which item parameters have already been estimated in a large calibration sample. By fixing the measurement model to a priori known parameters and only estimating the state-dynamics, one can radically reduce the dimensionality of the parameter space. This in turn results in easier convergence while using the MIF2 approach for estimation. Additionally, using an calibrated scale and adding a mean vector to the state trajectories would allow the idiographic trajectories of the states to be comparable across participants.

## 2.2. State Estimation via Particle Filtering

There are two halves to the estimation of state-space models. The first half is the estimation of the state values themselves, given a set of parameter estimates and the observed data. Estimating

the expected value of a state at time $t$ (i.e., $\mathbf{x}_t$) given the parameter values and observed variables up to and including time $t$ is known as *filtering*. In state-space modeling, many different approaches to filtering have been developed. One set of approaches can be classified as analytic filtering, and includes the classical Kalman filter, as well as many of its newer variants. Unfortunately, analytic filtering approaches rely on the filtering distribution (i.e., $\mathbf{x}_t|\mathbf{y}_{1:t}, \boldsymbol{\theta}$ with $\boldsymbol{\theta}$ standing for the entire parameter set) being analytically tractable. In the case of our ordinal outcome state-space models, the filtering distribution is not analytically tractable, and a different approach is needed.

The issue of intractable latent variable distributions in cross-sectional item response theory models is well known, and has been solved by either using numeric integration methods or Bayesian estimation approaches  (Baker & Kim, 2004) . Here, we use *particle filtering*  (Moral & Doucet, 2014) , an estimation method conceptually related to Bayesian MCMC, which we describe generally in Algorithm 1.

---

### Algorithm 1
#### Basic Particle Filter for State Estimation

---

Initialize $K$ particles $\mathbf{x}_0^k \sim f_0$
**for** t in 1:T **do**
    Sample $\mathbf{x}_t^k \sim f(\mathbf{x}_{t-1}^k|\boldsymbol{\theta})$
    Calculate $w_k = p(\mathbf{y}_t|\mathbf{x}_t^k, \boldsymbol{\theta})$
    Normalize $w_k : w_k^* = w_k / \sum_k w_k$
    Sample with replacement $K$ particles from $\mathbf{x}_t^k$ with probability $w_k^*$ with replacement.
**end for**

---

A particle filter is a simulation-based state estimation method that iteratively takes informed guesses at the values of $\mathbf{x}_t$ using the state dynamics model $f(\mathbf{x}_t|x_{t-1}, \boldsymbol{\theta})$, determines how good those guesses were by evaluating the likelihood of the observed data given the particles ($w_k = p(\mathbf{y}_t|\mathbf{x}_t^k, \boldsymbol{\theta})$), and then selects the best performing particles at each timepoint via resampling with replacement. While computationally expensive, particle filtering is a turnkey solution to the filtering problem, requiring only the state dynamics and measurement distributions, while making no assumptions about the form of them. In the case of the GR state-space models of Eqs. 5, 6, and 7, $f(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta})$ is Eq. 5, while $p(\mathbf{y}_t|\mathbf{x}_t^i, \boldsymbol{\theta})$ is Eq. 8. The distribution of the initial values $\mathbf{x}_0$ is $f_0 = \text{MVN}(\mathbf{0}, \mathbf{I})$. Finally, $\mathbb{E}[\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t, \boldsymbol{\theta}]$ is simply $\sum^K w_k^* \mathbf{x}_t^k$.

### 2.3. Parameter and State Estimation via the MIF2 Approach of Ionides et al. (2015)

While the particle filter described by Algorithm 1 can be used to estimate the expected value of the states, careful readers will notice that the model parameters $\boldsymbol{\theta}$ are assumed to be fixed and known. While this is not an issue in common engineering applications of state-space models where the model parameters are known, in the analysis of psychological or behavioral data the estimation of the parameters is of greatest importance. To tackle this side of the estimation problem, we employ yet more particle filtering.

To simultaneously estimate parameters and states, we use the multiple iterated filtering approach of Ionides et al. (MIF2; 2015). This algorithm is a general estimation routine for partially observed Markov processes (i.e., most state-space models) and can be conceptualized as a particle filter combined with an optimization technique called simulated annealing. Simulated annealing is a method to find the global optima of a surface by sequentially perturbing an estimate and updating the estimate when the perturbation results in a higher value. Over the course of the optimization, how much the estimate is perturbed is reduced to 0, resulting in the final estimate

that has "frozen" at the global optima (Kirkpatrick et al., 1983). The authors of Ionides et al. (2015) have developed an optimized implementation of the MIF2 algorithm within their `pomp` R package (King et al., 2023, 2016), and our implementation of the GR model makes use of it. One key advantage of the MIF2 algorithm is that it does not require any analytic derivations of gradients or hessians, and only requires the ability to simulate from implied state distributions and evaluate the log-likelihood of the data given a set of parameter estimates. This makes the approach able to work with arbitrary measurement and state distributions.

The MIF2 algorithm applies particle filtering to both the states and parameter values, sampling, evaluating and resampling both over many iterations. The state filtering component of the algorithm is similar to what is described in Algorithm 1, while the parameter filtering component makes use of a sampling distribution that progressively reduces its variance over the iterations. This causes the estimation routine to first search the parameter space, then freeze in an area of high likelihood, ideally the MLE. This converging behavior would not occur if the MIF2 algorithm used a set perturbation variance (Ionides et al., 2015). The MIF2 algorithm is described in Algorithm 2.

---

### Algorithm 2
MIF2 Algorithm of Ionides et al. (2015)

---

Initialize $K$ particles $\boldsymbol{\theta}_0^{0,k} = \boldsymbol{\theta}^*$
**for** m in 1:M **do**
    Sample $\boldsymbol{\theta}_0^{m,k} \sim \text{MVN}(\boldsymbol{\theta}^{m-1,k}, c^m \boldsymbol{\Theta})$ for $k \in 1:K$
    Sample $\mathbf{x}_0^{m,k} \sim f_0(\boldsymbol{\theta}^{m,k})$ for $k \in 1:K$
    **for** t in 1:T **do**
        Sample $\boldsymbol{\theta}_t^{m,k} \sim \text{MVN}(\boldsymbol{\theta}_{t-1}^{m,k}, c^m \boldsymbol{\Theta})$ for $k \in 1:K$
        Sample $\mathbf{x}_t^{m,k} \sim f(\mathbf{x}_t^{m,k}|\mathbf{x}_{t-1}^{m,k}, \boldsymbol{\theta}_t^{m,k})$ for $k \in 1:K$
        Calculate $w_k = p(\mathbf{y}_t|\mathbf{x}_t^{m,k}, \boldsymbol{\theta}_t^{m,k})$ for $k \in 1:K$
        Normalize $w_k : w_k^* = w_k / \sum_k w_k$
        Sample with replacement $K$ particles from $\{\boldsymbol{\theta}_t^{m,k}, \mathbf{x}_t^{m,k}\}$ with probability $w_k^*$.
    **end for**
**end for**

---

In Algorithm 2, $\boldsymbol{\Theta}$ is a diagonal covariance matrix with analyst-specified variances, and $c^m$ is a positive and decreasing sequence such that $c^m \to 0$ as $m \to M$. Ionides et al. (2015) implement this as a geometrically decreasing sequence. The tuning values for this algorithm are the number of particles $K$, the number of iterations $M$, the parameter perturbation covariance matrix $\boldsymbol{\Theta}$ and the cooling sequence $c^m$. Ionides et al. (2015) show that the MIF2 algorithm converges to the MLE under mild regularity conditions, making this approach a general estimation technique for state-space models that does not rely on tractable analytics or numerical derivatives. Now we can map the terms in the MIF2 algorithm to the specifics of the GR state-space model, which are contained in Table 1.

We implement our GR state-space models in the `genss` R package (Falk et al., 2023). This package is under active development, and will be expanded with additional measurement models and state dynamics in the near future.

Note that the discrimination parameters $\alpha_i$ are not estimated in this set, and are instead assumed to be fixed and known. In this manuscript, we fixed $\alpha_i = 1$ (i.e., a homogeneous Graded Response Model) as the MIF2 algorithm had considerable difficulties estimating $\alpha_i$ freely. This is not due to model identification issues, but rather due to the nature of $\alpha_i$. Specifically, the relative

TABLE 1.
GR state-space to MIF2 mapping.

| MIF2 terms | GR state-space terms |
|---|---|
| $\boldsymbol{\theta}$ | $\mathbf{A}, \{\beta_{ij} \ \forall \ i, j\}$ |
| $f_0(\boldsymbol{\theta}^{m,k})$ | $\text{MVN}(\mathbf{0}, \mathbf{I})$ |
| $f(\mathbf{x}_t^{m,k}\|\mathbf{x}_{t-1}^{m,k}, \boldsymbol{\theta}_t^{m,k})$ | $\text{MVN}(A_t^{m,k}\mathbf{x}_{t-1}^{m,k}, \boldsymbol{\Sigma}_\mathbf{t}^{\mathbf{m},\mathbf{k}})$ |
| $p(\mathbf{y}_t\|\mathbf{x}_t^{m,k}, \boldsymbol{\theta}_t^{m,k})$ | $p(\mathbf{y}_{it} = j\|\mathbf{x}_t^{m,k}) = p(\mathbf{y}_{it} > j\|\mathbf{x}_t^{m,k}) - p(\mathbf{y}_{it} > j - 1\|\mathbf{x}_t^{m,k})$ |
| | $p(\mathbf{y}_{it} > j\|\mathbf{x}_t^{m,k}) = \frac{1}{1+\exp[-\alpha_i(\delta_i(\mathbf{x}_t^{m,k})-\beta_{ij}^{m,k})]}$ |

magnitude of $\alpha_i$ only makes a difference to the likelihood of the data when close to 0, while for example $\alpha_i = 5$ and $\alpha_i = 10$ would result in nearly identical likelihood surfaces. This appears to cause difficulty with MIF2's exploration of the likelihood surface. This is a limitation of the estimation approach, and we discuss possible solutions in the Discussion section.

### 2.4. Standard Errors via Slice Likelihoods

While the MIF2 algorithm converges to the MLE under mild regularity conditions, it does not produce the information necessary to estimate the standard errors of the parameters. Calculating standard errors requires some knowledge of the curvature of the likelihood surface at the MLE, and as MIF2 does not use gradient information, the Fisher information matrix needs to be approximated. One potential way of doing this is to calculate numerical derivatives, however this is complicated by the stochastic nature of particle filtering. Here, we take the slice likelihood approach described by Ionides et al. (2006) to approximate standard errors.

Slice likelihood relies on the *local asymptotic normality* (LAN; Le Cam & Lo Yang, 1990) property of the likelihood surface near the MLE. This property says that the likelihood surface is approximately multivariate normal in the area around the MLE. This in turn allows one to estimate the curvature of the likelihood surface at the MLE using a series of quadratic regressions, which can then be used to approximate the Fisher information matrix. Slice likelihood involves evaluating the likelihood at a series of slices surrounding the MLE, usually along each parameter (hence, "slicing" along each parameter axis). Those values are then used in the series of quadratic regressions to estimate the curvature of the likelihood surface. The details of the implementation can be found in the Supplementary Materials of Ionides et al. (2006), and we include our implementation in the `genss` package (Falk et al., 2023). These standard error estimates are approximations and are known to be more liberal (i.e., smaller) than the true standard errors. We evaluate the performance of these slice likelihood standard errors in a simulation study below. Other methods for calculating standard errors when using particle filtering methods for estimation is an active area of research, and we discuss other potential options in our Discussion section.

## 3. A Simulation Study

With the estimation details of the GR state-space model established, we now describe a simulation study that serves two purposes: 1) to evaluate how MIF2 and slice-likelihood standard errors serve in fitting GR state-space models to data generated from a GR state-space model, and 2) to evaluate how a normal measurement state-space model (the *linear approximation* approach) performs when fit to data generated from a GR state-space model. Code to replicate the simulation study and empirical example are available at https://osf.io/fa69p/.

| Simulation factor | Values |
|---|---|
| Number of timepoints | 100, 500 |
| Number of items per state | 3, 6 |
| Number of responses per item | 3, 7 |
| AR parameter value | .3, .7 |
| CR parameter value | 0, .25 |
| GRM thresholds | Equal, Offset |

## 4. Methods

### 4.1. Data Generating Model

In all conditions, the data generating model was a GR state-space model with 2 states and a varying number of indicators per state. The number of states was fixed at two to avoid radically increasing the number of conditions and computational cost of fitting the data, while still being able to demonstrate the estimation of cross-regressive effects. The form of the $\mathbf{A}$ matrix is $\begin{bmatrix} AR & CR \\ 0 & AR \end{bmatrix}$ with $AR$ referring to the autoregressive effects, and $CR$ referring to the cross-regressive effect of the first state on the second state. The $\mathbf{\Sigma}$ used to generate the data is determined in accordance to the identification constraints described previously. In all cases, each observed GR variable is associated with a single state, which we assume is known when we fit the models to the simulated data. The $\alpha_i$ parameters are set to 1 and are considered fixed and known during estimation. The number of items per state varies by condition, but is the same for each of the two states. Finally, $\beta_{ij}$ values were determined in two ways. In the *equal* GRM threshold condition, $\beta_{ij}$ were calculated by centering the sequence $[1, 2, \ldots, j]$ on 0, so that, for example, an item with 7 response categories would have $\beta_{ij} = [-2.5, -1.5, -.5, .5, 1.5, 2.5]$. In this *equal* condition, the $\beta_{ij}$ sets for each observed variable are the same, corresponding to the situation where each observed variable has identical measurement properties. In the *offset* GRM threshold condition, the $\beta_{ij}$ sets for each item within a given state were offset from one another by either the number of items divided by the number of $\beta_{ij}$ per item, if that offset is less than 1.25, and 1.25 if it was calculated to be greater than 1.25. In the case of three items per state, with 7 response options per item, this results in the following set of $\beta_{ij}$, $\beta_{1j} = [-3, -2, -1, 0, 1, 2]$, $\beta_{2j} = [-2.5, -1.5, -.5, .5, 1.5, 2.5]$, $\beta_{3j} = [-2, -1, 0, 1, 2, 3]$. This condition corresponds with more typical measurement in psychometric scales, where the items were designed to cover a range of latent values.

Table 2 lists the simulation factors and values. All factors were fully crossed, which resulted in 64 conditions. A total of 300 iterations per condition were generated.

### 4.2. Analysis Models

All generated datasets were fit with a GR model and a linear approximation model. For both the GR model and linear approximation model, $\mathbf{A}$ was unconstrained and $\mathbf{\Sigma}$ was calculated using the identification constraint described above. Recall that this allows for direct comparison of the estimated values of $\mathbf{A}$ between the GR and linear approximation models. In both cases, the initial values for $\mathbf{A} = \begin{bmatrix} .1 & 0 \\ 0 & .1 \end{bmatrix}$. In the fit GR models, all $\beta_{ij}$ are freely estimated. The initial values for

$\beta_{i1}$ were set to $-2$, and the initial values for the intervals between subsequent $\beta_{ij}$'s were set to .36.

In the linear approximation models, the matrix $\mathbf{C}$ from Eq. 2 was constrained to freely estimate non-0 state-item loadings, corresponding to a known measurement structure with unknown parameter values. The initial values for nonzero loadings were set to 1. Initial values for diagonal elements of $\mathbf{\Psi}$ were set to 1.

Finally, 4 separate runs of the MIF2 algorithm were performed on for each model as per Ionides et al. (2015) and the estimated coefficients were averaged into a final estimate. Slice likelihood standard errors were calculated based on the averaged estimates. In all cases, 1000 particles were used with 250 MIF2 iterations. The cooling proportion per 50 iterations was set to.05 (i.e., after 50 iterations, the variance of the parameter perturbation distribution, $\mathbf{\Theta}$ is.05 of the starting values). For all parameters, the initial perturbation standard deviation was set to .3.

### 4.3. Outcomes

As our interests were chiefly in evaluating the GR models performance in recovering state dynamics and comparing its performance against the linear approximation model, our outcomes focus on the $\mathbf{A}$ matrix. While $\beta_{ij}$ parameters were estimated, we do not present outcomes related to them in this manuscript due to space concerns.

We present and discuss the following outcomes:

- **True-Estimated State Correlation**—We operationalize this outcome as the Spearman's correlation between the true state values, and the expected values of the states calculated via MIF2. This is a general measure of model performance, with values close to 1 indicating perfect recovery of the unobserved states, and values away from 1 indicating poor recovery.
- **Relative Bias of the Autoregressive Parameters**—This signed relative bias evaluates how well the models recover the autoregressive parameters. Values close to 0 indicate unbiased recovery, while nonzero values indicate bias.
- **Bias of the Cross-Regressive Parameter**—As one of our CR conditions set the parameter to 0, to examine the bias of the CR parameter we use bias rather than relative bias.
- **95% Wald Type Confidence Interval Coverage**—To evaluate how accurate our slice likelihood standard errors were, we calculated a simple 95% confidence interval (i.e., $\theta \pm 1.96 \times SE$) and determined the coverage. This was calculated for both the AR and CR parameters.

Additionally, we present plots and tables of slice standard errors in the Supplementary Materials.

## 5. Results

### 5.1. True-Estimated State Correlations

Figure 2 shows the true-estimated state correlations across all conditions. Median values per condition are contained in Supplementary Materials Tables S1 and S2. There are several notable findings here. First, in every condition the GR state-space model resulted in better overall recovery of the state values, with the maximum median recovery of $\rho_s = .849$ occurring in the 500 timepoint, AR: 0.7/CR: 0.25, equal thresholds, 6 items per state with 7 response categories. That the best recovery of the state values occurs in this condition is no surprise, given that it is the most generous data-quality wise. The median recovery for the linear approximation model for the same condition was $\rho_s = .749$. The worst recoveries for the GRM model were in the 100 timepoint, AR: 0.3/CR: 0 or.25, offset threshold, 3 items per state, 3 response categories per item
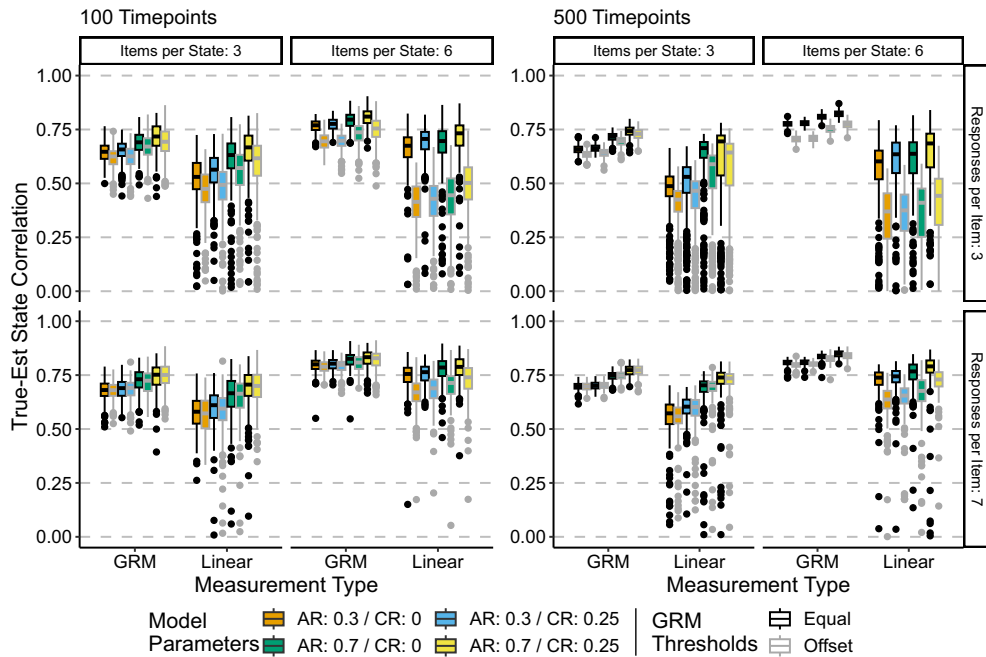
FIGURE 2.
True-estimated state Spearman's correlations. Central black line on boxplot denotes median, box denotes 25–75% interquartile range, whiskers denote $1.5 \times IQR$. AR refers to the autoregressive parameter, while CR refers to the cross-regressive parameter.

condition, with a median recovery of $\rho_s = .625$. The linear approximation model had median recoveries of $\rho_s = .474$ and $.491$. In addition to the GR models performing better than the linear approximation models with respect to overall level of state recovery, the GR models also had less variable recoveries across simulation iterations than the linear approximation models. The linear approximation models showed substantial negative skew in the true-estimated state correlations, suggesting that the linear approximation model struggled to recover the state values more often than the GR model.

In terms of the simulation factors effects on the state recovery, both GR and linear approximation models had the same general pattern of effects, with the notable exception of the effect of number of timepoints. In the case of the GR model, increasing the number of timepoints from 100 to 500 uniformly decreased the variability of the true-estimated state correlations within a condition and increased the overall true-estimated state recovery. However, for the linear approximation model in the 3 response categories per item conditions (top row of Fig. 2), increasing the number of timepoints from 100 to 500 led to an increase in variability and an overall slight decrease in the median level of recovery. This was not the case for the linear approximation models in the 7 response categories per item conditions (bottom row of Fig. 2), though both the reduction in variability and improvement in median recovery was modest compared to that of the GR model.

Increasing the number of items per state generally improved the recovery of the state values, as did increasing the number of response categories per state. Increasing the magnitude of the AR and/or CR parameters also generally increased the median recovery across all other conditions. Interestingly, having the GR threshold parameters (i.e., $\beta_{ij}$) be equal across items led to better overall recovery than having the GR threshold parameters be offset. Additionally, there appears to be an interaction between the offset of the GR thresholds and the number of items per state,
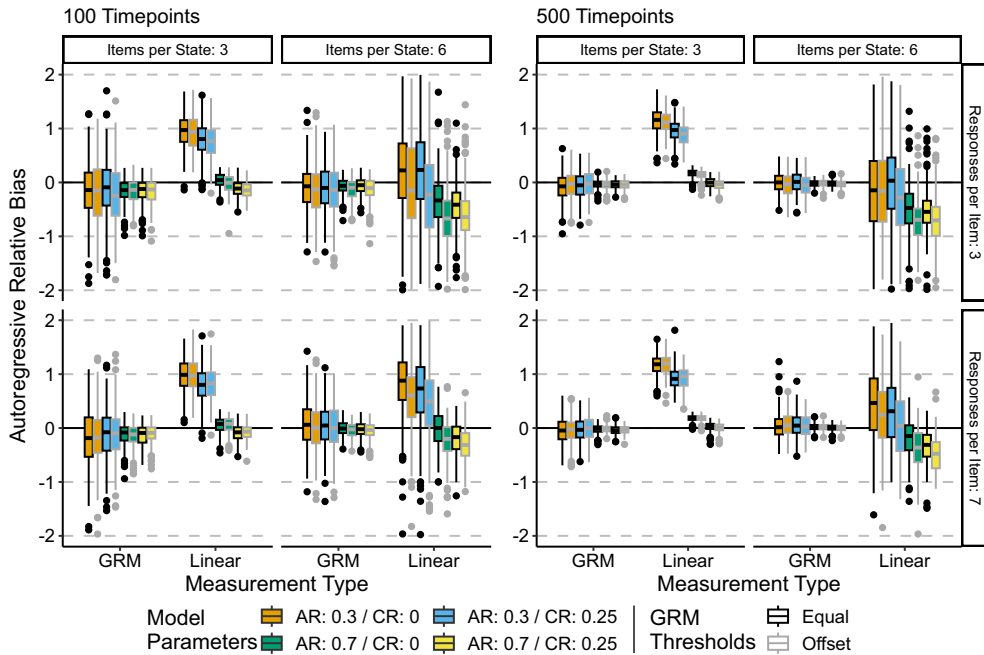
FIGURE 3.
Autoregressive parameter relative bias. Central black line on boxplot denotes median, box denotes 25–75% interquartile range, whiskers denote $1.5 \times IQR$. AR refers to the autoregressive parameter, while CR refers to the cross-regressive parameter.

where the reduction in recovery when GR threshold parameters were offset is greater with more items per state. This result is most noticeable in the linear approximation model, but occurs with lower magnitude in the GR models as well.

### 5.2. Relative Bias of the AR Parameter

Figure 3 visualizes the relative bias of the AR parameters. Tables S1 and S2 in the Supplementary Materials contain median relative bias values across all conditions.

First, the results shown in Fig. 3 suggest that the GR model results in asymptotically unbiased estimates of the AR coefficients. Across all conditions, the median relative bias is close to 0, and the variability of this bias decreases with increased numbers of timepoints. As one would expect, larger magnitudes of the AR coefficient result in lower relative bias.

However, the linear approximation model did not result in unbiased estimates of the AR parameters. Several patterns of findings stand out. First, when the AR parameters were small ($AR = 0.3$), the linear approximation model resulted in AR parameters that were positively biased, with the greatest median relative bias of approximately .75 to 1.25 occurring in the 3 items per state conditions. In the 6 items per state conditions, the variability of the relative bias was radically increased relative to the variability in the 3 items per state condition. For stronger AR effects ($AR = .7$), the linear approximation models resulted in relatively unbiased estimates only in the 3 items per state conditions, while for the 6 items per state conditions, there was a consistent negative bias for the AR coefficients. Notably, the variability of the bias was minimally impacted by increasing the number of timepoints from 100 to 500.
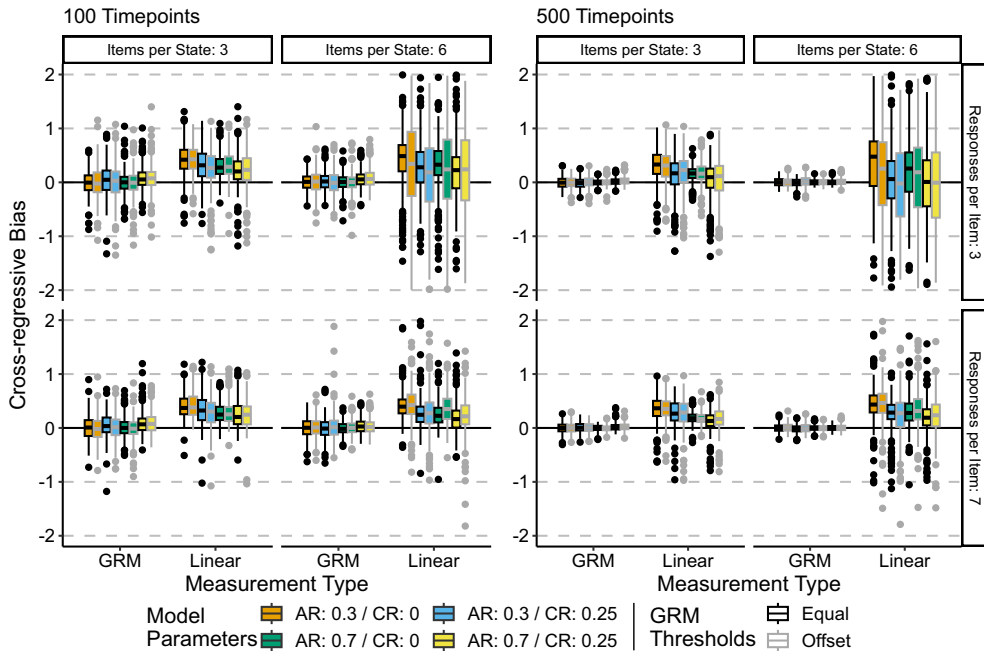
FIGURE 4.

Cross-regressive parameter bias. Central black line on boxplot denotes median, box denotes 25–75% interquartile range, and whiskers denote $1.5 \times IQR$. AR refers to the autoregressive parameter, while CR refers to the cross-regressive parameter. Note that this figure shows bias rather than relative bias.

### 5.3. Bias of the CR Parameter

Figure 4 shows the bias of the CR parameter across all conditions. Tables S3 and S4 in the Supplementary Materials contain the median values.

The results shown in Fig. 4 suggest that the GR model results in unbiased estimates of the CR parameter, with variance decreasing and median value becoming more 0 centered with an increase in the number of time points. For the linear approximation model, the bias of the CR parameter is consistently positive, and this bias does not improve with increased sample sizes. Consistent with the AR parameter, the 6 items per state/3 response categories per state conditions exhibited increased variability in the bias for the linear approximation models relative to the 3 items per state conditions. It appears that increasing the magnitude of the CR parameter does slightly improve bias; however, this difference is slight. Finally, there is a less pronounced effect of the GRM thresholds being equal vs. offset. The offset conditions appeared to result in slightly higher variance in the bias for the linear approximation models, but there is little difference in the median bias.

### 5.4. Parameter Coverage

Figure 5 shows the coverage of the 95% Wald type confidence interval using the slice likelihood SEs. Tables S1 and S2 in the Supplementary Materials list the coverage values per condition.

First, we note that across most conditions the coverage of the AR parameter in the GR models was consistently close to 75%, with the exception of the 500 timepoints, 6 items per state, 7 responses categories per state conditions, where the coverage was approximately 70%. There were minimal differences between coverage with respect to parameter values or GRM thresholds for the GR model. This is consistent with the GR models resulting in unbiased estimates of the
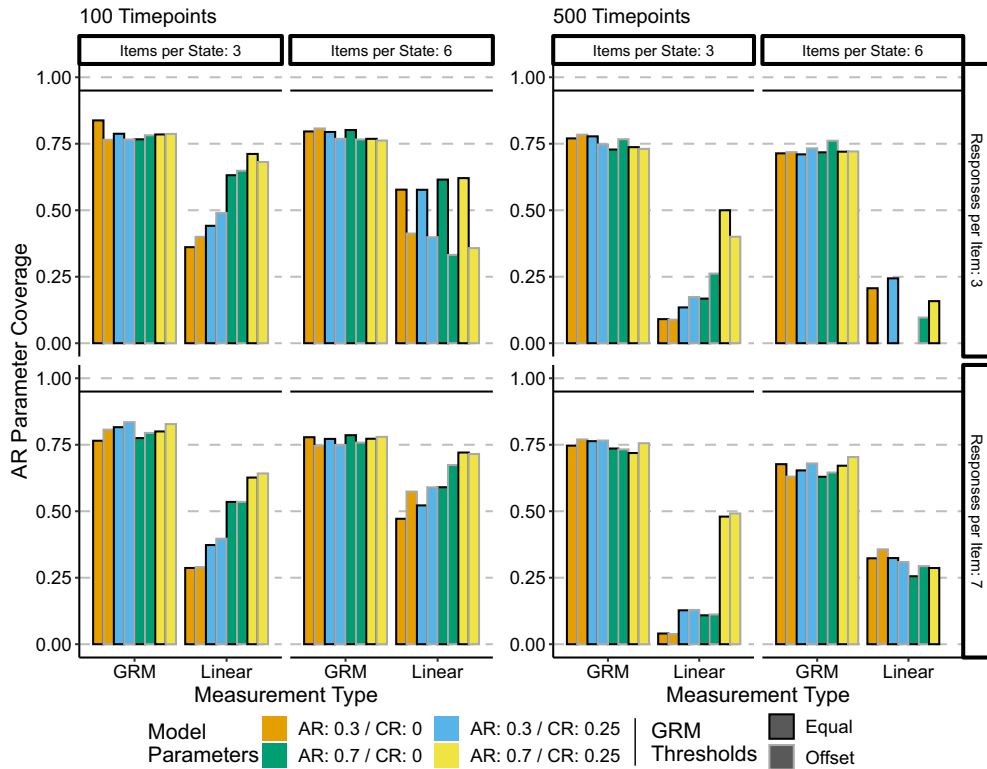
FIGURE 5.
Autoregressive parameter coverage. Height of bars represents coverage.

parameters, with the lower than nominal coverage consistent with the slice-likelihood SE being a liberal estimate.

On the other hand, the coverage of the AR parameter when using the linear approximation model was less than ideal. The coverage was not consistent across conditions, with stronger magnitude parameters having better coverage overall than smaller parameter conditions. These results are consistent with the patterns of bias shown by the linear approximation models, combined with the liberal nature of the slice likelihood standard errors.

Figure 6 shows the coverage of the CR parameter over all conditions. Tables S3 and S4 in the Supplementary Materials contain the values of coverage.

The coverage results for the CR parameter are consistent with the coverage results for the AR parameter. Overall the coverage of the GR models was approximately.75% with the exception of the 500 timepoint, 6 items per state, 7 responses per item conditions, where the coverage values range from approximately 55% to 70%. Again, this is consistent with unbiased parameter estimates combined with overly liberal standard errors.

Compared to their coverage of the AR parameter, the linear approximation models have considerably worse coverage of the CR parameter. Coverage is best when the CR parameter is nonzero; however, coverage decreases with increased sample size. This is consistent with the biased estimates that the linear approximation models provide.
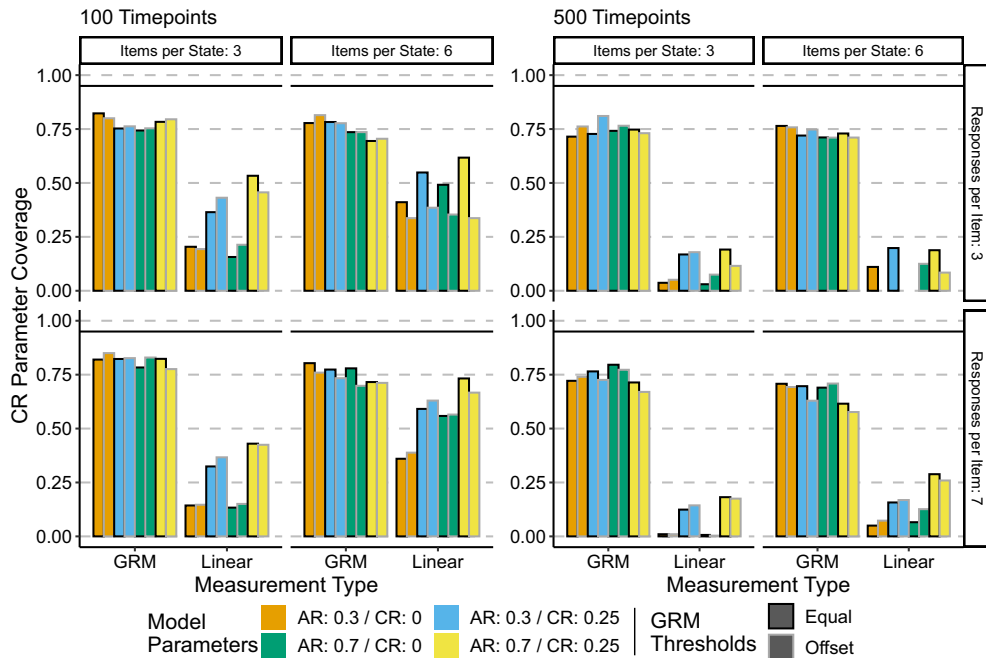
FIGURE 6.
Cross-regressive parameter coverage. Height of bars represents coverage.

## 5.5. Slice Likelihood Standard Errors

Figures S1 and S2 in the Supplementary Materials, as well as Tables S1, S2, S3, and S4, display and contain the slice likelihood standard error estimates. We briefly describe the overall pattern of findings here.

For both the AR and CR coefficient, two interesting patterns emerge. In the 3 items per state conditions, the median slice likelihood SEs of the linear approximation models are slightly less than that of the GR models, though this difference becomes negligible at 500 timepoints. Conversely, in the 6 items per state conditions, the linear approximation models have greater median slice likelihood SEs and more variability in the distribution of SEs than the GR model. This is reduced but not eliminated at 500 timepoints.

## 6. Empirical Example

Here, we apply the GR state-space approach as well as the linear approximation approach to the publically available dataset of Kossakowski et al. (2017), which consists of an EMA study of a single individual diagnosed with major depressive disorder. 1478 timepoints were collected over 239 consecutive days (with 10 timepoints a day collected) and during a specific phase of data collection the participant decreased their dosage of anti-depressants. In the following analyses, we examine the dynamics between three items measuring mood, and three items measuring self esteem. For the sake of simplicity, we did not take into account the phase of the study (i.e., before, during or after the reduction in antidepressants).

TABLE 3.
A priori estimated measurement parameters.

| Item | $\alpha$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | Loading |
|------|------|------|------|------|------|------|------|------|
| M1 | 1.888 | −0.708 | 0.607 | 1.42 | 2.075 | 3.244 | 4.345 | 0.803 |
| M2 | 1.324 | −5.367 | −4.835 | −2.068 | 1.2 | 2.713 | 3.607 | 0.315 |
| M3 | 4.349 | −3.306 | −2.856 | −0.387 | 1.047 | 2.221 | 3.134 | 0.679 |
| S1 | 3.397 | −3.185 | −3.092 | −1.979 | −0.607 | 3.173 | – | 0.409 |
| S2 | 2.99 | −3.661 | −3.025 | −2.515 | −1.959 | −1.174 | – | 0.34 |
| S3 | 3.462 | −3.123 | −2.388 | −0.596 | 0.7 | 3.251 | – | 0.605 |

$\alpha$ and $\beta$ are the discrimination parameter and threshold parameters for the graded response model, respectively. Loadings are the estimated factor loadings from the fit confirmatory factor model.

### 6.1. Measure Description, Preprocessing, and Model Choice

Out of the available items that measured mood and self-esteem, we chose three for each construct that showed the greatest range of response categories. For mood (for which items were measured on a 7 point Likert scale), we chose M1. "I feel relaxed" (reverse coded), M2. "I feel down" and M3. "I feel irritated." For self-esteem (also measured on a 7 point Likert scale), we chose S1. "I like myself," S2. "I am ashamed of myself" (reverse coded), and S3. "I can handle anything."

We initially attempted to fit GR state-space and linear approximation state-space models where the measurement parameters were estimated, as in the simulation study previously described. Due to empirical under-identification (very few observations at the extreme ends of the response options), we encountered substantial difficulties with convergence for both model types. As such, we chose to estimate measurement models separately using a multivariate graded response model using the R package `mirt` (Chalmers, 2012) and confirmatory factor model using the R package `lavaan` (Rosseel, 2012). Table 3 contains the estimated measurement parameters. In both cases, we set the mean of the latent variables to 0 and their variances to 1, and allowed correlation between the latent variables. The correlation between the latent variables was estimate to be −.803 for the multivariate graded response model and −.701 for the confirmatory factor model. Readers should note that these measurement parameters are estimated using the data from one individual for the purposes of this example, and should not be taken as accurate in the larger population.

Following the estimation of the measurement parameters, we fit the GR state-space model and linear approximation model, in both cases estimating the state dynamics matrix **A**. While not necessary from an identification standpoint, as we fixed the measurement parameters, we retained the constraint on marginal variance of the states (fixing the marginal variance to 1 by making the state innovation variances a function of the state dynamics) so as to make the comparison between the GR state-space model and the linear approximation model estimates as direct as possible. As with the simulation study, 4 separate MIF2 models were fit per model type, each with 1000 particles, 500 iterations, a cooling rate per 50 iterations of .75, and a starting proposal standard deviation of .05 for all estimated parameters. The cooling rate and proposal standard deviation were chosen via a small grid search, and resulted in the best overall convergence. After the 4 MIF2 runs were complete, the final parameter estimates were calculated as the average of the final iteration of the 4 runs, and slice likelihood SEs were obtained.

Finally, we quantified goodness of fit at the item level as follows: we first calculated the expected state trajectory for the GR state-space model and the linear approximation model using the parameters previously estimated (measurement and state-dynamics). We then simulated 1000

TABLE 4.
State dynamic parameter estimates.

| Parameter | GR Model Est | Slice SE | Linear Model Est | Slice SE |
|-----------|--------------|----------|------------------|----------|
| $\mathbf{A}_{11}$ | 0.199 | 0.023 | 0.168 | 0.018 |
| $\mathbf{A}_{12}$ | −0.310 | 0.018 | −0.339 | 0.018 |
| $\mathbf{A}_{21}$ | −0.190 | 0.014 | 0.071 | 0.024 |
| $\mathbf{A}_{22}$ | 0.651 | 0.015 | 0.505 | 0.024 |

measurement time series using the estimated measurement models and the expected state trajectories. We calculated the item specific mean squared error from the observed data for each of these 1000 simulated timeseries and calculated means and standard deviations across the 1000 iterations. This provides an item-level index of fit that is comparable across model types.

### 6.2. Results

Table 4 contains the estimated state dynamics parameters for both the GR state-space and linear approximation models. For reference, state 1 can be considered as the *mood* state, with higher values reflecting more negative mood, while state 2 can be considered as the *self-esteem* state, with higher values reflecting more self-esteem.

The estimated state dynamics were fairly consistent between the GR state-space model and the linear approximation model, with one notable exception being the cross-lagged effect of self-esteem on mood. In both models, the autoregressive effects ($\mathbf{A}_{11}$ and $\mathbf{A}_{22}$) suggest that mood ($\mathbf{A}_{11} = .199 \text{ and } .168$) returned to equilibrium faster than self-esteem ($\mathbf{A}_{11} = .651 \text{ and } .505$). The cross-lagged effect of mood on self-esteem ($\mathbf{A}_{12} = −.310 \text{ and } −.339$) was negative, suggesting an inverse relation between mood and self-esteem over time such that more negative mood at time $t$ was associated with less self-esteem at time $t + 1$. For both models, the slice standard errors for $\mathbf{A}_{11}, \mathbf{A}_{12}$, and $\mathbf{A}_{22}$ were small relative to the magnitude of the parameter estimate. Even with the more liberal nature of the slice standard errors, the relative difference in magnitude suggests that these effects would be considered statistically significant and unlikely due to random chance.

Where the GR state-space and the linear approximation models diverge is in the estimate of the cross-lagged effect of self-esteem on mood. In the GR model, this effect was negative ($\mathbf{A}_{21} = −.190$) with a small slice standard error, suggesting that there is a small inverse relation from self-esteem to mood over time, such that high self-esteem at time $t$ are related to lower negative mood at time $t + 1$. However, for the linear approximation model, the effect was positive and small ($\mathbf{A}_{21} = 0.071$) with a relatively large slice standard error of .024. Given the slice standard error's known underestimation of the uncertainty in the parameter estimate, this implies that there is little to no cross-lagged relation from self-esteem to mood when analyzed with the linear approximation approach. As the linear approximation approach is a priori known to be a miss-specified model relative to the GR state-space model, this in turn suggests that the linear approximation model is underestimating the strength of this relation and providing a false negative finding.

Goodness of fit measures are given in Table 5. These item level goodness of fit measures are consistent in magnitude across the GR state-space and linear approximation models, with the GR state-space model exhibiting better fit for five out of the six items. The linear approximation model showed better fit of item M3 ("I feel irritated"), with an MSE of .218 to the GR state-space model's MSE of .302. Taking these goodness of fit measures together with the estimated state

TABLE 5.
Model fit measures.

| Item | GR MSE | SD | Linear MSE | SD |
|------|--------|-----|------------|-----|
| M1 | 1.448 | 0.055 | 1.468 | 0.046 |
| M2 | 0.878 | 0.036 | 0.925 | 0.03 |
| M3 | 0.302 | 0.011 | 0.218 | 0.007 |
| S1 | 0.283 | 0.011 | 0.316 | 0.01 |
| S2 | 0.364 | 0.022 | 0.479 | 0.015 |
| S3 | 0.456 | 0.015 | 0.571 | 0.018 |

MSE refers to average mean squared error over the 1000 simulation iterations, and SD is the standard deviation of the MSE over the 1000 simulation iterations.

dynamics suggest that the GR state-space model results in a better fit to the observed data overall, and that this fit allows it to better detect relatively weak state dynamics.

## 7. Discussion

In this manuscript, we present an estimation strategy for state-space models with graded response measurements, suitable for use with ordinal measurements such as Likert scales. We demonstrate that the estimation routine results in unbiased parameter estimates, and compare the performance of our model with that of a linear approximation approach, consistent with fitting a dynamic structural equation model (in an $N = 1$ setting) to ordinal measurements. The findings of our simulation study suggest that the linear approximation approach results in biased estimates of the state dynamics, and should be used with considerable caution. Our empirical example, consisting of an analysis of the dynamics of mood and self- esteem across 1478 timepoints, showed that while the GR state-space and linear approximation models were broadly consistent in their findings, the linear approximation model underestimated the cross-lagged relation from self-esteem to mood and had overall worse fit than the GR state-space model.

While there are a number of methods for estimating state-space type models with ordinal outcomes, such as Bayesian approaches in DSEM (Asparouhov et al., 2018) or the work of van Rijn (van Rijn, 2008) using generalized linear modeling, the estimation approach we describe here is innovative in a number of ways. First, the identification constraint we develop here involves only the parameters for state dynamics, allowing comparable estimates to be obtained from models with differing measurement models. Second, the use of particle filtering, and more generally the MIF2 algorithm (Ionides et al., 2015), provides a general estimator that does not require the use of gradients that is suitable for any measurement model and/or any mix of measurement models, while both DSEM and the approach of van Rijn are restricted to specific classes of outcomes. While the estimation technique itself relies on simulation akin to standard methods in Bayesian estimation, the estimation is developed in a frequentist framework, requires less computation than full Bayesian estimation, and we provide a means of calculating approximate standard errors using slice-likelihoods. We emphasize that the MIF2 algorithm is not specific to ordinal measurement outcomes, and can be used to estimate models with any form of measurement (and for that matter state dynamics). Finally, we provide an open-source implementation of the GR and linear approximation models used in this manuscript in the R package genss (Falk et al., 2023).

Our simulation study demonstrates that the GR model results in unbiased estimates of the state-dynamics and that the slice-likelihood standard errors result in consistent coverage across conditions. However, this coverage was not at the nominal level of 95%, indicating that the slice-

likelihood standard errors are overly liberal. There are a number of approaches for obtaining accurate SE estimates from models that are estimated using a particle filtering approach such as profile likelihoods (Ionides et al., 2015) or particle filter approximations of the Hessian (Spall, 2005; Chada et al., 2022). However, these approaches are orders of magnitude more computationally expensive than slice-likelihood, requiring in the case of the profile likelihoods re-estimation of thousands of models with slight variations in fixed parameter values. Refining the standard error estimates for these models is of paramount interest, as accurate SEs are necessary for accurate inference. That being said, the slice-likelihood standard errors are fairly consistent in their coverage in the GR model across conditions, suggesting that one can alleviate these issues by using more strict Wald type confidence intervals. For example, we found that a 99.8% Wald CI (i.e., $\theta \pm 3.09 \times SE$) provided approximate 95% coverage in the GR models. As the slice-likelihood SEs are more liberal, conservative statistical test thresholds should be used to help reduce the chance of false positive findings. We advocate for the use of 99.8% confidence intervals and that authors explicitly note the use of the slice-likelihood SEs along with their limitations.

The linear approximation models performance in our simulation study revealed considerable issues with bias in both the AR and CR parameters, and that there was no set of conditions that completely alleviated the bias. Overall, it appeared that more response options per item (7 vs. 3), resulted in less variable estimates, but only minimally different estimates with respect to bias. This is somewhat consistent with the idea that an ordinal measurement with increasing number of response categories will converge to a continuous distribution (e.g., a 0–100 slider is more continuous than a 1–7 Likert scale), however the negligible improvement in bias suggests that the influence of increasing the number of response categories needs to be studied more. We also note that while increasing the number of timepoints improved bias for the GR model, this was not the case in the linear approximation models. Of course, this was not expected as increasing the sample size when fitting a mis-specified model does not make for less biased estimates, just more misplaced confidence in the biased estimates one obtains. Here, we suggest extreme caution is warranted in fitting continuous measurement state-space models to ordinal measurements, and strongly advocate for the use of an appropriate model such as the GR models presented here or other ordinal outcome state-space models discussed previously. We note that our findings of bias in the case of the linear approximation model are in line with previous work evaluating the performance of normal theory models applied to ordinal data. In an excellent and thorough study of this topic, Rhemtulla et al. (2012) comprehensively reviewed related literature which suggested that at around seven response categories, treating ordinal variables as continuous resulted in minimal, but still nonzero bias in parameter estimates. Rhemtulla et al. (2012) additionally showed that estimating confirmatory factor models on Likert type data resulted in biased parameter estimates and overly liberal standard errors particularly for measurement distributions that had small numbers of response categories and highly skewed measurement distributions. However, they did note that at seven response options the bias was minimal, though consistent, and more severe when the measurement distribution was skewed. While our simulations show fairly severe bias in estimating low magnitude state-dynamic parameters even with seven response categories, our findings are consistent in that large magnitude state-dynamic parameters had minimal bias when estimated with the linear approximation model. Our empirical example is also consistent with Rhemtulla et al. (2012), showing small but meaningful differences in fit and parameter estimates between the GR state-space and the linear approximation models. Taking our findings with those of Rhemtulla et al. (2012), we can synthesize the following note of caution: while it is certainly the case that ordinal variables with large numbers of response options could be treated as continuous and the consequences of that approximation on inference *can* be minimal, the linear approximation is still a miss-specified model with guaranteed bias, and it appears that the consequences of the approximation increase with model complexity. Again, we advocate for caution in using the linear

approximation state space models, and call for more research on robust estimation for state-space models with ordinal measurements.

There are a number of limitations to the modeling strategy presented here. First and foremost is the computational expense of the estimation. As the estimation routine is a simulation based method, and requires iterated particle filters, any model fit using MIF2 takes a considerable amount of time. In our simulations, it took approximately 15–20 min for a single model to fit, though that time heavily depends on the computational environment (15–20 min on a Linux desktop computer outfit for simulation studies, 20–30 min on a more typical Windows laptop). There is also the limitation of needing to determine appropriate setting for the estimation itself (i.e., number of particles, cooling rate, number of iterations and the standard deviation of the perturbation distribution), along with the need to fit multiple runs of the same model to the data to determine if a global maximum has been obtained. These limitations can be ameliorated by optimizing the implementation, but should be taken into account when considering using this estimator.

Another limitation of the approach we develop here is that the state-space models are $N = 1$, suitable for purely idiographic analysis. There are a couple of subtleties to this limitation that need to be discussed, as they are relevant for readers interested in fitting these models idiographically. First, because there is no sharing of information between multiple participant's models, both the measurement model and state dynamics will be specific to a given participant. However, state dynamics parameters can be compared between participants' models. Furthermore, as the expected values of the state distribution are assumed to be 0 as per the identification constraints, the estimated state trajectories are not comparable between participants unless measurement parameters were assumed fixed and known. If the $\beta_{ij}$ values were estimated per participant, the state trajectories are not directly comparable between participants, as the measurement parameters would contain information as to the location of the participants' state values relative to the population. This limitation can be alleviated by using a calibrated set of items with known parameters, and modifying the GR state-space model to estimate participant specific mean vectors of states. If the same set of measurement parameters is used for all subjects in addition to constraining the marginal variance of the states to 1, then the state trajectories will be comparable between participants (and suitable for the calculation of percentiles relative to the population).

Additionally, a limitation that arose when applying these models to empirical data was that tuning the estimation procedure to the vagaries of empirical data is a difficult and time consuming process. We encountered difficulty in getting both the GR state-space and linear approximation model to converge when we allowed the measurement components to be freely estimated, and we believe that this difficulty has to do with the nature of the MIF2 algorithm combined with empirical under-identification issues. While the MIF2 algorithm is guaranteed to eventually converge to the MLE, the rate of convergence is unknown and highly dependent on the tuning parameters and the specifics of the analyzed data. Carefully tuning the performance of this estimator is possible, but would have to be performed for every dataset and model structure separately, with few to no general guidelines possible to recommend. As such, this estimation approach and therefore the GR state-space model are not turn-key analysis methods and require a substantial amount of effort and expertise to implement correctly.

The limitations of the current model suggest a number of fruitful future directions. First, a means of analyzing multiple participants is needed. The estimation method is not suitable for analyzing multiple participants' data simultaneously in the same model due to computational expense, so we are investigating the use of meta-analytic techniques for analyzing the participant-specific parameter estimates. This approach would preserve the idiographic nature of the models while allowing for an analysis of population heterogeneity/homogeneity. Second, as the MIF2 algorithm is a general estimator that allows for arbitrary measurement models, other measurement models such as distributions for counts, zero-inflated distributions, and other IRT measurement

models, should be implemented and their performance evaluated. A low hanging fruit in this direction is that of zero-inflated distributions, which would improve our ability to model time series of rarer behaviors like substance use. Finally, the MIF2 algorithm itself should be studied to improve its speed and convergence. One core component of the MIF2 algorithm is the use of a symmetric perturbation distribution, which could be improved via integrating gradient information ala Hamiltonian MCMC (Neal, 1996) or by sharing information between particles ala a genetic optimization algorithm (de Lima & Krohling, 2011). Improving the convergence of the MIF2 approach would relieve much of the burden on analysts to carefully tune the estimator, making these models more accessible to applied researchers. Finally, improved standard error estimates must be developed, potentially using the profile likelihood or particle filter estimation of the Hessian methods mentioned previously.

The analysis of intensive longitudinal data in psychology is only becoming more prominent, and researchers need analysis techniques that can appropriately account for the measurement properties of their data. We aim that the work presented in this manuscript provides a general estimation infrastructure for state-space modeling of ordinal psychological data and look forward to expanding its capabilities in the near future.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 253*, 359–388. https://doi.org/10.1080/10705511.2017.1406803

Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques* (1st ed.). London: CRC Press. https://doi.org/10.1201/9781482276725

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112518*, 859–877. https://doi.org/10.1080/01621459.2017.1285773

Browne, M.W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of autoregressive moving average time series models. In *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 415–452). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Chada, N. K., Jasra, A., & Yu, F. (2022). Unbiased estimation of the hessian for partially observed diffusions. *Proceedings of the Mathematical, Physical, and Engineering Sciences*. https://doi.org/10.1098/rspa.2021.0710

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 486*, 1–29. https://doi.org/10.18637/jss.v048.i06

Chaubert, F., Mortier, F., & Saint André, L. (2008). Multivariate dynamic model for ordinal outcomes. *Journal of Multivariate Analysis, 998*, 1717–1732. https://doi.org/10.1016/j.jmva.2008.01.011

Chow, S.-M., Hamaker, E. L. J., Fujita, F., & Boker, S. M. (2009). Representing time-varying cyclic dynamics using multiple-subject state-space models. *British Journal of Mathematical and Statistical Psychology, 62*, 683–716. https://doi.org/10.1348/000711008X384080

de Lima, L. M., & Krohling, R. A. (2011). Particle filter with differential evolution for trajectory tracking. In A. Gaspar-Cunha, R. Takahashi, G. Schaefer, & L. Costa (Eds.), *Soft computing in industrial applications* (pp. 209–219). Berlin: Springer. https://doi.org/10.1007/978-3-642-20505-7_18

Ditzen, B., Hoppmann, C., & Klumb, P. (2008). Positive couple interactions and daily cortisol: On the stress-protecting role of intimacy. *Psychosomatic Medicine, 708*, 883–889. https://doi.org/10.1097/PSY.0b013e318185c4fc

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods (No. 38)* (2nd ed.). Oxford: Oxford University Press.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models (Softcover version of original hardcover edition 2001 ed.)* (2nd ed.). New York: Springer.

Falk, A., Slipetz, L., & Henry, T. (2023). netlabUVA/genss: genss v0.1.0 - Initial Prerelease. Zenodo. https://doi.org/10.5281/zenodo.7887019

Fisher, Z. F., Chow, S.-M., Molenaar, P. C. M., Fredrickson, B. L., Pipiras, V., & Gates, K. M. (2022). A Square-Root Second-Order Extended Kalman Filtering Approach for Estimating Smoothly Time-Varying Parameters. *Multivariate Behavioral Research, 571*, 134–152. https://doi.org/10.1080/00273171.2020.1815513

Ionides, E. L., Bretó, C., & King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences, 10349*, 18438–18443. https://doi.org/10.1073/pnas.0603181103

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., & King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences, 1123*, 719–724. https://doi.org/10.1073/pnas.1410597112

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering, 821*, 35–45. https://doi.org/10.1115/1.3662552

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering, 831*, 95–108. https://doi.org/10.1115/1.3658902

Kim, C.-J., & Nelson, C. R. (2017). State-space models with regime switching: Classical and gibbs-sampling approaches with applications. *The MIT Press*. https://doi.org/10.7551/mitpress/6444.001.0001

King, A.A., Ionides, E.L., Bretó, C., Ellner, S.P., Ferrari, M.J., Funk, S., & Wood, S. N. (2023). pomp: Statistical inference for partially observed Markov processes.

King, A. A., Nguyen, D., & Ionides, E. L. (2016). Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v069.i12

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 2204598*, 671–680. https://doi.org/10.1126/science.220.4598.671

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association, 82400*, 1032–1041. https://doi.org/10.2307/2289375

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics, 51*, 1–25. https://doi.org/10.2307/1390750

Kossakowski, J. J., Groot, P. C., Haslbeck, J. M. B., Borsboom, D., & Wichers, M. (2017). Data from 'Critical slowing down as a personalized early warning signal for depression'. *Journal of Open Psychology Data, 5*, 1. https://doi.org/10.5334/jopd.29

Le Cam, L., & Lo Yang, G. (1990). Locally asymptotically normal families. In *Asymptotics in statistics* (pp. 52–98). Springer US. Series Title: Springer Series in Statistics https://doi.org/10.1007/978-1-4684-0377-0_5

Levinson, C. A., Williams, B. M., Christian, C., Hunt, R. A., Keshishian, A. C., Brosof, L. C., & Ralph-Nearman, C. (2023). Personalizing eating disorder treatment using idiographic models: An open series trial. *Journal of Consulting and Clinical Psychology, 91*, 14–28. https://doi.org/10.1037/ccp0000785

Liddell, T., & Krushke, J. (2017). *Analyzing ordinal data with metric models: What could possibly go wrong?* (Technical Report). SSRN. https://doi.org/10.2139/ssrn.2692323

McGee, L.A., & Schmidt, S. F. (1985). *Discovery of the Kalman filter as a practical tool for aerospace and industry* (Technical Report No. 86847). Ames Research CenterNASA.

Moral, P. D., & Doucet, A. (2014). Particle methods: An introduction with applications. *ESAIM: Proceedings, 44*, 1–46. https://doi.org/10.1051/proc/201444001

Mote, J., & Fulford, D. (2020). Ecological momentary assessment of everyday social experiences of people with schizophrenia: A systematic review. *Schizophrenia Research, 216*, 56–68. https://doi.org/10.1016/j.schres.2019.10.021

Muthén, L.K., & Muthén, B. O. (1998). Mplus user's guide (8th ed.).

Neal, R.M. (1996). Monte Carlo implementation. In *Bayesian learning for neural networks* (Vol. 118, pp. 55–98). Springer New York. Series Title: Lecture Notes in Statistics https://doi.org/10.1007/978-1-4612-0745-0_3

Oud, J. H., Jansen, R. A. R. G., van Leeuwe, J. F. J., Aarnoutse, C. A. J., & Voeten, M. J. M. (1999). Monitoring pupil development by means of the Kalman filter and smoother based upon sem state space modeling. *Learning and Individual Differences, 112*, 121–136. https://doi.org/10.1016/S1041-6080(00)80001-1

Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika, 652*, 199–215. https://doi.org/10.1007/BF02294374

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS ®): Depression, anxiety, and anger. *Assessment, 183*, 263–283. https://doi.org/10.1177/1073191111411667

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 173*, 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 482*, 1–36. https://doi.org/10.18637/jss.v048.i02

Samejima, F. (1997). Graded response model. (Vol. 35, pp. 85–100). https://doi.org/10.1007/s10464-005-3397-z

Saulnier, K. G., Saulnier, S. J., & Allan, N. P. (2022). Cognitive risk factors and the experience of acute anxiety following social stressors: An ecological momentary assessment study. *Journal of Anxiety Disorders, 88*, 102571. https://doi.org/10.1016/j.janxdis.2022.102571

Seidman, A. J., George, C. J., & Kovacs, M. (2022). Ecological momentary assessment of affect in depression-prone and control samples: Survey compliance and affective yield. *Journal of Affective Disorders, 311*, 63–68. https://doi.org/10.1016/j.jad.2022.05.015

Song, H., & Ferrer, E. (2009). State-space modeling of dynamic psychological processes via the Kalman smoother algorithm: Rationale, finite sample properties, and applications. *Structural Equation Modeling, 162*, 338–363. https://doi.org/10.1080/10705510902751432

Sorenson, H. W. (Ed.). (1985). *Kalman filtering, theory and application*. New York: IEEE Press.

Spall, J. C. (2005). Monte Carlo computation of the fisher information matrix in nonstandard settings. *Journal of Computational and Graphical Statistics, 144*, 889–909. https://doi.org/10.1198/106186005X78800

Stumpp, N. E., Southward, M. W., & Sauer-Zavala, S. (2023). Do you see what i see? researcher-participant agreement on single-item measures of emotion regulation behaviors in borderline personality disorder. *Assessment, 301*, 102–110. https://doi.org/10.1177/10731911211044216

van Rijn, P. (2008). *Categorical time series in psychological measurement* (Dissertation). University of Amsterdam.

Vatiwutipong, P., & Phewchean, N. (2019). Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process. *Advances in Difference Equations, 20191*, 276. https://doi.org/10.1186/s13662-019-2214-1

Wan, E., & Van Der Merwe, R. (2000). The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium* (cat. no.00ex373) (pp. 153–158). IEEE. https://doi.org/10.1109/ASSPCC.2000.882463

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics, 71*, 126–153.

Wright, A. C., Browne, J., Skiest, H., Bhiku, K., Baker, J. T., & Cather, C. (2021). The relationship between conventional clinical assessments and momentary assessments of symptoms and functioning in schizophrenia spectrum disorders: A systematic review. *Schizophrenia Research, 232*, 11–27. https://doi.org/10.1016/j.schres.2021.04.010