ESTIMATION APPROACHES FOR GENERALIZED LINEAR FACTOR ANALYSIS MODELS WITH SPARSE INDICATORS

Sierra A. Bainter

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology and Neuroscience.

Chapel Hill
2016

Approved by:

Patrick Curran

Daniel Bauer

Kenneth Bollen

Amy Herring

Andrea Hussong

David Thissen

ABSTRACT

Sierra A. Bainter: Estimation Approaches for Generalized Linear Factor Analysis Models with Sparse Indicators
(Under the direction of Patrick J. Curran)

Substance use research involves a number of methodological challenges that require advanced data analysis techniques. Generalized linear factor analysis (GLFA) is a general latent variable modeling framework useful for substance use research that can be applied to continuous or categorical measures. Unfortunately, substance use data is characterized by a large proportion of zeros (sparseness), and sparse endorsement can cause maximum likelihood estimation of GLFA models to fail. However the extent of estimation problems caused by sparseness has not previously been well studied. Because of the great need to improve reliability for estimating models with items with low endorsement, in this study I evaluated Bayesian estimation as an alternative to maximum likelihood estimation for GLFA models with sparse, categorical indicators. I found that the use of priors in Bayesian estimation eliminated extreme parameter estimates, improved estimate efficiency, increased empirical power to detect true effects, and provided meaningful results when models do not converge using ML estimation. I also found that the gains in efficiency and empirical power using Bayesian estimation depend on specifying adequately concentrated priors (i.e. adequate information to constrain inferences), and the increased overall efficiency and empirical power were also tied to a trade-off with overall unbiasedness. In sum, my proposal to use Bayesian estimation with prior information to estimate GLFA models with sparse indicators provides a much needed alternative for substance use researchers who wish to make inferences with sparse data.

To Grandfather Paul, the first quantitative psychologist in the family.

ACKNOWLEDGEMENTS

This project was funded by NIDA (Grant F31 DA035523). I would like to thank the members of the L.L. Thurstone lab, both faculty and graduate students, for their collective mentorship and creating an encouraging academic environment. I am grateful to my mentor, Dr. Patrick Curran, for guidance and for always looking out for my best interests. I am very fortunate to have had such a stellar committee, the likes of which could not have been assembled at any other university. Finally, I am grateful to my husband Matt and daughter Hazel for helping me keep it all in perspective.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Research aimed at understanding the developmental factors of substance use and addiction is characterized by a number of methodological challenges. Specifically, a developmental investigation demands a longitudinal approach to separate causes from consequences of substance use, substance use outcomes are categorical, measures may have different meanings at different ages as age norms change, and it is important to consider influences from multiple levels (e.g. family and peer contexts, biological risk) which may operate over different time intervals (i.e. early versus proximal influences) and which may also change over time (Chassin, Presson, Lee, & Macy, 2013). All of these important considerations create demands for complex data collection and analysis, and many sophisticated statistical approaches have been developed for these problems involving specialized statistical models (e.g. Bauer et al., 2013; Bauer & Hussong, 2009; MacKinnon & Fairchild, 2009).

Additionally, studying the development of substance use requires collecting data on individuals before outcomes develop, and it is well known that substance use data is characterized by a large proportion of zeros, or non-users. For example, cocaine use among 8th graders is rare, below 2% (Johnston, O'Malley, Miech, Bachman, & Schulenberg, 2015), and even in large samples endorsement will be sparse —defined here as low endorsement frequencies for individual items or categories. Yet, researchers cannot completely avoid research with sparse items because it is important to study cases such as twelve-year-olds using drugs. Research in psychology is notoriously underpowered in general (Maxwell, 2004), and this problem of low statistical power is compounded in substance use research by the additional

challenges of studying rare behaviors and the need for complex data analysis techniques (Curran & Hussong, 2009).

An enduring problem for substance use researchers is collecting a sample large enough to observe sufficient numbers of cases of rare behaviors, such as early alcohol involvement or use of illicit drugs besides marijuana (Chassin, Presson, Lee, & Macy, 2013). This need has encouraged data sharing and spurred the development of approaches to simultaneously analyze data from independent studies (Hussong, Curran, & Bauer, 2013). However, what constitutes a sample that is "large enough" depends on the requirements of the appropriate statistical analysis technique. Given that sparseness is a significant issue in substance use research, lack of statistical procedures appropriate for sparse data substantially limits the inferences that can be made by substance use researchers.

Generalized linear factor analysis (GLFA; Bartholomew, Knott, & Moustaki, 2011; Skrondal & Rabe-Hesketh, 2004) is a broad class of models useful for research in the development of substance use disorders, encompassing traditional factor analysis and item response theory models. GLFA models may be recast as growth curve models to analyze longitudinal data or embedded in more comprehensive structural equation models (e.g. moderated nonlinear factor analysis, Bauer & Hussong, 2009). Although theoretically useful for addressing research questions related to substance use, a number of simulation studies have found that common estimation approaches for GLFA models – maximum likelihood and limited-information approaches – perform poorly in conditions that are characterized by sparseness (Forero & Maydeu-Olivares, 2009; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Olsson, 1979; Muthén et al., 1997, Rhemtulla, 2012). This is because the desirable properties of currently-available estimators are based on large-sample theory, which necessarily breaks down when observations are limited (Wasserman, 2005). For GLFA models with sparse categorical

indicators (e.g., substance use), sparseness depends not only on overall sample size but also endorsement frequencies for individual items and individual categories.

Though less familiar to social science researchers, GLFA estimation can also be approached from a Bayesian framework. In comparison to current estimation approaches, Bayesian estimation does not necessarily rely on large-sample theory and has a number of potential advantages for limited-data settings (Gelman et al., 2013). The potential advantages to Bayesian estimation are counterbalanced by a number of computational challenges and some aspects of Bayesian estimation, especially the specification of prior distributions, are subject to controversy. Though not previously studied for sparseness in GLFA models with categorical indicators, theory suggests that Bayesian estimation may be a beneficial alternative for GLFA when available estimation approaches break down.

Because of the great need to improve reliability for estimating models with sparse item responses, for my dissertation I investigated Bayesian estimation as an alternative estimation approach for GLFA models with sparse categorical indicators. In the next sections I will present the generalized linear factor analysis model for categorical indicators, survey traditional estimation approaches as well as Bayesian estimation methods, and discuss how sparseness is expected to influence each estimator. Next I will introduce the methods, design, and results for my dissertation project. I close by discussing implications of this work and future directions.

Generalized Linear Factor Analysis

In this section I review generalized linear factor analysis (GLFA; Skrondal & Rabe-Hesketh, 2004), a general psychometric modeling framework that is well-suited for research questions related to substance use. I present this general framework because it unifies two common psychometric modeling techniques: factor analysis and item response theory.

Historically, factor analysis (FA) was developed to explain dependence among continuous items[1] (e.g. scores on a battery of ability tests) by positing that they arise from one or more unobserved latent factors. Similarly, item response theory (IRT) was historically motivated to measure (unidimensional) latent ability from categorical test items. The historical distinction between FA and IRT has gradually blurred as FA has been extended to categorical items and IRT has been expanded to multiple latent dimensions, and both models can be derived as special cases of GLFA which is appropriate for categorical or continuous indicators.

Using notation adopted from the generalized linear model (McCullagh & Nelder, 1989), a univariate GLFA model for responses to item *i* for respondent *j* (y_{ij}) consists of three components: (1) a linear predictor (μ_{ij}), (2) a conditional response distribution, and (3) a link function *g* to relate the linear predictor to the probability of response.

For continuous y_{ij} the linear predictor is defined as an item intercept ν_i plus a factor loading λ_i expressing the regression of item *i* on the continuous latent factor η_j

$$\begin{aligned} y_{ij} \mid \eta_j &\sim N(\mu_{ij}, \sigma_i^2) \\ \mu_{ij} &= \nu_i + \lambda_i \eta_j \\ \eta_j &\sim N(\alpha, \psi) \end{aligned} \quad . \qquad 1$$

Here the response distribution of y_{ij} conditioned on η_j is univariate normal, and σ_i^2 is the item-specific residual variance. Specifying uncorrelated item-specific residual variances leads to the important property that the indicators are assumed independent, conditioned on the latent factor. For continuous indicators the identity link function is used, $g(\mu_{ij}) = \mu_{ij}$, which directly relates the linear predictor to the conditional response distribution for each item *i*.

[1] Note that I use the terms "item" and "indicator" interchangeably.

Because the factor scores η_j are unobserved they are modeled as randomly varying over individuals, and in order for the parameters in this model to be identified, restrictions must be imposed on α, ψ, and λ_i. The model is usually scaled either by fixing one item loading per factor to 1 and its intercept to 0, or by setting the mean and variance of the latent factor to 0 and 1, respectively.

Without adapting the conditional response distribution and link function, applying the GLFA model to categorical indicators creates an automatic misspecification − the categorical responses cannot be linear functions of the continuous factors. Ignoring the categorical nature of the data results in biased estimates, standard errors, and fit statistics (e.g., Dolan, 1994). As the number of categories increases, categorical variables approach continuity and bias generally decreases (Dolan, 1994; Rhemtulla et al., 2012). However in general with categorical data having four categories or fewer, continuous modeling strategies are not an optimal choice (Dolan, 1994; Rhemtulla et al., 2012).

In order to model categorical responses, the GLFA model is adapted in two important ways. First, a normal conditional response distribution is no longer appropriate. For binary items (e.g. yes/no or true/false item responses) a Bernoulli response distribution can be specified for each item as

$$y_{ij} \mid \eta_j \sim Ber(\mu_{ij}). \qquad 2$$

Further, the identity link function is no longer appropriate for relating the linear predictor to the expected value of y_{ij}. Instead, a function is needed to map the range of the linear predictor (-∞ to ∞) onto the permissible range for the conditional response distribution, which can only take on the values [0, 1]. One natural choice is the logit (inverse logistic) function, defined as

$$g(\mu_{ij}) = \ln\left[\mu_{ij}(1-\mu_{ij})^{-1}\right] \qquad 3$$

and plotted in Figure 1. Also plotted in Figure 1 is the probit function, which can be scaled to form a similar curve and is derived from the inverse cumulative distribution function of the normal distribution. Because the distributions align so closely, the choice of function usually depends on convenience. In this case, choosing the logit link for the GLFA yields

$$\nu_i + \lambda_i \eta_j = \ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right), \qquad 4$$

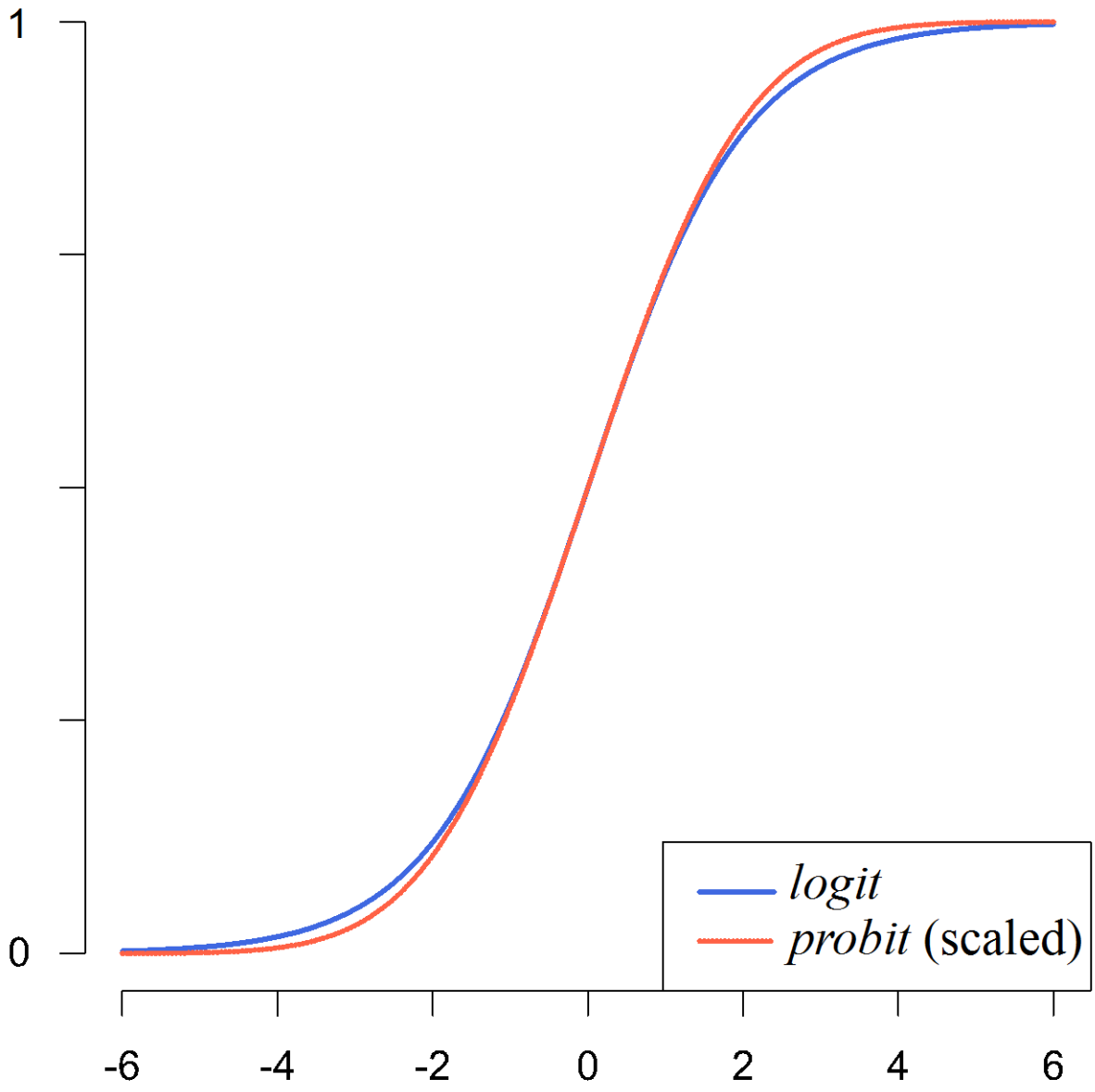
and η_j is specified as in Equation 1. Equivalently this model is expressed as

$$\mu_{ij} = \frac{1}{1+\exp[-(\nu_i + \lambda_i \eta_j)]} \qquad 5$$

which is an alternative expression of the well-known 2PL IRT model (Takane & de Leeuw, 1987). Using this specification for binary indicators, uncertainty is modeled only through the response distribution (rather than through residual variances). This completes the model specification for binary indicators, and as before, the model implies that all indicators are mutually independent given scores on η_j.

The preceding specification follows a factor analysis parameterization for the item parameters. Equivalently, each item parameter can be expressed using an IRT parameterization. Whereas the IRT is a nonlinear model for probabilities and estimates item difficulty and discrimination, the GLFA is a linear model for the logit (or probit) and estimates item threshold parameters and factor loadings. Parameters in the IRT and GLFA models can be directly transformed from one parameterization to the other; only their interpretations differ (see Wirth & Edwards, 2007 for conversion formulas). In the next section I discuss estimation approaches that have been developed for GLFA and consider challenges caused by sparseness.

Figure 1. Cumulative density functions for logit and scaled probit link functions



GLFA Estimation and Challenges with Sparseness

There are two families of traditional estimation approaches for GLFA models that include categorical indicators. These are limited-information estimators (e.g., modified weighted least squares methods, Jöreskog & Sörbom, 2001; Muthén, du Toit, & Spisic, 1997; polychoric instrumental variable estimator, Bollen & Maydeu-Olivarez, 2007) and full-information maximum likelihood (ML) estimation. For many properly specified models with moderately large samples, differences between estimators should be negligible (Forero & Maydeu-Olivares, 2009); however there are a number of key strengths and weaknesses to each approach. Limited-information estimators are computationally faster than ML, especially for models with multiple correlated latent variables, and have well-established tests for model fit (Wirth & Edwards, 2007; Forero & Maydeu-Olivares, 2009). Some limited-information approaches may also be less sensitive to mild misspecification (Bollen & Maydeu-Olivarez, 2007). For these reasons, limited-information estimation methods are in widespread use (e.g. the default estimator in Mplus, WLSMV, is limited-information, see Muthen & Muthen, 2014).

While more computationally challenging, ML estimation is statistically preferable to limited-information approaches for the problem of sparse endorsement because limited-information estimators are sensitive to bivariate sparse frequencies[2] whereas ML estimation is sensitive to univariate (item-level) sparse frequencies (Wirth & Edwards, 2007). Previous research using simulation studies has shown that ML performs better than limited information approaches in conditions characterized by sparseness (Forero & Maydeu-Olivares, 2009). For this reason, I limit my focus to compare ML to Bayesian estimation in this project.

[2] Specifically, the polychoric correlation coefficients that are used in limited-information estimation approaches are sensitive to low frequencies in bivariate contingency tables (Olsson, 1979; Savalei, 2011).

Maximum Likelihood Estimation

Maximum likelihood is a natural estimator choice for GLFA because of its well-known properties; ML estimation is both asymptotically efficient and consistent for correctly specified models under weak regularity conditions – mainly that the true values do not lie at the boundary of the parameter space and that the number of parameters does not increase with sample size (see e.g. Skrondal & Rabe-Hesketh, 2004, Ch. 6).

The likelihood function following from the GLFA model specification, marginalized over the latent scores η_j, can be written as

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{N}\left[\int \pi(\eta_j \mid \boldsymbol{\varphi}) \prod_{i=1}^{P} f_i\left(y_{ij} \mid \eta_j, \boldsymbol{\varpi}_i\right) d\eta_j\right] \qquad 6$$

where the vector $\boldsymbol{\theta}$ contains model parameters ($\boldsymbol{\varphi}$ and $\boldsymbol{\varpi}$), $\boldsymbol{\varphi}$ contains parameters governing the univariate normal distribution of η_j, and $\boldsymbol{\varpi}_i$ denotes parameters for the conditional response distribution f_i for each item *i*.

For continuous y_{ij}, the response distribution for each item f_i is normal, which results in a simplified individual log-likelihood function which is relatively computationally simple, and maximum likelihood estimation can be carried out using well-established algorithms to minimize the log-likelihood function, notably the expectation-minimization (EM) algorithm (Dempster, Laird, & Rubin, 1977).

However, the model specification for categorical y_{ij} does not lead to a simplified likelihood function. No analytic solution exists to integrate the likelihood in Equation 6 over η_j, and approximations to the integral must be obtained instead. For categorical indicators, the model is estimated by finding the observed proportions of each full pattern of item responses and estimating a multinomial distribution for the probability of each response pattern governed by

the item parameters (ν_i and λ_i). ML estimation for GLFA with categorical indicators was originally introduced in the IRT framework by Bock and Lieberman (1970) and reformulated by Bock and Aitkin (1981), employing a strategy equivalent to the EM algorithm.

ML estimation for models with categorical indicators requires integration over a *P*-dimensional distribution, where *P* is the number of latent factors or traits. This integral is generally approximated using quadrature techniques. Quadrature-based integration in its simplest form essentially estimates the area under the curve using a series of rectangles (or trapezoids), making the integral much easier to compute. Besides rectangular numerical integration, Gauss-Hermite integration is another common approach to approximation which uses weighted polynomials between points. The number of points used to approximate the area for each dimension is termed the number of quadrature points; these may or may not be equally spaced. These algorithms may be constructed as adaptive, determining optimal placing for each quadrature point, or quadrature points may be fixed. Because quadrature-based integration is needed for each latent dimension, the total number of quadrature points increases exponentially with the number of factors (Wirth & Edwards, 2007). This computationally intensive integration has to be repeated at each iteration of the EM algorithm. Common defaults for the algorithm and number of quadrature points per dimension vary, for example rectangular numerical integration with 15 adaptive quadrature points (in Mplus; Muthén & Muthén , 2014) or Gauss-Hermite integration with 49 fixed quadrature points (in FlexMIRT; Houts & Cai, 2013).

Some promising developments have recently been introduced to reduce the computational complexity of ML for models with multiple factors. Markov chain Monte Carlo (MCMC) algorithms can be used to assist the integration (see Wirth & Edwards, 2007; Cai, 2010b). MCMC techniques are widely applied in Bayesian statistics to simulate the posterior distribution and will be discussed in more detail in the next section; however MCMC can also be

utilized as an integration aid in the traditional (frequentist) statistical framework. Another exciting development by Cai (2010a), termed the two-tier item factor analysis model, is a dimension-reduction reformulation technique for some models which significantly reduces the computational burden.

Model Fit. Assuming a model is estimable, researchers must also be able to evaluate the usefulness of a model. There are many ways to evaluate a model's usefulness. For GLFA, one strategy is to try to assess how closely a model fits or reproduces the observed data. Much work assessing model fit is based on comparing the deviance in the loglikelihood to the deviance in a saturated model with all means, variances, and covariances freely estimated. For continuous, normally distributed indicators, the difference between the deviances of these two models, $\hat{F}$, can be used to form the likelihood ratio test statistic as

$$T = (N-1)\hat{F} \qquad 7$$

where N is sample size. For large samples and properly specified models, T has a central chi-square distribution with degrees of freedom equal to the difference between the number of parameters in the saturated and hypothesized models. For misspecified models, T follows a non-central chi-square distribution with non-centrality parameter λ. This statistic is often used as the basis of testing absolute goodness of fit (i.e., does the model fit the data?) and relative goodness of fit for comparing nested models (i.e., does model A fit worse than model B?). Other goodness-of-fit statistics such as the RMSEA (Steiger & Lind, 1980) and Comparative Fit Index (Bentler, 1990) are based on these deviance values. These and other fit indices for models with continuous indicators (normal and non-normal) have been heavily studied and are in widespread use.

It is difficult to extend these statistics to models with categorical indicators using ML estimation, though model fit tests are available for limited-information estimators. The unrestricted multinomial model for the frequency table of observed response patterns can be used

for the saturated model to compute the statistic; however the finite-sample properties of this statistic in realistic models are not acceptable (Koehler & Larntz, 1980). New promising developments are limited information methods for goodness-of-fit testing, especially the M_2 for dichotomous responses (Maydeu-Olivares & Joe, 2006) and M_2* for polytomous responses (Cai & Hansen, 2013). Instead of testing for goodness-of-fit against the entire multinomial distribution, these statistics collapse across cells to test against tables for the marginal distributions, yielding better Type I error control and better power (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2006). Despite the importance of assessing model fit, I do not focus on this issue in my project because I investigate estimator performance for properly specified models.

Regardless of approach to ML estimation, practical constraints in finite samples affect the quality of the solution.

General Performance of ML in Finite Samples. The desirable properties of ML estimation for GLFA models are based on large-sample (asymptotic) theory, and it is important to consider the quality of ML solutions in finite samples. Many factors including indicator type (categorical versus continuous), sample size, number of indicators, number of factors, magnitudes of factor loadings, and additional sources of model complexity (e.g. cross-loadings) are important for expected convergence to a proper solution (i.e., solution propriety) and stability of parameter estimates. For GLFA with continuous indicators, solution propriety generally increases with sample size, the number of indicators per factor, and the strength of factor loadings (Gagné & Hancock, 2006; see also Anderson & Gerbing, 1984; Marsh et al., 1998). Categorical variables necessarily contain less statistical information than continuous variables. Therefore, larger sample sizes are needed to obtain similar solution propriety when indicators are

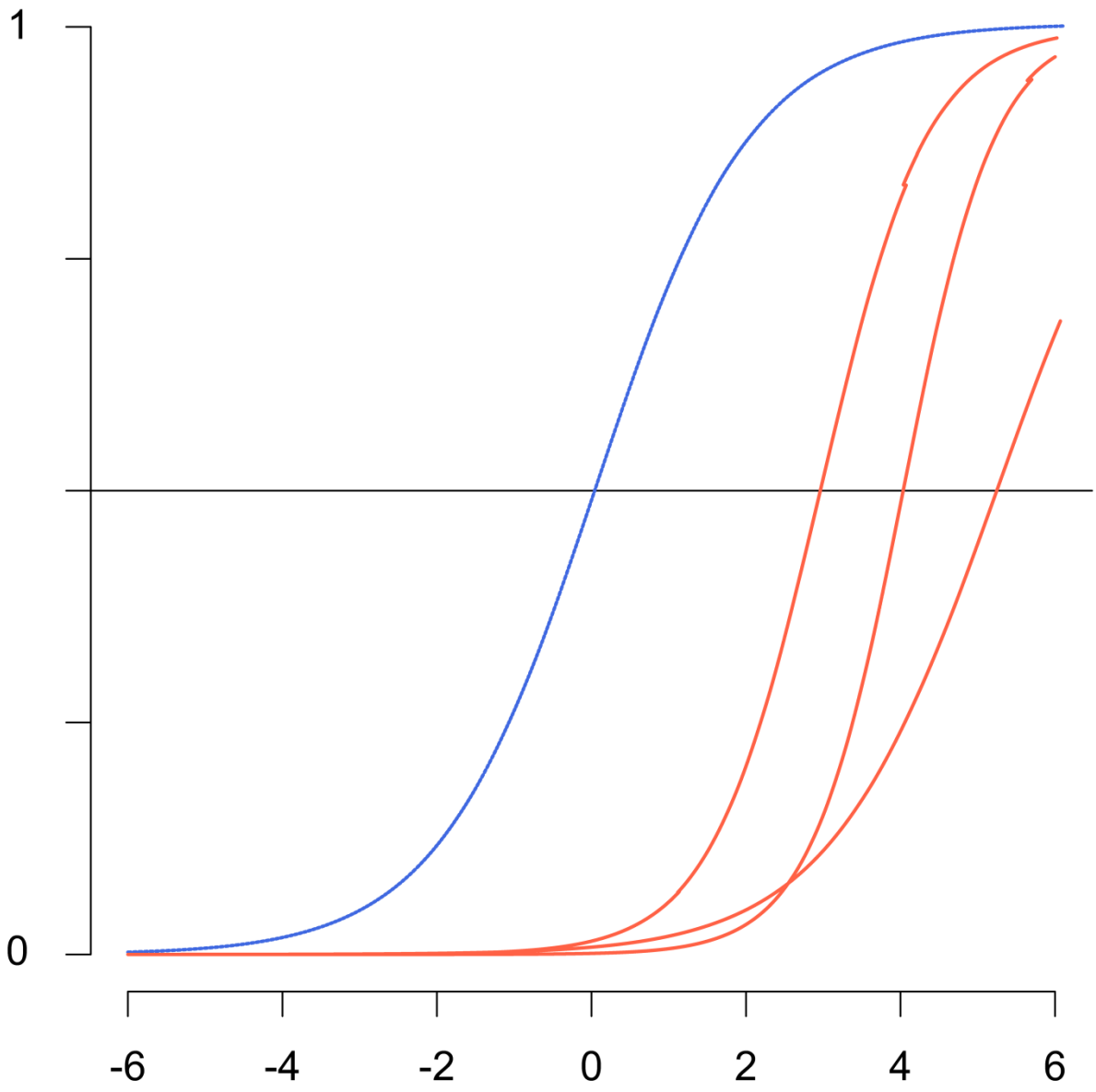
categorical (Moshagen & Musch, 2014; Wirth & Edwards, 2007). Besides this, sparseness – as an issue distinct from sample size – is a concern for models with categorical indicators.

Complications due to Sparseness. Sparse endorsement is expected under a variety of combinations of factor loadings and thresholds. Strictly following from the definition of marginal response probability in Equation 4 – high thresholds, low factor loadings, or a combination of both can lead to low probabilities of endorsement and therefore sparse observed frequencies in finite samples.[3] Example trace lines for item characteristic curves with high thresholds that could lead to sparseness are plotted in Figure 2. Typically in applications the values of factor loadings and thresholds are not independent, for example items with a high threshold may commonly also have a high loading (e.g. a rare but extreme behavior or a very difficult test question). In substance use research, low endorsement for rare behaviors such as early drug use is more likely consistent with high thresholds coupled with substantial loadings, meaning a high level of the latent trait is required to endorse an item, rather than very low factor loadings, which would signify a weak relationship with the latent construct.

Though sparseness has not been specifically studied for ML estimation of GLFA models with categorical indicators, a limited literature suggests that ML estimation performance is poor in conditions characterized by sparseness – namely smaller sample sizes combined with smaller probabilities for item endorsement (Forero & Maydeu-Olivares, 2009; Moshagen & Musch, 2014; Wirth & Edwards, 2007). Forero and Maydeu-Olivares (2009) found that ML estimation failed in small samples (200 observations) for binary items with low endorsement (10%), especially with fewer items per factor and low factor loadings. Moshagen and Musch (2014) found that ML estimation of GLFA models in smaller samples could yield highly distorted

[3] The reverse is also true, for example very low thresholds could lead to an item that is almost always endorsed and non-endorsement is sparse.

Figure 2. Item characteristic curves for one standard item and three items that could lead to sparseness.



parameter estimates and standard errors in smaller samples, even when ML estimation converges.

These previous simulation studies were not specifically motivated to study sparseness, and sparseness in these studies was confounded with other important factors. In Moshagen & Musch (2014), binary items had a 50% probability of endorsement, and sparseness was a result of small samples. The problems observed by Moshagen and Musch (2014) and Forero and Maydeu-Olivarez (2009) were also associated with models that were poorly-determined with few indicators per factor and low factor loadings. Research has not yet determined what levels of sparseness are problematic for ML estimation even in well-determined models (e.g. specific marginal probabilities or item frequencies), the impact of number or proportion of sparse items, the impact of sparseness for different item loadings, or the implications of different patterns of sparseness across latent factors. For example, it is not known if having half of all items sparse, spread across two factors, has a different impact compared to having all sparse items on one factor. Theory suggests that sparseness becomes an issue in ML estimation of GLFA models with categorical indicators in two key ways.

First, it is likely more difficult to obtain stable parameter estimates for items with low endorsement in finite samples. One reason for this can be inferred from the issues of quasi-complete or complete separation in logistic regression analysis with sparse outcomes (see Agresti, 2012, Ch. 6). This occurs when the outcome separates or nearly separates some combination of predictors with the result that discrimination is perfect, the maximum likelihood solution does not exist, and any obtained estimates will be untrustworthy. Similarly, sparseness may suggest parameter values near the boundary of the parameter space, which breaks the important regularity conditions for properties of the ML estimator (see e.g., Agresti, 2012, Ch 1).

Secondly, probabilities of response patterns involving sparse items become small. Because the probabilities of each response pattern are modeled as a function of the independent item parameters, the sparse multinomial distribution is not directly estimated. Any empty cells in the multinomial table are not predicted. Many very small cells however may be difficult to predict by extreme model parameters, but this issue is largely unexplored.

Further, especially in models with categorical indicators, there is likely interplay between sample size, model complexity, sparseness, and estimation challenges. More complex models combined with modest sample sizes and rare endorsement are expected to compound the problem of sparseness, and it is easy to build models that are more complex than data can support. Models where estimation challenges arise are not needlessly complicated; examples include latent curve models with multiple indicators for improved measurement (see Bollen & Curran, 2006, Ch. 8), multiple-group models (see Bollen, 1989, Ch. 8), and moderated nonlinear factor analysis (Bauer & Hussong, 2008). These are just a few examples of theoretically justified increases in model complexity, especially for substance use research; yet increased complexity, when combined with categorical indicators and finite sample sizes, may lead to empirical underidentification and estimation challenges. Researchers currently facing these estimation challenges must combine items, collapse item categories (if more than two categories), or drop items, potentially sacrificing information. For example, Hussong, Huang, Serrano, Curran, & Chassin (2012) report combining items assessing drug use other than marijuana due to sparseness, and Hussong, Flora, Curran, Chassin, and Zucker (2008) report dichotomizing ordinal items because sparse endorsement led to estimation problems.

In sum, ML estimation is satisfactory for GLFA in some cases, but ML is not designed to work well for finite samples with sparse data. In many domains of psychology and especially substance use research, it is not always an option to avoid sparse items when the pool of items is

limited, sample size is limited, or items are particularly important to comprehensively measure a construct. For example, if the intended measure is a tendency towards self-harm, a rare behavior, it may be theoretically important to include some items about extreme self-harm behaviors, even if they have low base rates. Next I introduce Bayesian estimation as an alternative when ML estimation breaks down.

Bayesian Estimation

Bayesian estimation is based on a historically distinct approach to statistical inference from frequentist-based methods such as ML. Some advantages for the estimation of GLFAs with categorical data may exist in a Bayesian framework[4]; however these potential advantages are balanced with an increase in methodological complexity. Further, these have yet to be studied specifically for the case of sparse items in GLFA.

In Bayesian statistics, parameters are random variables (rather than fixed, true values as in classical statistics). A Bayesian estimation approach requires selection of an appropriate prior distribution for each parameter in the model. The prior distribution $\pi(\theta)$ is combined with the model likelihood function $L(y;\theta)$ — the same likelihood maximized by ML estimation— to arrive at the posterior distribution $\pi(\theta \mid y)$ via Bayes' theorem:

$$\pi(\theta \mid y) = \frac{\pi(\theta)L(y;\theta)}{\int \pi(\theta)L(y;\theta)d\theta} \propto \pi(\theta)L(y;\theta). \qquad 8$$

It is on this posterior distribution that inferences are based; specifically detailed information is available about the distributions of individual parameters.

This is an important distinction between a Bayesian estimation approach and more traditional frequentist approaches. Because the posterior distribution of the parameters is

[4] The Bayesian approach I focus on is not the only possible approach. Maximum a posteriori (MAP or modal Bayes) estimation pairs prior distributions from Bayesian statistics with a method of estimation similar to ML estimation (Mislevy, 1985). I focus on "full" Bayesian inference and MCMC to describe the posterior distribution in part for its generality and potential to scale to higher dimensional problems.

available, standard errors or credible intervals (the Bayesian analogue to confidence intervals) are based on the percentiles of the posterior, which can have any distributional shape (e.g., symmetric, asymmetric, skewed). In contrast, a maximum likelihood approach assumes that the asymptotic distribution of a parameter estimate is normal, an assumption based on large-sample theory. Because it does not rely on large-sample theory, Bayesian estimation can be advantageous for fitting models to small samples. However, there are important tradeoffs and assumptions inherent in either approach. In a Bayesian analysis, inferences may be dependent on choices made about the prior distribution, whereas in ML estimation, asymptotic properties may not hold in finite samples.

Important components of a Bayesian analysis are: prior specification, model specification, posterior computation, and evaluating the posterior solution. The model specification does not differ in a Bayesian analysis, so I focus on the other three components in the next three sections. For this introductory material, I borrow from *Bayesian Data Analysis* by Gelman et al. (2013), to which I refer interested readers for further details on all aspects of Bayesian inference.

Prior Specification. Prior distributions for each model parameter can be used to express prior knowledge or information about parameter values, even if the information only concerns permissible parameter values. This prior knowledge is combined with the information in the data by Bayes' theorem to arrive at the posterior distribution in a process known as Bayesian updating. The process of selecting priors is extremely flexible; priors may vary in distributional form and shape. Conjugate priors use distributions that, when combined with the likelihood, yield a posterior distribution of the same form. Conjugate priors have historically been useful for computational simplicity, but this restriction is not necessary and different parametric or non-

parametric distributions may be chosen. The parameters (scale, location, etc.) governing the prior distributions of parameters are called hyperparameters.

Priors can be diffuse or have relatively more mass near a range of plausible values, and the level of diffusion in the prior is usually expressed by the hyperparameter values. Many flat priors do not have "proper" probability distributions, meaning they do not integrate to 1. For example a uniform distribution on the real line ($U(-\infty,\infty)$) is improper. The use of improper priors can lead to an improper posterior distribution, invalidating inference, therefore using improper priors requires care to ensure that the posterior distribution is proper. Prior distributions and their hyperparameters can be chosen from prior knowledge, certain default values, or from the data (data-dependent priors). Priors may also have hyperpriors governing the distribution of the hyperparameters. Sometimes priors are labeled as informative/subjective or uninformative/objective for peaked and diffuse priors, respectively. However I avoid this labeling because it can be misleading as a flat prior may be highly informative for some purposes, and the level of information in a particular prior varies case-by-case (see Zhu & Lu, 2004).

Flat priors can also be used to obtain results consistent with maximum likelihood estimation, using Bayesian estimation methods simply as a computational tool (Gelman et al., 2013). With little prior information and adequate sample size, Bayesian and ML estimation converge on the same solution; this means that Bayesian estimation can be expected to perform as well as ML estimation when ML is converging to a stable solution (See Gelman et al., 2013, Ch. 4; Wasserman, 2005). Including prior information can improve an analysis by building on existing knowledge and is a way to be transparent about prior beliefs, incorporating hypotheses into the analysis. It is fairly common to at least restrict parameter values to their admissible range, for example constraining variances to be positive (Gelman et al., 2013). One concern is

that such restrictions may mask misspecification, because a negative variance may be a symptom of misspecification (Kolenikov & Bollen, 2012).

Although in some cases strongly concentrated priors may produce misleading results, this is not problematic for properly specified models[5] as long as there is non-zero probability at the true values with enough data, even using relatively concentrated but inaccurate priors (Depaoli, 2014). With limited sample sizes, parameter estimates are more sensitive to prior values (Berger & Bernardo, 1992; Kass & Wasserman, 1996). There are also hazards to relying on default priors of any kind, including default flat priors (Kass & Wasserman, 1996).

For Bayesian estimation of the GLFA model defined earlier, priors are needed for the parameters governing the distribution of the latent factors, factor loadings, item intercepts, and any thresholds. Priors are not assigned for any fixed parameters. An example prior specification for a univariate model with binary indicators is as follows:

$$\begin{aligned} \pi(\lambda_i) &\sim U(-\infty, \infty) \\ \pi(\nu_i) &\sim U(-\infty, \infty) \end{aligned} \qquad 9$$

where the model is scaled by setting the mean and variance of the latent factor to (0,1). However there is a reasonable basis to restrict these priors. General ranges and typical values of these parameters are known. If theory would strongly dictate that all items should be positively related to the latent variable, the prior distribution could favor positive values. Truncated priors may be used to constrain ranges for parameters. For example if the variance of η is estimated, a normal distribution truncated at zero (half-normal) would constrain estimated variance to positive values. Setting this variance to a large value (e.g. 100) for a half-normal distribution would form a very flat prior constrained to positive values, whereas a half-normal (0,1) distribution would express a prior .95 belief that values should be between 0 and 1.96. Because thresholds ν_j are

[5] The influence of concentrated prior distributions, correct and incorrect, for misspecified GLFA models is an important area of future research.

expected to range from about negative 4.5 to 4.5, a reasonable prior could be normal with variance focused in this range. With multiple ordered threshold categories, it is also necessary to constrain their order in the priors and estimation. More specific priors may also be specified for individual items, for example for a self-harm scale, thinking about harming oneself could have relatively lower prior probability ranges for thresholds than an item about repeatedly injuring oneself.

Even when reasonable prior specification guidelines are given, and especially without useful prior information, a sensitivity analysis should be conducted to see whether the results are robust to prior specification (e.g. Song & Lee, 2012, Ch. 3). This can be done for example by perturbing the prior hyperparameter values or by considering other prior choices. After specifying the prior distributions for each parameter, a Bayesian analysis proceeds by describing the posterior, usually by MCMC simulation.

Posterior Simulation. The posterior distribution is usually impossible to describe analytically. Consequently, Bayesian estimation of most interesting models, including GLFA, only became feasible with the introduction of Markov chain Monte Carlo (MCMC) simulation methods which provide an approach for generating samples from the posterior distribution (Tanner & Wong, 1987; Gelfand & Smith, 1990). Whereas traditional Monte Carlo algorithms take independent samples from a target distribution directly, Markov chain Monte Carlo methods generate correlated samples that asymptotically converge to the target posterior distribution. MCMC simulations are initialized with starting values and require a burn-in period of draws before the chain has reached the target distribution (i.e., the chain has converged). After convergence, subsequent draws will be approximately from the target posterior distribution. The posterior distribution is then summarized from these samples. For a clear overview of some common MCMC algorithms and practical issues in implementation, see Edwards (2010).

Most existing work for Bayesian GLFA (both FA and IRT models) has focused on two types of MCMC algorithms: Gibbs and Metropolis-Hastings (Albert & Chib, 1993; Béguin & Glas, 2001; Edwards, 2010; Patz & Junker, 1999, Song & Lee, 2002, 2012; Lee & Tang, 2006). Gibbs sampling (Geman & Geman, 1984) is useful when it is impossible to sample from the full posterior for all parameters in a model (θ), but θ can be partitioned into two or more conditional distributions in convenient forms for sampling. The Gibbs sampler is set up to sample iteratively from each of the conditional distributions of a subvector of θ given the observed data $\boldsymbol{y}$ and current values of the other parameters. Under mild regularity conditions, these samples converge to the target stationary distribution, the posterior of θ (Geman & Geman, 1984). Although simple to program and useful for many models, prior choice and model choice are restricted in order to arrive at a posterior that can be partitioned into convenient conditional distributions. For example, priors are usually restricted to the class of conjugate priors, and the choices for prior variance can have biasing influences on the posterior distribution (Gelman, 2006). Gibbs sampling for GLFA models is not sufficient on its own if categorical indicators are included (Lee & Song, 2012).

Metropolis-Hastings (MH; Metropolis et al., 1953; Hastings, 1970) is a much broader family of algorithms for posterior simulation, actually including Gibbs sampling as a special case (see Gelman et al., 2013, p. 318). MH algorithms sample a value from a convenient proposal distribution (e.g., normal) and accept that proposed value with probability carefully defined to form a chain that converges to the posterior. For GLFA estimation, more general MH algorithms are used in the MCMC chain to sample from any nonstandard distributions when Gibbs is not an option (Lee & Song, 2012). MH sampling for GLFA models can be implemented an infinite number of ways, making it much more general. However the rules controlling implementation require careful oversight and fine-tuning in order to effectively explore the parameter space, and

convergence for high-dimensional target distributions can be effectively impossible (Gelman et al, 2013). Often, MCMC algorithms are written specifically for a particular model and prior specification and even tailored to perform well for different data. Given these essential properties of MCMC, there are some major barriers to widespread use of MCMC techniques for Bayesian estimation for GLFA.

Gibbs and MH sampling depend on "random walk" behavior to converge to and explore the target distribution. This random walk, while accomplishing its designed purpose, is also inherently inefficient: simulations may zigzag erratically through the target distribution for many iterations. An alternative to Gibbs and MH algorithms designed to suppress random walk behavior is Hamiltonian Monte Carlo (HMC, sometimes called Hybrid Monte Carlo). HMC is based on methods for studying molecular dynamics in physics, specifically Hamiltonian dynamics (Duane, Kennedy, Pendleton, & Roweth, 1987). Whereas other MCMC algorithms use a probability distribution to propose future states in the Markov chain, HMC algorithms use physical state dynamics, specifically Hamiltonian dynamics.

To understand the intuition of Hamiltonian dynamics – and by extension HMC– I borrow a description of the physical interpretation of Hamiltonian dynamics from Radford Neale (2010):

> In two dimensions, we can visualize the dynamics as that of a frictionless puck that slides over a surface of varying height. The state of this system consists of the *position* of the puck, given by a 2D vector q, and the *momentum* of the puck (its mass times its velocity), given by a 2D vector p. *The potential energy, U* (q), of the puck is proportional to the height of the surface at its current position, and its *kinetic energy*, K (p) is equal to $|p|^2/(2m)$, where m is the mass of the puck. On a level part of the surface, the puck moves at a constant velocity, equal to p/m. If it encounters a rising slope, the puck's momentum allows it to continue, with its kinetic energy decreasing and its potential energy increasing, until the kinetic energy (and hence p) is zero, at which point it will slide back down (with kinetic energy increasing and potential energy decreasing).

Whereas the physical interpretation of Hamiltonian dynamics is used to describe objects moving through space, these concepts can also be translated to describe the movement of parameters through the posterior distribution. In this interpretation, the position corresponds to the

parameters of interest, the potential energy relates to the probability distribution of the parameters of interest, and momentum variables are added for each parameter of interest to describe these dynamics.

The Hamiltonian dynamics are expressed by a system of differential equations that must be approximated, specifically by discretizing time and proceeding through time in steps. In each series of steps, the momentum, position, and potential energy for the system are updated. HMC algorithms simulate this process.[6] Certain properties of Hamiltonian dynamics make it especially useful for MCMC; essentially during the simulation it represents and preserves volume of the posterior distribution, and uses this representation of the posterior distribution to guide exploration. Because of preservation of volume and simulation of momentum, HMC can be used to move more efficiently through the parameter space than Gibbs or MH sampling (Neal, 1993, Chapter 5). Although more efficient, HMC requires tuning of parameters to guide the chain, and this complicated tuning process has discouraged widespread implementation. However, the No-U-Turn sampler (NUTS; Hoffman & Gelman, 2014) effectively automates this tuning process.

There have been many efforts to make software for general-purpose Bayesian estimation, most using combinations of MH and Gibbs sampling. Some programs have either been inflexible– not applicable to a wide range of models, data, or priors (e.g. Mplus) – or general at the risk that MCMC may be inefficient and fail to converge (see Carpenter et al., 2015). Use of MCMC in a canned statistical package is somewhat risky, as it is challenging to implement MCMC correctly, and further it is necessary to ensure that all aspects of the MCMC estimation were successful before making inferences (MacCallum, Edwards, & Cai, 2012). One recent attempt to create general software for Bayesian estimation is the Stan programming language

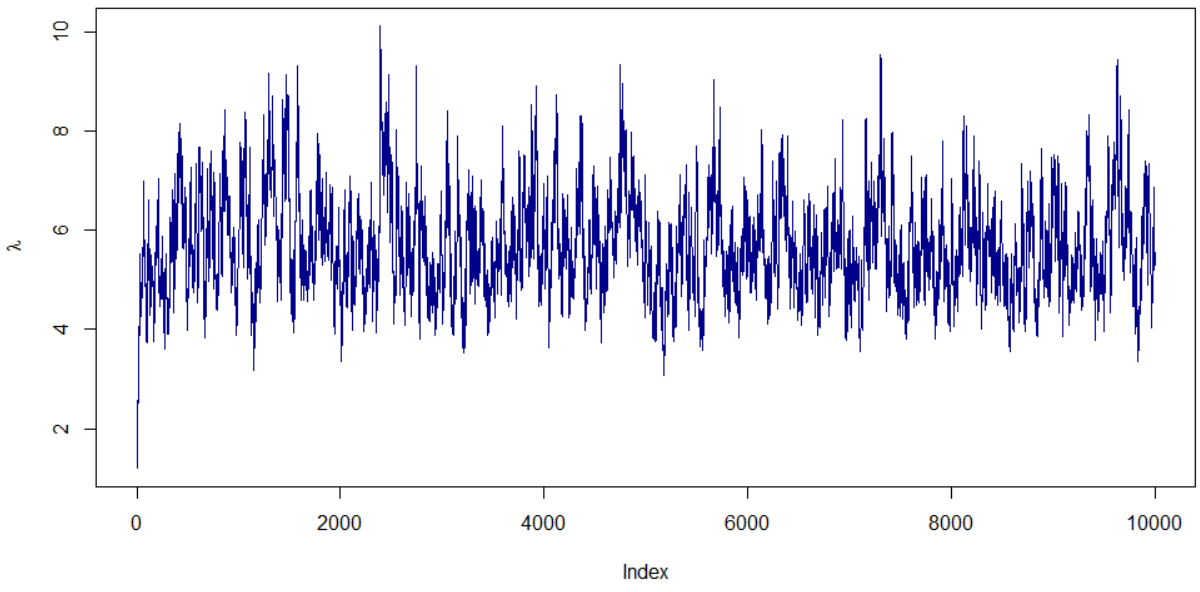
[6] Because many concepts of Hamiltonian dynamics and HMC are unfamiliar to non-physicists, a detailed description of HMC is beyond the scope of this project. I refer interested readers to Neal (2010) and Gelman et al. (2013, pp. 300-308) for more details, however note that this material is necessarily technical.

(Stan Development Team, 2015), which uses Hamiltonian Monte Carlo for efficient posterior exploration and the NUTS sampler to automatically tune the algorithm.

Posterior Evaluation. After MCMC sampling, it is necessary to evaluate the samples for convergence and summarize the posterior to make inferences. There are many techniques to help assess MCMC convergence (see Gelman et al., 2013, for a review). However it is generally impossible to know for sure that any single chain has converged, because methods for monitoring convergence assess necessary but not sufficient conditions for convergence.

One good practice is to run multiple chains from different starting values and check that the chains appear to converge to the same solution (Gelman et al., 2013). A useful visual diagnostic tool is a traceplot which shows the iteration number plotted against the sampled values for a parameter; an example traceplot is shown in Figure 3. In these plots good mixing, lack of periodicity and clear movement from the starting values to a stable target distribution are all evidence of convergence.

Figure 3: Example MCMC diagnostic trace plot



Because the draws from the posterior are not independent the "effective number of simulation draws" is less than the total number of draws. The number of effective draws depends on the autocorrelation of the simulation draws. Asymptotically the number of effective samples if there are *n* draws from *m* chains is

$$n_{eff} = \frac{mn}{1+2\sum_{t=1}^{\infty}\rho_t} \qquad 10$$

where ρ_t is the autocorrelation of the sequence ϕ at lag *t*. Computing the effective sample size in practice requires estimating the infinite sum of the autocorrelations from a positive partial sum, $\sum_{t=1}^{T}\hat{\rho}_t$ using variance and covariance information from within and between sequences (see Gelman, et al., 2013, pp. 284-87 for complete computational details). A measure of effective sample size is useful to measure efficiency of the chain and determine whether sufficient uncorrelated samples have been drawn for posterior inference.

Additionally, the potential scale reduction statistic ($\hat{R}$; Gelman and Rubin, 1992) can be computed to help monitor whether a chain has converged to the equilibrium distribution. The potential scale reduction statistic compares variability within a sequence to variability between other randomly initiated chains as

$$\hat{R} = \sqrt{\frac{\text{var}^{+}(\phi \mid y)}{W}} \qquad 11$$

where $\text{var}^{+}(\phi \mid y)$ is an estimate of the marginal posterior variance of the estimand, and W is an estimate of within-sequence variance (see Gelman et al., 2013, pp. 284-285 for full details).

If the value of $\hat{R}$ is one, this is evidence of convergence, while values above one suggest that the chain has not converged. Importantly, all parameters in a model must show evidence of convergence before it is suitable to make inferences from the posterior distribution.

Rather than a point estimate and large-sample based confidence intervals, Bayesian estimation produces posterior distributions for each parameter. Often it is useful to examine the posterior means and quantiles, including 95% posterior intervals to make inferences about each parameter.

Model Fit Assessment. Evaluating goodness of fit for Bayesian models is an active area of research. Posterior predictive checking (PPC; Gelman, Meng, & Stern, 1996) can be used to compare the value of any test statistic for the observed data to values computed for simulated data obtained from draws from the posterior distribution. The expectation is that, for well-fitting models, data simulated from draws from the posterior (which is based on the hypothesized model for $\boldsymbol{y}$), should be similar to $\boldsymbol{y}$. A posterior predictive *p*-value is often calculated as the proportion of simulated replications for which the test statistic equals or exceeds its realized value. Posterior predictive checking is popular in applied Bayesian analyses and has been demonstrated for GLFA models (Béguin & Glas, 2001). However, PPC has been criticized because the observed data will be more consistent with the posterior distribution, which it was used to compute, than random draws from the posterior (e.g., Yuan & Johnson, 2012). This double-use of the data is theoretically problematic and sacrifices power to detect misfit. Further, the posterior predictive *p*-values are not uniformly distributed under the proposed model, making their interpretation difficult (Bayarri & Berger, 2000). Yuan & Johnson (2012) propose an alternative methodology, involving comparisons of what they term pivotal discrepancy measures, which are uniformly distributed and have higher statistical power to detect misfit.

Advantages of Bayesian Estimation for Sparse GLFAs. Though Bayesian estimation has been profitably used to estimate complex GLFA models (e.g. Edwards, 2010; Song & Lee, 2012), it has not been studied for the problem of estimating GLFA models with sparse, categorical indicators. However, theory suggests that Bayesian estimation should be a useful

alternative when ML breaks down. Incorporating prior information has been shown to be especially useful in sparse data settings (Dunson & Dinse, 2001; Peddada, Dinse, & Kissling, 2007). Dunson and Dinse (2001) suggest a Bayesian method for studying tumor incidence rates, which are rare events and often difficult to predict because of small sample sizes. By incorporating historical data as prior information, their method leads to more interpretable results and can improve detection of small but biologically important changes in incidence rates (Dunson & Dinse, 2001; Peddada, Dinse, & Kissling, 2007).

Introducing priors to an analysis should be an advantage for dealing with sparseness in GLFA, both theoretically and computationally. The prior should have a stabilizing, shrinkage effect on parameters with little data available for their estimation. Often applied researchers prefer the unbiasedness property of maximum likelihood estimation, but in cases of sparseness, it may be better to prefer estimation with some bias in exchange for lower variance to avoid overfitting. This rationale (i.e., increased stability at the cost of some bias) is the same used for regularized regression methods such as ridge regression or lasso regression (Tibshirani, 1996), which are used in a frequentist framework but also have Bayesian interpretations (Park & Casella, 2008). The stabilizing effect of reasonable priors should also be beneficial for computational problems arising from sparse categorical data because the priors can be used to avoid improper solutions and aid convergence.

The prior may thus provide more information than the data for some parameters in some cases. This prior influence may be problematic for some circumstances and depending on the purposes for specific model inferences, however in general if reasonable priors are chosen, prior-driven stabilization may be advantageous. In the case of thresholds nearing extreme values due to sparse data, shrinking these extreme values may be computationally advantageous and more reliable.

In summary, Bayesian inference is remarkably flexible and can be adapted to provide good performance even in challenging or less than ideal circumstances with large models, small samples, missing data, or sparseness. As such, Bayesian estimation is a promising alternative to ML estimation for GLFA with sparse indicators; however it is important to evaluate computational challenges and sensitivity to prior specification.

Current Research

Sparse categorical indicators commonly arise in substance use research due to finite sample sizes and the potential for extreme items. In the current work I evaluated the impact of sparseness on ML estimation of GLFA and investigated Bayesian estimation as an alternative to ML estimation for sparse indicators, to stabilize estimates and aid convergence. Although theory suggests that using priors to stabilize estimates may be preferable to ML estimation for sparse items in GLFA, it is not possible to compare these approaches analytically for finite samples. Therefore, to accomplish these aims, I conducted a simulation study centered on the following theoretically derived hypotheses:

1. Maximum likelihood estimation for GLFA models with sparse, categorical indicators was expected to fail to consistently produce converged, reasonable solutions with a higher proportion of sparse items, decreasing probability of endorsement, and lower item loadings. Efficiency of solutions was expected to be poor even for converged replications.
2. In conditions where maximum likelihood estimation performs well, I hypothesized that Bayesian estimation would perform as well or better, specifically in terms of efficiency of parameter estimates.

3. Bayesian estimation was expected to outperform maximum likelihood as sparseness increases in terms of convergence to reasonable solutions, efficient parameter estimates, and empirical power.

I varied levels of item sparseness, item loadings, and patterns of sparse items for a two-factor GLFA model with binary indicators. Specifically, I studied 2 levels of sparseness, 2 factor reliabilities, and 3 patterns of sparse items in a simulation design with (2x2x3) 12 cells, in addition to examining 2 baseline (even endorsement) conditions, one for each level of item loading.[7] In Study 1, I determined conditions where ML estimation is impaired due to sparseness. In Study 2, I examined Bayesian estimation where ML performs well and in a subset of conditions determined in Study 1 where ML estimation performs poorly.

[7] Note that this simulation design is not fully crossed, because baseline conditions with even endorsement on all items do not cross with the manipulations for sparse items.

CHAPTER 2: STUDY 1 – MAXIMUM LIKELIHOOD ESTIMATION

Simulation Study Design

Model Design

To evaluate the impact of sparseness for ML estimation of GLFA models, I simulated data consistent with a two-factor GLFA with 5 binary indicators per factor. I chose a multidimensional model in order to study the effects of patterns of item sparseness across factors and bias and efficiency in the estimated correlation between factors. The correlation between factors was moderate, $\psi_{12} = .3$ for all conditions. Sample size was constant, N =500 for each of 500 replications per condition. This value was chosen to be representative of a modestly large sample size, a sample with which substantive researchers would typically feel confident estimating and interpreting structural equation models. I did not vary sample size because this would confound marginal endorsement rates and cell frequencies, and there was no expected interaction between marginal endorsement and sample size. Larger sample sizes, holding constant the item parameters, should improve convergence, estimates, and standard errors. I manipulated item parameters to induce sparseness and determine conditions where ML estimation is meaningfully affected by sparseness. I examined parameter estimate convergence, bias, efficiency, confidence interval coverage, and empirical power as outcomes.

Design Factors

I examined model convergence, parameter estimate bias and efficiency, confidence interval coverage, and empirical power for the given model specification and sample size, for different item loading values and levels and patterns of item sparseness.

Item loadings. I evaluated the effects of sparseness for two item loading parameter values, $\lambda_i = 1.5$ and $\lambda_i = 2.0$, corresponding to communalities of .41 and .55. These item loadings parameter values were informed by a review of parameters encountered in practice (e.g. Hussong, Flora, Curran, Chassin, & Zucker, 2008) and simulation studies for similar models (e.g. Cai, 2010; Edwards, 2010; Curran et al., in preparation).

Item thresholds. I varied threshold parameters to induce sparseness, examining a baseline (even endorsement) condition and two conditions with high thresholds. For the baseline conditions endorsement was even on all items (all $\nu_i = 0$). To induce sparseness, I set threshold parameters to $\nu_i = 3.85$ and $\nu_i = 4.9$ (logit-scaled). For conditions with $\lambda_i = 1.5$, this corresponds to marginal probabilities of *p*=.05 and *p*=.02, respectively, and for $\lambda_i = 2.0$ this results in marginal probabilities of *p*=.075 and *p*=.035. The marginal probabilities of endorsement for different thresholds were derived by integrating over the distribution of η in Equation 4; this integration was done by simulating a large number of draws (i.e. 10^7) from a standard normal distribution, and calculating the probability of response given each value of η using Equation 4. This yields expected marginal frequencies of 25; 10 (when $\lambda_i = 1.5$), and 37.5; 17.5 (when $\lambda_i = 2.0$) for the sample size of 500.

Pattern of sparse items. In addition to baseline conditions with no sparse items, I examined three patterns of sparse indicators in the model. To determine if the effect of sparseness depended on the pattern of sparse items across factors, I compared two conditions with a total of four sparse indicators distributed differently across factors. In one condition, all four sparse indicators were on the same factor, and in a second condition two sparse indicators were distributed evenly on each factor. I also examined a high sparseness condition, with four of five indicators sparse on both factors.

Summary of simulation design. The simulation factors described formed a fractional factorial design, because all possible combinations of levels of each factor were not fully crossed. Fractional designs have been recommended to remove redundancy in simulation study designs, especially when higher-order interactions among the design factors are not of interest (Skrondal, 2000).There were a total of 14 conditions in the simulation design, and the conditions are summarized in Figure 4.

Figure 4. Summary of simulation design and factorial design matrices for meta-models

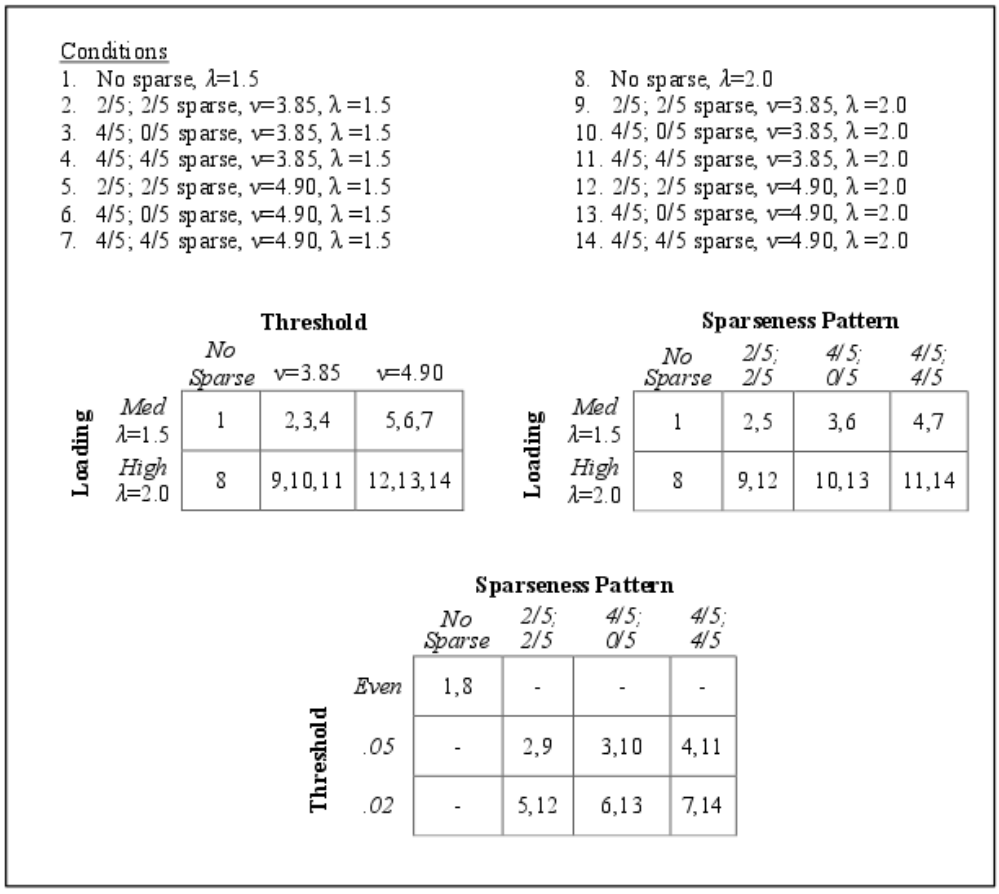


Figure 4. Descriptions of 14 simulation conditions. E.g., "2/5; 2/5 sparse" means 2 of 5 items sparse on factor 1 and factor 2, and "ν = " gives threshold.

Data Generation

Data was generated in matrix form within R (R Core Team, 2015) from a distribution with fixed population values using the following three step algorithm. First, I generated random standard normal latent variable values for both factors from a bivariate normal distribution with a correlation =.30 between factors. Second, I calculated probabilities of responses given parameter values, latent factor scores, and the defined model and logit link function (i.e., Equation 4). Third, I simulated item responses as draws from a Bernoulli distribution with probabilities calculated in the previous step. If endorsement on any item was zero, the replication was discarded and replaced with a new replication until 500 replications were simulated with non-zero endorsement for all items[8]. This resulted in a 500 x 10 (N x P) data matrix for each of the 500 replications for each cell of the simulation design. Note that the design of the simulation study, with fixed population values, is consistent with a traditional (frequentist) specification, whereas a Bayesian specification would draw from a distribution of population values.

Estimation

I estimated the correct model for each replication using full information maximum likelihood as programmed in Mplus version 7 with a logit parameterization and default start values, convergence criteria, and the default integration method of adaptive numerical integration with 15 integration points. The default integration method and number of integration points is well-suited for a GLFA with 2 latent factors, though alternative methods of integration are preferable for more complex models with more latent factors (Wirth & Edwards, 2007). Estimation for each replication was automated using the MplusAutomation R package (Hallquist & Wiley, 2014). In order to estimate the model, the latent factors were identified by setting the

[8] Not allowing zero endorsement technically changes the population parameter for the probability of item endorsement. However, the impact is trivial because the probability of observing no endorsement for an item with 2% probability of endorsement is less than <.0001 for a sample size of 500, even with 8/10 items sparse.

variance to unity for each factor and estimating all factor loadings[9]. The program syntax is provided in Appendix A.

Evaluation Criteria

I evaluated performance of maximum likelihood estimation in terms of convergence, bias, efficiency, confidence interval coverage, and empirical power.

Convergence and extreme estimates. I monitored convergence of replications to proper solutions in each condition, as defined by the algorithm in Mplus. Convergence failures are reported as errors in the output file. However, Mplus may also give warnings and errors that do not necessarily indicate non-convergence (e.g., warning that an estimate has been fixed). I monitored all warnings and errors to screen for serious errors indicating nonconvergence versus ignorable warnings. Mplus may fix threshold estimates if they reach boundaries (e.g., logit thresholds outside [-15,15]) at certain points in the estimation routine, but estimates outside of this range may also be reported (Muthén & Muthén, 2014). In addition to convergence to proper maximum likelihood solutions, I also monitored solutions for extreme estimates which would seem suspicious in practice.

Raw bias. Raw bias was calculated for all parameters $(\lambda_i, \nu_i, \psi_{12})$. Raw bias is calculated generally for parameter θ by subtracting the true value from the *r*th estimate $(\hat{\theta}_r)$ and averaging across the total number of replications in the cell (*R*):

$$\frac{\hat{\theta}_r - \theta}{R}. \qquad 12$$

Raw bias for estimates within each replication was computed for meta-models of the simulation design, and average bias was used to interpret bias for parameters within each condition.[10]

[9] This model specification is only locally identified (Bollen & Bauldry, 2010; Loken, 2005), as there is a sign indeterminacy for the factor loadings on one or both factors. For the estimation routines used in Mplus for these models and data, the sign indeterminacy is not an issue and leads to solutions with a majority of positive factor loadings.

Because the mean is sensitive to extreme values, I also calculated median bias and recorded minimum, 5th quantile, 95th quantile, and maximum values for parameters in each condition.

Efficiency. I examined root mean square error (RMSE) as a measure of parameter estimate efficiency for each parameter, computed generally for parameter θ as

$$\sqrt{\frac{\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)^2}{R}}. \qquad 13$$

RMSE is a measure of both sampling variability and squared bias, with larger values reflecting greater variability in estimates relative to the true value. When estimates are unbiased, the RMSE can be thought of as the empirical standard error. When bias is present, efficiency measured by RMSE reflects overall accuracy. Because RMSE is sensitive to extreme values, the median absolute deviation about the median (MAD) was also included as a robust measure of efficiency (Huber & Ronchetti, 2009), calculated for each parameter in replication *r* as

$$MAD_k = Median\left\{\left|\hat{\theta}_r - M_k\right|\right\} \qquad 14$$

where

$$M_k = Median\left\{\hat{\theta}_r\right\}.$$

Confidence interval coverage. As an indicator of bias in standard errors, I computed 95% confidence intervals for parameters in each replication and examined the proportion of estimated confidence intervals that contained the true population parameter. If parameter estimates and standard errors are unbiased, the 95% confidence interval should contain the true population value in 95% of replications. Collins, Schafer, and Kam (2001) consider coverage values that fall below 90% to be problematic.

---

[10] I do not include standardized bias as an outcome in this simulation because a key comparison is between thresholds for even endorsement ($\nu_i = 0$) and sparse endorsement conditions, and standardized bias is not defined for parameters with a true value of zero.

Empirical power. Empirical power was computed by recording the proportion of significant estimates for each parameter according to a standard alpha level of .05. In simulations with properly specified models and a large number of replications, empirical power is a highly accurate estimate of power.

Meta-Models

I analyzed the factors of the simulation using a general linear model (GLM) predicting raw bias to examine interaction and main effects among the design factors. The GLMs used were weighted to account for the fractional factorial design of the simulation study. Two-way design tables for the three factors of the study are provided in Figure 4. Because the GLM has high power to detect significant effects, I used partial η^2 values as an effect size measure to screen for meaningfully large effects. Partial η^2 is computed as

$$\frac{SS_{Between}}{SS_{Between} + SS_{Within}} \quad (15)$$

where $SS_{Between}$ and SS_{Within} are the sums of squared deviations from the mean, representing between-group and within-group variability respectively. Corresponding to a conventional medium effect size (Cohen, 1988), I planned to examine significant effects that produced a partial η^2 value of at least .06. I did not have specific hypotheses about systematic parameter estimate bias in these properly specified GLFA models. Meta-models were only used to investigate factors predicting bias. Because other outcome measures of interest did not vary within cells of the simulation design (i.e., RMSE, MAD), I investigated these outcomes descriptively.

Results

Tables 1 through 4 summarize results of all converged replications for each condition, organized with all results for conditions with medium item loadings in Table 1 and Table 2

(ν=3.85 and ν=4.90 conditions, respectively), and results for high item loading conditions in Table 3 and 4 (ν=3.85 and ν=4.90 conditions, respectively). To simplify the presentation, results are grouped for item loadings and thresholds on items with 50/50 endorsement (λ, ν) and loadings and thresholds for sparse items (λ_{SP}, ν_{SP}). In the following sections I evaluate results for model convergence, parameter estimate bias, efficiency, confidence interval coverage, and empirical power.

Model Convergence and Extreme Values

Model convergence rates are summarized in Table 5. In both baseline conditions (i.e. no sparse items) convergence was 100%, and in all 5% sparseness conditions, convergence was above 99%. Non convergence was generally not an issue and only notable in the 2% sparseness conditions. In the most extreme condition, with ν=4.90 for 8/10 items and $\lambda_i = 1.5$, convergence was 91.2%. Of conditions with ν=4.90, convergence improved slightly with higher item loadings (98.8% with 2% sparseness for 8/10 items and $\lambda_i = 2.0$), but overall convergence rates were high. All convergence failures encountered were due to the estimator reaching a saddle point, meaning a stationary point that is not a local extremum of the likelihood.

Table 1. *Recovery of population generating values when λ = 1.5 with 5% endorsement for sparse items using ML estimation.*

	True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	95% CI	Sig
	No Sparse Items (R=500)												
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.14	0.19	0.41	0.51	0.95	1.00
λ	1.5	1.53	1.51	0.23	0.03	0.23	0.15	0.85	1.18	1.93	2.66	0.96	1.00
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.42	-0.21	0.22	0.47	0.95	0.05
	2/5; 2/5 Sparse Items (R=499)												
ψ_{12}	0.3	0.30	0.30	0.08	0.00	0.08	0.05	0.07	0.15	0.41	0.55	0.94	0.95
λ	1.5	1.54	1.50	0.31	0.04	0.31	0.18	0.74	1.11	2.10	3.45	0.95	1.00
λ_{SP}	1.5	1.60	1.52	0.53	0.10	0.54	0.30	0.35	0.92	2.52	5.73	0.96	0.99
ν	0	0.00	0.00	0.13	0.00	0.13	0.08	-0.42	-0.20	0.21	0.46	0.96	0.04
ν_{SP}	3.85	4.02	3.87	0.70	0.17	0.72	0.34	2.75	3.24	5.25	9.78	0.95	1.00
	4/5; 0/5 Sparse Items (R=500)												
ψ_{12}	0.3	0.30	0.29	0.09	0.00	0.09	0.06	0.06	0.15	0.45	0.61	0.94	0.93
λ	1.5	1.56	1.51	0.41	0.06	0.34	0.17	0.65	1.11	2.05	6.81	0.95	0.99
λ_{SP}	1.5	1.58	1.49	0.58	0.08	0.58	0.31	0.25	0.86	2.57	5.76	0.96	0.97
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.46	-0.21	0.21	0.89	0.95	0.04
ν_{SP}	3.85	4.01	3.85	0.75	0.16	0.76	0.35	2.77	3.19	5.32	10.59	0.95	0.99
	4/5; 4/5 Sparse Items (R=498)												
ψ_{12}	0.3	0.30	0.30	0.11	0.00	0.11	0.07	-0.06	0.12	0.50	0.68	0.93	0.80
λ	1.5	1.70	1.52	0.78	0.20	0.80	0.32	0.57	0.95	2.94	8.64	0.95	0.91
λ_{SP}	1.5	1.58	1.50	0.58	0.08	0.58	0.31	0.15	0.82	2.55	7.03	0.95	0.97
ν	0	0.00	0.00	0.15	0.00	0.15	0.09	-0.84	-0.23	0.23	1.14	0.95	0.05
ν_{SP}	3.85	4.01	3.87	0.74	0.16	0.75	0.37	2.76	3.19	5.26	11.92	0.94	0.99

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5^{th} and 95^{th} quantile estimates, 95% CI is the coverage for the 95% confidence interval, and Sig is the proportion of significant estimates.

Table 2. *Recovery of population generating values when λ = 1.5 with 2% endorsement for sparse items using ML estimation.*

	True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	95% CI	Sig
	No Sparse Items (repeated from previous table for reference, R=500)												
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.14	0.19	0.41	0.51	0.95	1.00
λ	1.5	1.53	1.51	0.23	0.03	0.23	0.15	0.85	1.18	1.93	2.66	0.96	1.00
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.42	-0.21	0.22	0.47	0.95	0.05
	2/5; 2/5 Sparse Items (R=495)												
ψ_{12}	0.3	0.29	0.29	0.08	-0.01	0.08	0.06	0.07	0.16	0.42	0.53	0.94	0.96
λ	1.5	1.55	1.50	0.33	0.05	0.33	0.19	0.74	1.10	2.16	3.38	0.96	1.00
λ_{SP}	1.5	1.90	1.53	2.41	0.40	2.41	0.44	-0.29	0.68	3.85	56.09	0.96	0.71
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.52	-0.22	0.20	0.51	0.95	0.04
ν_{SP}	4.9	5.79	4.95	5.38	0.89	5.30	0.58	3.35	4.03	9.03	143.2	0.94	0.94
	4/5; 0/5 Sparse Items (R=489)												
ψ_{12}	0.3	0.29	0.28	0.11	-0.01	0.11	0.07	-0.04	0.12	0.47	0.74	0.92	0.80
λ	1.5	1.64	1.52	0.59	0.14	0.44	0.18	0.31	1.12	2.68	5.49	0.95	0.91
λ_{SP}	1.5	2.29	1.47	5.10	0.79	5.14	0.47	-1.51	0.44	4.05	69.52	0.94	0.52
ν	0	0.00	0.00	0.14	0.00	0.14	0.09	-0.65	-0.23	0.22	0.55	0.96	0.04
ν_{SP}	4.9	6.76	4.90	11.17	1.86	11.29	0.58	3.35	3.92	9.18	154.9	0.91	0.92
	4/5; 4/5 Sparse Items (R=458)												
ψ_{12}	0.3	0.30	0.28	0.16	0.00	0.16	0.10	-0.12	0.07	0.58	0.98	0.89	0.54
λ	1.5	2.21	1.69	1.46	0.71	1.61	0.65	0.19	0.74	4.95	9.42	0.93	0.41
λ_{SP}	1.5	3.29	1.46	55.94	1.79	39.96	0.49	-2259	0.45	3.99	1751	0.94	0.50
ν	0	0.00	0.00	0.19	0.00	0.19	0.10	-1.23	-0.31	0.28	0.97	0.97	0.03
ν_{SP}	4.9	11.92	4.89	133.9	7.02	93.59	0.61	3.34	3.90	9.16	5792	0.91	0.91

Note. Section in gray is repeated from previous table to facilitate comparison. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5th and 95th quantile estimates, 95% CI is the coverage for the 95% confidence interval, and Sig is the proportion of significant estimates.

Table 3. *Recovery of population generating values when λ = 2 with 7.5% endorsement for sparse items using ML estimation.*

	True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	95% CI	Sig
No Sparse Items (R=500)													
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.13	0.21	0.40	0.50	0.94	1.00
λ	2	2.03	2.00	0.28	0.03	0.28	0.18	1.23	1.61	2.53	3.67	0.96	1.00
ν	0	0.01	0.00	0.15	0.01	0.15	0.10	-0.50	-0.24	0.25	0.52	0.95	0.05
2/5; 2/5 Sparse Items (R=500)													
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.12	0.20	0.41	0.49	0.94	1.00
λ	2	2.04	2.00	0.34	0.04	0.34	0.21	1.19	1.56	2.65	4.40	0.96	1.00
λ_{SP}	2	2.11	2.01	0.65	0.11	0.65	0.30	0.98	1.42	3.07	16.73	0.96	1.00
ν	0	0.00	0.00	0.15	0.00	0.15	0.10	-0.64	-0.24	0.25	0.58	0.95	0.05
ν_{SP}	3.85	4.01	3.87	0.85	0.16	0.84	0.34	2.79	3.22	5.19	24.79	0.96	1.00
4/5; 0/5 Sparse Items (R=499)													
ψ_{12}	0.3	0.30	0.30	0.07	0.00	0.07	0.04	0.08	0.19	0.42	0.51	0.95	0.99
λ	2	2.07	2.01	0.40	0.07	0.37	0.19	1.05	1.59	2.68	6.00	0.96	1.00
λ_{SP}	2	2.07	2.01	0.49	0.07	0.50	0.29	0.70	1.4.	2.96	5.02	0.95	1.00
ν	0	0.00	0.00	0.15	0.00	0.15	0.10	-0.66	-0.25	0.24	0.71	0.96	0.04
ν_{SP}	3.85	3.95	3.85	0.60	0.10	0.61	0.35	2.71	3.17	5.00	8.14	0.94	1.00
4/5; 4/5 Sparse Items (R=500)													
ψ_{12}	0.3	0.29	0.30	0.08	-0.01	0.08	0.05	0.06	0.16	0.42	0.51	0.95	0.95
λ	2	2.27	2.07	0.79	0.27	0.82	0.36	1.03	1.43	3.76	7.61	0.96	0.97
λ_{SP}	2	2.07	2.01	0.51	0.07	0.51	0.30	0.70	1.39	2.96	5.85	0.95	1.00
ν	0	0.01	0.00	0.16	0.01	0.16	0.11	-0.59	-0.24	0.27	0.70	0.97	0.03
ν_{SP}	3.85	3.97	3.88	0.62	0.12	0.63	0.36	2.64	3.18	5.08	9.57	0.95	1.00

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5^{th} and 95^{th} quantile estimates, 95% CI is the coverage for the 95% confidence interval, and Sig is the proportion of significant estimates.

Table 4. *Recovery of population generating values when λ = 2 with 3.5% endorsement for sparse items using ML estimation.*

	True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	95% CI	Sig
	No Sparse Items (repeated from previous table for reference, R=500)												
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.13	0.21	0.40	0.50	0.94	1.00
λ	2	2.03	2.00	0.28	0.03	0.28	0.18	1.23	1.61	2.53	3.67	0.96	1.00
ν	0	0.01	0.00	0.15	0.01	0.15	0.10	-0.50	-0.24	0.25	0.52	0.95	0.05
	2/5; 2/5 Sparse Items (R=498)												
ψ_{12}	0.3	0.30	0.31	0.07	0.00	0.07	0.05	0.11	0.19	0.40	0.50	0.96	0.99
λ	2	2.05	2.00	0.36	0.05	0.37	0.22	1.23	1.53	2.69	4.15	0.95	1.00
λ_{SP}	2	2.31	2.02	1.46	0.31	1.42	0.40	0.62	1.30	4.02	29.70	0.96	0.93
ν	0	0.00	0.00	0.15	0.00	0.15	0.10	-0.69	-0.24	0.25	0.60	0.96	0.04
ν_{SP}	4.9	5.47	4.97	2.47	0.57	2.42	0.57	3.49	3.99	8.12	53.14	0.95	0.95
	4/5; 0/5 Sparse Items (R=491)												
ψ_{12}	0.3	0.30	0.29	0.07	0.00	0.07	0.05	0.10	0.18	0.43	0.56	0.97	0.98
λ	2	2.14	2.03	0.58	0.14	0.45	0.22	0.94	1.58	2.98	6.16	0.96	0.95
λ_{SP}	2	2.27	1.98	2.10	0.27	1.96	0.41	0.66	1.19	3.77	40.98	0.94	0.95
ν	0	0.01	0.01	0.16	0.01	0.16	0.10	-0.59	-0.25	0.26	0.93	0.95	0.05
ν_{SP}	4.9	5.45	4.88	4.00	0.55	3.69	0.55	3.32	3.90	7.83	82.27	0.93	0.97
	4/5; 4/5 Sparse Items (R=494)												
ψ_{12}	0.3	0.30	0.29	0.10	0.00	0.1	0.07	0.04	0.15	0.47	0.63	0.93	0.89
λ	2	2.65	2.18	1.46	0.65	1.57	0.62	0.57	1.28	5.32	12.82	0.95	0.69
λ_{SP}	2	2.29	1.99	2.55	0.29	2.29	0.40	0.21	1.18	3.70	95.50	0.95	0.95
ν	0	-0.01	-0.01	0.19	-0.01	0.19	0.11	-0.88	-0.31	0.29	0.76	0.97	0.03
ν_{SP}	4.9	5.49	4.91	5.02	0.59	4.46	0.56	3.26	3.92	7.60	194.3	0.94	0.97

Note. Section in gray is repeated from previous table to facilitate comparison. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5th and 95th quantile estimates, 95% CI is the coverage for the 95% confidence interval, and Sig is the proportion of significant estimates.

Table 5. *Convergence rates and number of converged solutions without extreme parameter estimates in each condition.*

		$\lambda_i = 1.5$					
		ν=3.85			**ν=4.90**		
	No Sparse	**2/5 & 2/5**	**4/5 & 0/5**	**4/5 & 4/5**	**2/5 & 2/5**	**4/5 & 0/5**	**4/5 & 4/5**
Converged (*max*=500)	500	499	500	498	495	489	458
Given converged, all estimates in range: ν_i [-15,15], λ_i [-8,8]	500	499	500	497	462	445	376

		$\lambda_i = 2.0$					
		ν=3.85			**ν=4.90**		
	No Sparse	**2/5 & 2/5**	**4/5 & 0/5**	**4/5 & 4/5**	**2/5 & 2/5**	**4/5 & 0/5**	**4/5 & 4/5**
Converged (*max*=500)	500	500	499	500	498	491	494
Given converged, all estimates in range: ν_i [-15,15], λ_i [-8,8]	500	499	499	500	479	483	467

Parameter estimates were not automatically fixed to boundary values during estimation (e.g., thresholds to +/- 15). Parameters were occasionally fixed to their estimated value (i.e. no standard error is reported); this occurred in less than 1% of replications. Although convergence criteria were technically satisfied for most solutions, extreme values were frequently reported in conditions with sparseness. While any single set of thresholds for extreme values is necessarily arbitrary, it is illustrative to consider how frequently extreme estimates arose. As one measure of the number of extreme solutions, included in Table 5 are the numbers of replications that converged with threshold estimates between [-15,15] and item loadings between [-8,8]. In practice, estimates this large would be considered suspicious. For conditions with sparseness, extreme estimates were rare in the conditions with ν=3.85. In conditions with ν=4.90 and medium item loadings, solutions with extreme values were more common, for example 445/500 (89%) with 4 sparse items on one factor and 376/500 (75%) with 4 sparse items on both factors.

Examining these extreme values further, I investigated loading estimates that corresponded to extreme threshold estimates and vice versa. I also looked for characteristics in the simulated data that were associated with extreme estimates. For the chosen ranges of extreme values, extreme thresholds and item loadings were almost equally common (49% extreme loadings, 51% extreme thresholds). Of note, high loadings were commonly observed with high thresholds: 75% of items with extreme threshold estimates also had extreme loading estimates. These ICCs were essentially step functions shifted high on the range of the latent variable. Low item loadings, implying essentially flat ICCs, were relatively rare and generally did not correspond to high threshold estimates. This means that items with low endorsement were generally estimated to be strongly related to the latent variable. High threshold and loading estimates often corresponded to observed endorsement of about 5 cases (1% for sample size of 500).

In summary, although convergence rates to proper ML solutions were high, there were a large proportion of replications with improbably high estimates in conditions with severe sparseness. These solutions would likely by met with suspicion in practice. All technically converged solutions were included with the results in Tables 1-4 and are included in subsequent sections describing bias, efficiency, coverage, and empirical power. Robust statistics (e.g. medians and MAD) are useful for considering performance without undue influence from extreme observations; however note that "extreme" estimates occurred frequently in conditions with high sparseness.

Raw Bias

Meta-model results predicting raw bias for each parameter are summarized in Table 6. There was no evidence of meaningful, systematic bias in any of the conditions studied, and none

of the simulation factors in the model predicted bias with a medium effect size or larger. All partial η^2 values were less than .01. Estimates from all converged solutions are included in the meta-model results shown. I also fitted the models excluding extreme values and the results did not differ meaningfully. The lack of systematic bias can also be seen in Tables 1 through 4. Some mean estimates are biased due to extreme values; however median estimates were very close to the true values.

Table 6. *Results from meta-models fitted to raw bias of estimates using ML estimation*

	Correlation (ψ_{12})			Loading (λ)			Threshold (ν)		
Factor (df)	***F* (6907)**	***p***	**η^2**	***F* (69195)**	***p***	**η^2**	***F* (69195)**	***p***	**η^2**
Loading (1)	2.17	.14	<.001	4.35	.04	<.001	12.60	<.001	<.001
Threshold (2)	2.86	.06	<.001	7.55	.00	<.001	11.16	<.001	<.001
Pattern (3)	2.41	.07	.001	2.57	.05	<.001	3.97	.008	<.001
*Load*Sparse* (2)	0.83	.43	<.001	3.04	.05	<.001	8.12	<.001	<.001
*Sparse*Pattern* (2)	1.55	.21	<.001	4.81	.01	<.001	11.57	<.001	<.001
*Load*Pattern* (3)	1.74	.16	<.001	2.18	.09	<.001	7.41	<.001	<.001
Sparse Item (1)				1.99	.16	<.001	10.18	.001	<.001

Note. GLM results for fractional design. η^2 denotes partial η^2. Denominator degrees of freedom shown below *F* in ( ). Loading is value of λ (1.5 or 2), Threshold is value of ν (0.0, 3.85, 4.9), Pattern is distribution of sparse items across factors, and Sparse Item is an item-level main effect for loadings or thresholds on sparse items. Meta-models include all converged solutions and do not exclude replications with extreme values.

Efficiency

Average RMSE and MAD for each parameter type in each condition are also shown in Tables 1 through 4. Because RMSE and MAD are summary statistics for parameters in each cell of the design (i.e., they do not vary within condition), I did not fit meta-models for measures of efficiency. Instead, I describe differences in RMSE and MAD qualitatively. Efficiency for the estimated correlation between factors ψ_{12} was identical for both baseline conditions (RMSE = 0.06, MAD= 0.04), but RMSE/MAD for estimates of item loadings λ and thresholds ν was

slightly higher in the medium item loading baseline condition (e.g. RMSE = .13 versus .15 for all ν). As expected, sparseness lead to decreased efficiency for all parameter estimates. In general, RMSE and MAD increased with higher thresholds (ν=4.90 versus ν=3.85) and with more sparse items (4 versus 8). For example, the loss of efficiency from baseline to the high sparseness condition (8/10 items sparse) was an increase in RMSE from .06 to .08 (33%; λ=2) or from .06 to .11 (83%; λ=1.5) for the estimated correlation between factors, when sparseness was at the .05 level. This compares to a 167% increase in RMSE for the correlation estimate from the baseline to high sparseness condition at the ν=4.90 level (λ=1.5).

In terms of RMSE, efficiency was worse for the uneven sparseness conditions, for example RMSE rose 33% from .33 to .44 for item loadings on non-sparse items (ν=4.90, λ =1.5) and 113% from 2.41 to 5.14 for loadings on sparse items, however the differences in terms of MAD were less striking (.19 to .18, λ; .44 to .47, λ_{SP}), reflecting that extreme values were more common in the uneven sparseness conditions, but median efficiency was comparable.

Confidence Interval Coverage

For nearly all conditions studied, 95% confidence interval coverage was between 94-96%. The range widened slightly in conditions with ν=4.90 for 4/5 items on a single factor (93-97% and 91-96% for high and medium item loadings, respectively) and in high sparseness conditions (93-97% for ν=3.85 on 8/10 items; 89-97% for ν=4.90 on 8/10 items). These results suggest that confidence intervals were not substantially biased by sparse items.

Empirical Power

Empirical power to detect significant effects (ψ_{12}, λ, λ_{SP}, ν_{SP}) was lower in conditions with sparseness. This effect differed by threshold, with lower power for ν=4.90 versus ν=3.85. Empirical power for all parameters was higher when λ =2, for example 80% versus 95% of correlation estimates were significant in the medium versus high item loading conditions with

ν=3.85. Focusing on item loadings and the correlation between factors, empirical power was 80% or above for all conditions with ν=3.85. For conditions with λ =2, power fell below 80% only when ν=4.90 for 8/10 items (e.g., .69 for λ). Empirical power was lowest with ν=4.90 in models with λ=1.5. For example, 54% significant correlation estimates with sparseness for 8/10 items and 80% with uneven sparseness on 4/5 items on one factor (empirical power was higher, 96%, for sparseness on 2/5, 2/5 items).

Summary of Study 1 Results

Taken together, the results of study 1 showed that ML performed as expected by theory under conditions of sparseness. There was no evidence of biased estimates or confidence intervals in these properly specified models. In general, convergence problems were infrequent in the conditions studied; however improbably extreme estimates were common even in technically converged solutions. Lower parameter estimate efficiency and decreased empirical power to detect significant effects were the main effects of sparseness. As expected, these effects were more severe with lower item loadings (λ=1.5), with more extreme thresholds (ν=4.90), and with a majority of sparse items on one or both factors. Given these results, it is clear that ML estimation begins to break down in conditions with a high proportion of sparse items. If researchers wish to make inferences from a model with a high proportion of sparse items, they are likely to obtain suspicious parameter estimates and to lack power to detect significant effects.

From these results, I chose three conditions from Study 1 to investigate Bayesian estimation for GLFA models with sparse indicators in Study 2. Because I was interested in studying Bayesian estimation where ML performance is unacceptable, I chose two conditions where ML performance was worst. Specifically, from the models with λ=1.5, I chose the most extreme condition with 8/10 items having ν=4.90 (2% marginal endorsement), and the condition

with 4/10 items on a single factor having ν=4.90. I also selected a baseline condition as a comparison where ML performs well, with λ=1.5.

CHAPTER 3: STUDY 2 – BAYESIAN ESTIMATION

In Study 2 I evaluated Bayesian estimation for GLFA models with sparse, binary indicators. I compared Bayesian estimation to ML estimation on the same data sets for a subset of three conditions identified in Study 1: one where ML performs well and two where it performs poorly. I evaluated the performance of Bayesian estimation for these models under a variety of different priors.

I performed Bayesian estimation for subsets of replications identified in Study 1 using the Stan programming language implemented in R, using Hamiltonian Monte Carlo and the No U-turns sampler. The Stan programming language can be used with many interfaces, including R software, but is coded in C++ for efficiency. To write a Stan program, users define the statistical model and priors for each parameter, and the program adapts the sampling algorithm while still allowing a reasonable amount of flexibility in model and prior specification and oversight over the sampling. Using HMC in Stan, there is no computational advantage to choosing conjugate priors. Stan allows users to specify improper priors (i.e. integral of prior is infinity) and diagnoses improper posteriors automatically when parameters overflow to infinity during simulation (Carpenter et al., 2015). In contrast to other statistical programs that offer Bayesian estimation, using the Stan programing language allows the analyst flexibility in model and prior choice, oversight of MCMC convergence, and fast computation. The Stan program used to specify the GLFA model is provided in Appendix B.

Prior Specification

Because it is risky to rely on default priors (e.g., Kass & Wasserman, 1996), a central aim of Study 2 was to evaluate different priors for the GLFA model and HMC/NUTS sampler. For a range of priors, I evaluated model convergence and bias and overall accuracy of parameter estimates, and I evaluated the sensitivity of posterior inferences based on prior input. I evaluated three general types of priors. First, I included a condition with flat priors for the intercepts and item loadings. These priors were normal with extremely high variance, essentially uniform on the admissible range for all parameters:

$$\begin{aligned} &\pi(\lambda_i) \sim N(0, \sigma = 1000) \\ &\pi(\nu_i) \sim N(0, \sigma = 1000) \\ &\pi(\psi_{21}) \sim U[-1,1] \end{aligned} \qquad 16$$

Second, I evaluated moderately concentrated priors, with increased probability for plausible values.

$$\begin{aligned} &\pi(\lambda_i) \sim N(0, \sigma = 10) \\ &\pi(\nu_i) \sim N(0, \sigma = 10) \\ &\pi(\psi_{21}) \sim U[-1,1] \end{aligned} \qquad 17$$

Note that the moderately concentrated prior is general for applications in psychology. A third prior specification was more concentrated and constrained all factor loadings and the covariance between factors as positive:

$$\begin{aligned} &\pi(\lambda_i) \sim N(\lambda_i, \sigma = 3.57), \qquad \lambda_i \geq 0 \\ &\pi(\nu_i) \sim N(\nu_i, \sigma = 3.57) \\ &\pi(\psi_{21}) \sim U[0,1] \end{aligned} \qquad 18$$

The variance in the concentrated priors specifies 95% prior probability that item intercepts lie within [-7, 7], and 97.5% prior probability that factor loadings lie within [0, 7]. These restrictions more heavily limit the posterior for conditions with sparse data.

Posterior Simulation

The simulations were run using a large computing cluster for UNC Chapel Hill researchers located on UNC's campus. For each condition and prior, replications were submitted in parallel in sets of 20. Each submission was allowed to run for 7 days; submissions that did not complete in this time were terminated.

The method of identification used in Study 1 (setting each factor mean and variance to 0 and 1, respectively), although only locally identifying the model (Bollen & Bauldry, 2010) lead to all solutions with a majority of positive factor loadings (i.e. sign indeterminacy was not an issue using ML estimation for this model and data in Mplus). However, sign indeterminacy does become an issue using the same scaling in the Bayesian framework. Specifically, solutions with either all positive factor loadings or all negative factor loadings are log-likelihood equivalent. Similarly, a solution with all positive loadings for one factor and all negative loadings for the other factor, and a negative covariance between factors, is equivalent. This sign indeterminacy can be resolved using the alternate scaling: by setting a single indicator to 1 for each factor and estimating the variance of each factor. Using Bayesian estimation in Stan, choice of scaling had an impact on the efficiency of posterior simulation. Although scaling to an indicator has the advantage of solving sign indeterminacy, the efficiency of posterior simulation greatly decreased using this scaling. Specifically, for the baseline condition with no sparse items and moderate priors, scaling to a latent factor resulted in small estimated effective sample size (e.g., less than 10) for multiple parameters in approximately 10% of replications after 10,000 replications (half warm-up). Scaling by setting the factor variances to 1, however, resulted in higher estimated effective sample size (e.g., minimum 371) and sampling was twice as fast.

In order to maximize efficiency in posterior simulation, the more efficient scaling was used for Bayesian estimation (setting latent factor variances to 1), and "flipped" solutions were

post-processed after estimation to the preferred scaling for inference. Post-processing to an inferential parameterization has been used in a similar modeling context with continuous indicators (Ghosh & Dunson, 2009). In pilot simulations, I did not encounter any replications where a single chain switched from one solution (e.g. all positive loadings) to an opposite solution (e.g. all negative loadings), however estimating multiple chains for the same data did result in multiple solutions. Different solutions for each chain was also manifested in high estimated $\hat{R}$. To avoid opposite solutions within replication, a single chain with 20,000 iterations (half warm-up) was run for each replication, and $\hat{R}$, which is calculated on split chains, was monitored for each chain to determine if any chains switched between solutions (i.e., $\hat{R}$ above 1 should signal switching within a chain).

Evaluation Criteria

Convergence Assessment. Convergence in an MCMC framework is theoretically guaranteed after infinite samples under certain assumptions, but with a finite number of MCMC samples it is impossible to guarantee convergence. Whereas ML has clear replications where models do not converge, for Bayesian estimation there are only degrees of confidence in convergence. Convergence was assessed by monitoring the estimated potential scale reduction factor and effective sample size estimates. Stan computes the potential scale reduction factor on split chains (Stan Development Team, 2015), so it is possible to monitor $\hat{R}$ even for a single chain. I also monitored MCMC plots for a small sample of replications.

For this simulation, effective sample size of at least 100 for all parameters was considered sufficient to interpret results for each replication. Replications with effective sample size below 100 for any parameter were not included in results tables. In practice, higher effective sample size may be preferable for any single replication (e.g. 1000 for increased precision for interpreting posterior intervals; see Gelman et al., 2013, p. 267). However it is not currently

possible to automate sampling until a desired effective sample size is reached using the HMC/NUTS algorithm in Stan.

Evaluation of bias, efficiency, coverage, and empirical power. The performance of Bayesian estimation under each prior specification was evaluated as in Study 1 based on posterior medians and posterior intervals. I assessed the performance of Bayesian estimation in terms of bias[11], using a meta-model to test for systematic bias as a function of condition and prior specification. The efficiency (RMSE and MAD) of estimates, credible interval coverage, and empirical power are presented in subsequent sections. For each outcome, I also compare the performance of Bayesian estimation to the results using ML estimation. Finally, based on the results of Bayesian estimation for different prior specifications and encountered difficulties with MCMC estimation, I detail the advantages and potential limitations of Bayesian estimation for GLFA models with sparse, categorical indicators.

Results

Convergence

For all conditions reported here, $\hat{R}$ was 1 for all parameters. Effective sample sizes for each condition, prior, and parameter are summarized in Table 7. Sampling did not complete within the time limit of 7 days for conditions with sparse items using flat priors, so results for these conditions are not reported. For the baseline condition with no sparse items, effective sample size was above 100 for all parameters in 498 replications using flat priors (99.6%), and in 100% of replications using moderate or concentrated priors. The median and 5th quantile of effective samples was similar across all prior specifications in the baseline condition.

For conditions with sparseness, effective sample size differed substantially using moderate versus concentrated priors. Whereas 10,000 post-warmup iterations was sufficient to

[11] Although parameters are not considered constant in Bayesian analysis, it is common to evaluate Bayesian methods using frequentist operating characteristics (e.g. Gelman et al., 2013, Ch. 4.4).

achieve 100 effective samples per parameter for most replications using concentrated priors, effective sample size was much lower using moderate priors. To obtain a larger number of replications with sufficient minimum effective sample size, I repeated the simulation for conditions with sparseness and moderate priors with 40,000 post-warmup iterations. Minimum effective sample size remained less than 100 using moderate priors for 28% and 42% of replications with 4/10 and 8/10 sparse items, respectively.

Table 7. *Median, minimum, and 5th quantile number of effective samples for each condition, prior, and parameter.*

No Sparse Items

	Flat (R=498) 10k iterations			**Moderate (R=500)** 10k iterations			**Concentrated (R=500)** 10k iterations		
	Med NEff	**Min NEff**	**.05Q NEff**	**Med N Eff**	**Min N Eff**	**.05Q N Eff**	**Med N Eff**	**Min N Eff**	**.05Q N Eff**
ψ_{12}	3038	14	2362	3075	1670	2389	2909	1943	2299
λ	3423	3	2319	3432	371	2367	3500	569	2463
ν	3905	5	2807	3967	848	2831	3897	1794	2794

4/5; 0/5 Sparse Items

	Moderate (R=361) 40k iterations			**Concentrated (R=491)** 10k iterations		
ψ_{12}	1500	3	11	930	80	210
λ	6083	5	29	4225	50	220
λ_{SP}	2073	3	14	2271	179	727
ν	9470	4	48	5055	139	1494
ν_{SP}	2329	3	26	2681	222	879

4/5; 4/5 Sparse Items

	Moderate (R=292) 40k iterations			**Concentrated (R=466)** 10k iterations		
ψ_{12}	653	3	11	488	25	111
λ	356	4	14	258	58	125
λ_{SP}	1408	3	14	2058	112	690
ν	2097	5	48	1885	184	479
ν_{SP}	1614	3	249	2397	180	843

Posterior simulation (20,000 total iterations) for sets of 20 replications completed in approximately 10 hours or less running on a single Intel Xeon Processor (2.93 GHz). This means that estimation for single replications could be expected to run in about 30 minutes on a personal computer, for this model and sample size. Computational time was generally faster in baseline conditions relative to conditions with sparse items and for more concentrated priors.

Because convergence is very different in the Bayesian and ML frameworks, it is problematic to directly compare "convergence rates" from the two frameworks. Even though effective sample size was lower than the specified cutoff for 9 and 34 replications with 4/5 sparse items on one or both factors, respectively, sampling for more iterations could be done to achieve the desired effective sample size. In these conditions with a high number of sparse items, using a concentrated prior specification, it is possible to examine solutions in cases where an estimate was not available using ML estimation, either by sampling for more iterations or by inspecting solutions with lower effective sample size.[12]

In Study 1, using ML estimation, extreme values were frequently encountered in technically converged replications (unrelated to systematic bias) in the sparseness conditions studied here using Bayesian estimation. However, in this study using Bayesian estimation, extreme values were related to prior specification and bias in parameter estimates. Therefore, I save treatment of extreme values from Bayesian estimation for the next section on bias in parameter estimates.

Raw Bias

Meta-models predicting raw bias for each parameter are summarized in Table 8. Only replications with effective sample size greater than 100 for all parameters were analyzed. Because posterior sampling failed to complete in the time allotted for conditions with sparse

[12] For the concentrated prior specification, I separately examined results for all replications, including replications with effective sample size below my preferred cutoff. The results did not differ meaningfully for any outcome.

items using flat priors, only results for moderate and concentrated priors were included in the meta-models.

There was a substantial effect of sparseness pattern on bias in correlation estimates ($F(2,2604) = 151$, $p<.0001$, $\eta^2 = .10$). Bias in item loadings depended on a number of interactions between factors. There were no factors predicting substantial bias in threshold estimates. To understand these patterns, I refer to the summarized results for each condition and prior (Tables 9 and 10).

Table 8. *Results from meta-models fitted to raw bias of estimates using Bayesian estimation for moderate and concentrated priors*

	Correlation (ψ_{12})			Loading (λ)			Threshold (ν)		
Factor (df)	***F* (2604)**	***p***	**η^2**	**F (26091)**	***p***	**η^2**	**F (26091)**	***p***	**η^2**
Pattern (2)	151.00	<.0001	.10	3587.38	<.0001	.22	0.40	.67	<.0001
Prior (1)	17.58	<.0001	.001	658.99	<.0001	.03	1155.08	<.0001	.042
Pattern Prior* (2)	5.88	.003	.005	1005.97	<.0001	.07	0.01	.99	<.0001
Sparse Item (1)				5946.87	<.0001	.19	679.73	<.0001	.025
*Sparse Item *Prior* (1)				2489.96	<.0001	.09	0.33	.568	<.0001
*Sparse Item *Pattern* (1)				1225.89	<.0001	.05	617.45	<.0001	.02

Note. η^2 denotes partial η^2. Denominator degrees of freedom shown below *F* in ( ). Pattern is distribution of sparse items across factors (none, 4/10, 8/10), and Sparse Item is an item-level effect for loadings or thresholds on sparse items. Meta-models include all solutions with effective sample size ≥ 100 for all parameters.

Table 9. *Recovery of population generating values using Bayesian estimation for baseline condition*

	True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	95% CI	Sig
Flat Prior (R=498)													
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.14	0.19	0.40	0.50	0.96	1.00
λ	1.5	1.56	1.54	0.24	0.06	0.25	0.16	0.86	1.20	1.98	2.83	0.95	1.00
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.43	-0.21	0.22	0.47	0.95	0.05
Moderate Prior (R=500)													
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.14	0.19	0.40	0.50	0.96	1.00
λ	1.5	1.56	1.54	0.24	0.06	0.25	0.16	0.86	1.19	1.98	2.81	0.95	1.00
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.43	-0.21	0.22	0.48	0.95	0.05
Concentrated Prior (R=500)													
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.14	0.19	0.40	0.50	0.96	1.00
λ	1.5	1.55	1.54	0.24	0.05	0.24	0.16	0.85	1.19	1.96	2.76	0.95	1.00
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.43	-0.21	0.22	0.47	0.95	0.05

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5^{th} and 95^{th} quantile estimates, 95% CI is the confidence coverage for the 95% credible interval, and Sig is the proportion of significant estimates.

Table 10. *Recovery of population generating values using Bayesian estimation with moderate and concentrated priors.*

	True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	95% CI	Sig
4/5; 0/5 Sparse, Moderate Prior (R=361)													
ψ_{12}	0.3	0.25	0.24	0.10	-0.05	0.11	0.06	-0.04	0.10	0.42	0.68	0.84	0.93
λ	1.5	2.56	1.57	2.86	1.06	1.46	0.20	0.47	1.16	10.87	12.41	0.84	1.00
λ_{SP}	1.5	1.79	1.42	1.26	0.29	1.29	0.55	-2.26	0.35	4.58	6.85	0.86	0.90
ν	0	0.00	0.00	0.21	0.00	0.18	0.10	-1.33	-0.28	0.27	1.42	0.95	0.05
ν_{SP}	4.9	5.65	4.90	1.97	0.75	2.10	0.62	3.38	3.95	10.32	13.83	0.88	1.00
4/5; 0/5 Sparse, Concentrated Prior (R=491)													
ψ_{12}	0.3	0.26	0.26	0.09	-0.04	0.10	0.06	0.05	0.13	0.41	0.70	0.93	1.00
λ	1.5	1.76	1.57	0.68	0.26	0.50	0.19	0.41	1.17	3.52	5.27	0.93	1.00
λ_{SP}	1.5	1.39	1.35	0.56	-0.11	0.57	0.38	0.22	0.55	2.39	3.41	0.95	1.00
ν	0	-0.01	0.00	0.14	-0.01	0.14	0.09	-0.67	-0.24	0.22	0.52	0.95	0.05
ν_{SP}	4.9	4.91	4.79	0.71	0.01	0.71	0.44	3.39	3.99	6.26	7.94	0.96	1.00
4/5; 4/5 Sparse, Moderate Prior (R=292)													
ψ_{12}	0.3	0.21	0.19	0.11	-0.09	0.14	0.07	-0.09	0.06	0.42	0.68	0.75	0.79
λ	1.5	6.82	9.20	4.47	5.32	6.95	2.32	-10.96	0.86	11.83	12.64	0.51	0.99
λ_{SP}	1.5	1.79	1.44	1.26	0.29	1.29	0.59	-4.98	0.39	4.49	6.32	0.86	0.89
ν	0	0.00	-0.01	0.42	0.00	0.42	0.17	-1.54	-0.77	0.66	1.44	0.93	0.07
ν_{SP}	4.9	5.66	4.93	1.94	0.76	2.08	0.68	3.36	3.92	10.23	14.22	0.87	1.00
4/5; 4/5 Sparse, Concentrated Prior (R=466)													
ψ_{12}	0.3	0.24	0.23	0.10	-0.06	0.12	0.07	0.04	0.10	0.44	0.72	0.92	1.00
λ	1.5	2.77	2.79	1.17	1.27	1.73	0.97	0.25	0.95	4.65	5.46	0.88	1.00
λ_{SP}	1.5	1.39	1.34	0.57	-0.11	0.58	0.40	0.19	0.54	2.42	3.25	0.95	1.00
ν	0	-0.02	-0.01	0.20	-0.02	0.20	0.12	-0.67	-0.35	0.30	0.63	0.93	0.07
ν_{SP}	4.9	4.93	4.80	0.74	0.03	0.74	0.47	3.34	3.95	6.34	7.97	0.96	1.00

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5^{th} and 95^{th} quantile estimates, 95% CI is the confidence coverage for the 95% credible interval, and Sig is the proportion of significant estimates.

Table 9 summarizes results for the baseline model with no sparse items, organized by prior specification. Results for conditions with sparse items are in Table 10. As in Study 1, results are grouped for item loadings and thresholds on items with 50/50 endorsement (λ, ν) and loadings and thresholds for sparse items (λ_{SP}, ν_{SP}). With no sparse items, there was no evidence of bias in any parameter under any of the three priors studied. Correlation estimates were downwardly biased when the models included sparse items, and this bias was more pronounced with more sparse items. For example, the mean correlation estimate was .26 and .24 (raw bias −.04 and −.06) with 4/10 and 8/10 items sparse, respectively, using concentrated priors.

Factors predicting bias in item loadings included the number of sparse items in the model, prior specification, and whether the item loading was for a sparse item. The effect of number of sparse items differed by prior specification ($F(2,2604) = 1005.97$, $p<.0001$, $\eta^2 = .07$), and the item-level effect of loading on a sparse item also depended on prior specification ($F(1,2604) = 2489$, $p<.0001$, $\eta^2 = .09$). These effects are illustrated in Figure 5, where median estimates for item loadings are plotted for each prior, condition, and for loading on sparse (versus non-sparse) items. The median estimate is only negligibly biased in the condition with 4/10 items sparse. With 8/10 items sparse, bias was substantial for the item loadings on non-sparse items, especially using the moderate prior specification.

Altogether, estimates were more biased using the moderate prior specification than with the more concentrated priors. However, extreme estimates were uncommon. Considering the ranges for what were considered extreme estimates from Study 1, there were no threshold estimates outside of +/- 15; item loadings outside of +/- 8 were only observed using moderate priors with 4/5 items sparse on both factors.

Figure 5. Median estimates of λ depending on condition, prior, and whether item was sparse

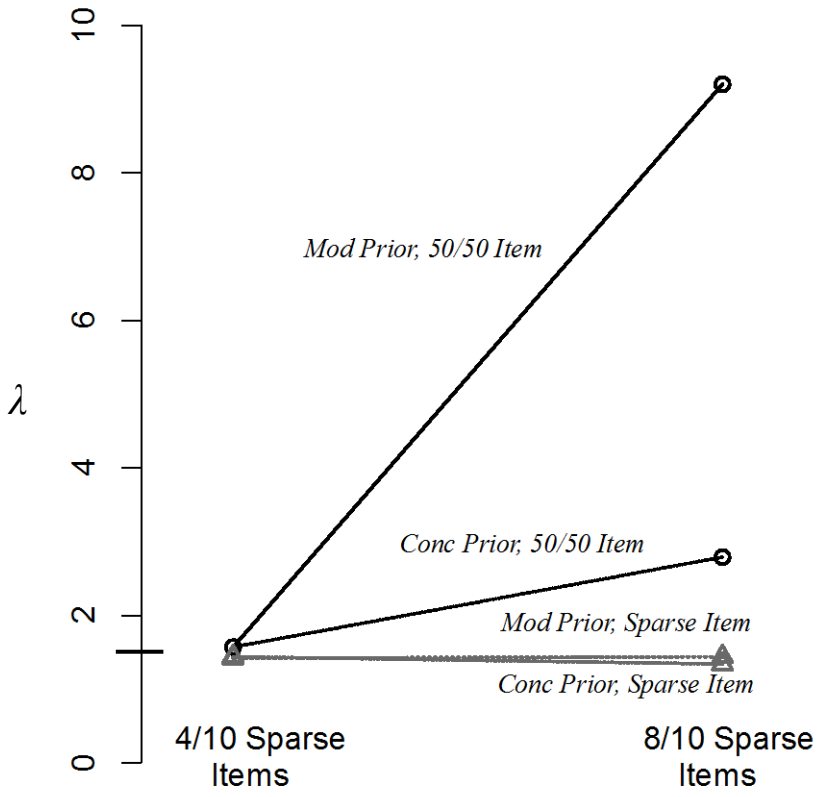


Figure 5. Median estimates for λ for moderate (Mod) and concentrated (Conc) priors in conditions with differing numbers of sparse items and for items with sparse endorsement. The true value of λ is 1.5 and marked on the *y*-axis.

Efficiency

With no sparse items, parameter estimate efficiency as measured by RMSE and MAD was essentially the same using each prior specification; the efficiency of parameter estimates also closely matched efficiency using ML estimation for this baseline condition. As expected, in conditions with sparse items, RMSE and MAD were larger with more sparse items, but smaller with more concentrated priors. As an example of decreased efficiency with more sparse items, using moderate priors, RMSE was .11 with 4 sparse items on one factor and .14 with 4 sparse items on both factors (compared to .06 RMSE in the baseline condition). The substantial

difference observed in efficiency for moderate versus concentrated priors was partially related to bias and partially related to variance. In the high sparseness condition, RMSE for item loadings was 1.29 (sparse item) and 6.95 (non-sparse item) using moderate priors, compared to 0.58 (sparse item) and 1.73 (non-sparse item) for concentrated priors.

Comparing efficiency of estimates across estimators, performance differed by parameter estimate. Figure 6 and Figure 7 compare MAD and RMSE in both sparseness conditions for each parameter using ML estimation and Bayesian estimation with a concentrated prior. For item loadings on non-sparse items, RMSE and MAD was higher using Bayesian estimation. For all other parameters, RMSE and MAD was higher using ML estimation or about equal. Note that different subsets of replications are included in this comparison, because the ML results are restricted to models that converged and Bayesian results are restricted to replications that met the minimum effective sample size for all parameters. Replications that did not converge using ML estimation were not the same replications with below threshold effective sample size using Bayesian estimation.

Credible Interval Coverage

For the baseline condition, coverage was between 95-96% for all parameters and all priors; this aligns with the coverage observed using ML. Coverage fell below 90% for several parameters in the sparseness conditions using moderate priors, as low as 51% coverage for item loadings on the non-sparse items in the high sparseness condition, which was related to high bias for this parameter estimate. With concentrated priors, coverage rates were comparable to those observed for ML estimation for the same conditions: 93-96% versus 91-96% for Bayesian and ML estimation, respectively with 4/5 items sparse on a single factor; 88-96% versus 89-97% for Bayesian and ML estimation, respectively with 4/5 items sparse on both factors.

Figure 6. MAD for ML and Bayesian estimation using concentrated priors for conditions with sparseness

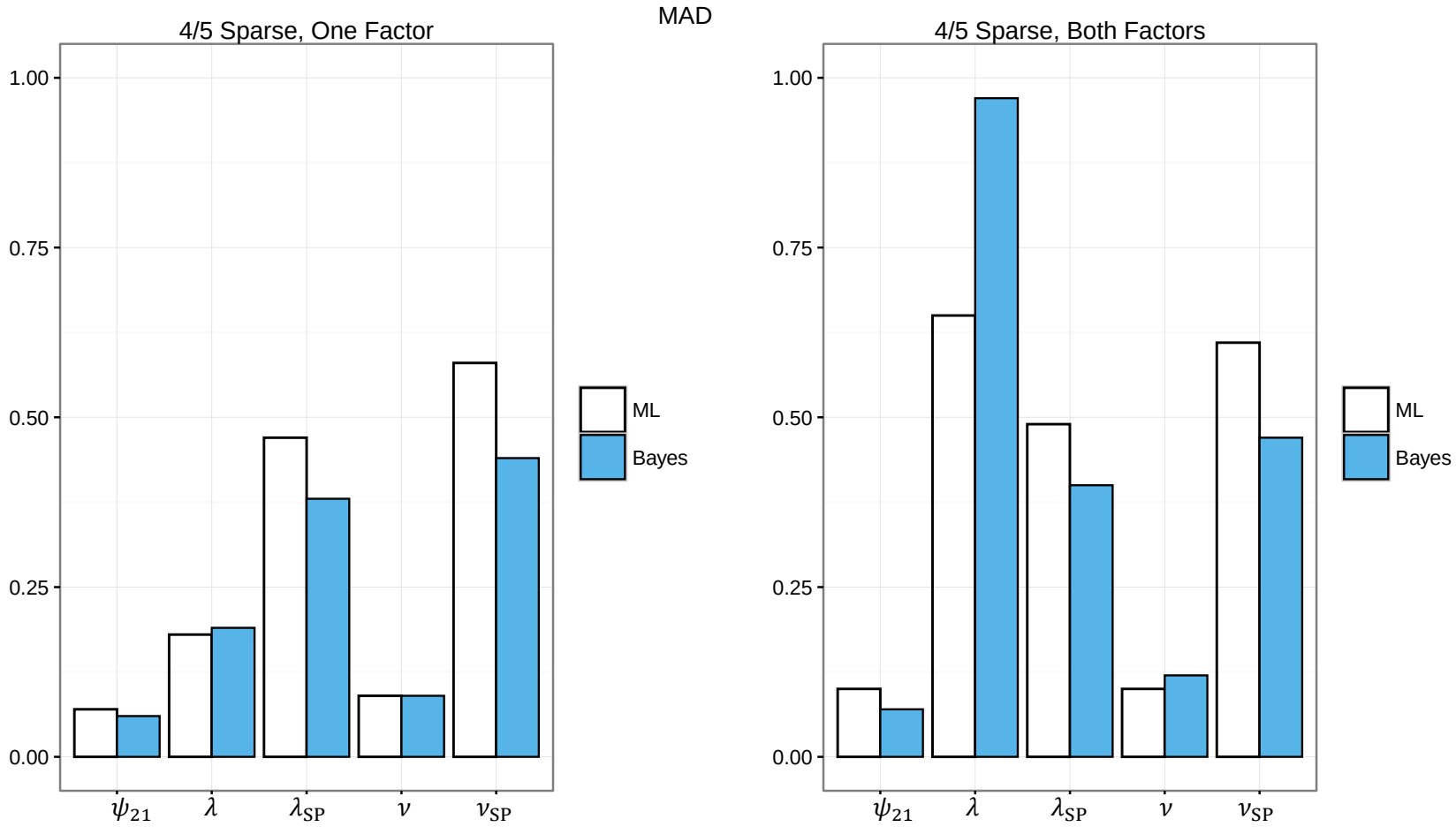


Figure 6. Median absolute deviation for parameter estimates using ML and Bayesian estimation with a concentrated prior specification. Results shown with 4/5 sparse items on one factor (Left) and with 4/5 sparse items on both factors (Right). Note that the results for ML estimation include only converged solutions and results for Bayesian estimation include solutions with above threshold effective sample size for all parameters, so the solution sets do not exactly overlap.

Figure 7. RMSE for ML and Bayesian estimation using concentrated priors for conditions with sparseness

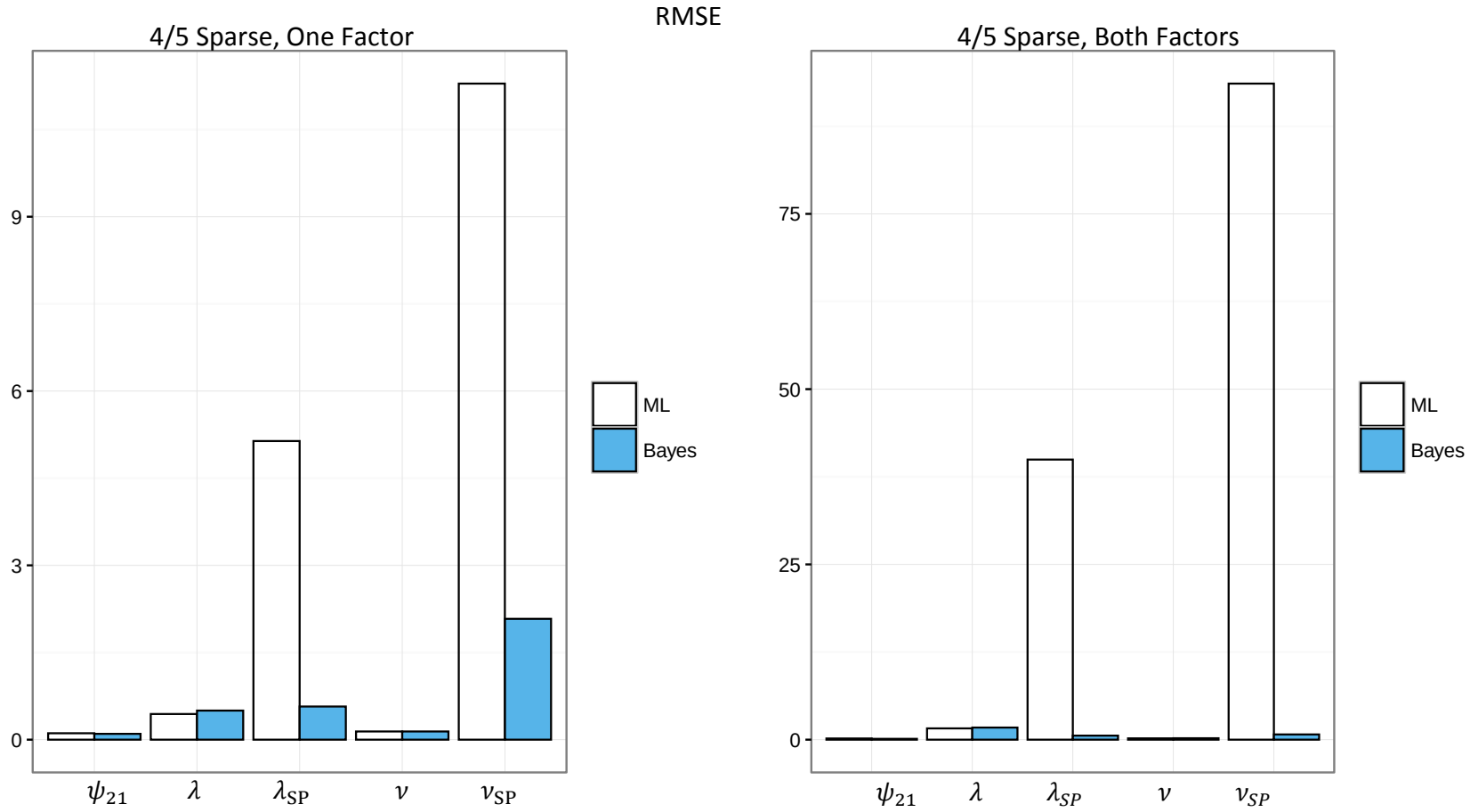


Figure 7. Note that the y-axes are different between plots, due to the extremely large discrepancy in RMSE values for each condition . Root-mean-square-error for parameter estimates using ML and Bayesian estimation with a concentrated prior specification. Results shown with 4/5 sparse items on one factor (Left) and with 4/5 sparse items on both factors (Right). Note that the results for ML estimation include only converged solutions and results for Bayesian estimation include solutions with above threshold effective sample size for all parameters, so the solution sets do not exactly overlap.

Empirical Power

Empirical power for different estimates is summarized in the last column of Table 9 and Table 10 for the baseline condition and conditions with sparse items. In the baseline condition, empirical power was 1.00 for true effects (correlation estimates and factor loadings) using all prior specifications. In the sparseness conditions, empirical power differed by prior specification. With concentrated priors, power to detect true effects was 1.00 for all parameters, matching empirical power in the baseline conditions. Using the moderate prior specification, empirical power differed by parameter, however in all cases empirical power was higher using Bayesian estimation than was observed for ML estimation. For example, power to detect the correlation between factors was 0.80 and 0.54 with 4/5 items sparse on one factor and two factors, respectively using ML estimation. This compares to 0.93 and 0.79 in the same conditions using Bayesian estimation (moderate priors).

Summary of Study 2 Results

Taken together, the results showed that the use of priors in Bayesian estimation can stabilize estimates in GLFA models with sparse, categorical data. The use of a concentrated prior specification eliminated extreme parameter estimates, improved estimate efficiency, and increased empirical power to detect true effects. Results also suggest that Bayesian estimation can be a useful alternative when models do not converge using ML estimation, although more iterations of posterior sampling may be needed to ensure an adequate number of effective samples. The gains in efficiency and empirical power using Bayesian estimation were found to be dependent on prior specification, with concentrated priors offering substantial improvement over more diffuse priors. However, increased overall efficiency and empirical power were tied to a trade-off with overall unbiasedness. Bayesian estimation performs similarly to ML estimation in a baseline condition with a moderate sample size and high endorsement on all items.

CHAPTER 4: DISCUSSION

I have evaluated a method for improving GLFA estimation with sparse, categorical indicators. Prior information about typical parameter values in psychological research is utilized in a Bayesian framework to decrease variability in parameter estimates, eliminate extreme estimates, and improve empirical power to detect true effects. In the first simulation study, I evaluated the performance of ML estimation in a range of GLFA models with sparse indicators. In the second study, I evaluated Bayesian estimation in conditions where ML performs poorly and in a comparison condition where ML performs well. Next, I will discuss how the simulation results align with my hypotheses about the performance of ML and Bayesian estimation for models with sparse indicators and compare the two approaches. Subsequently I will discuss the unique contributions of the present work and summarize my recommendations for applied researchers. I will end by reviewing limitations of the present work and provide recommendations for future research.

Performance of ML Estimation for Sparse Items

Because previous research has suggested that categorical estimation methods break down under conditions of sparseness (e.g., Forero & Maydeu-Olivares, 2009; Rhemtulla et al., 2012; Wirth & Edwards, 2007), I hypothesized that as the extent and severity of sparse items increased, ML estimation would start to break down and fail to reliably produce converged, reasonable solutions. I also hypothesized that efficiency would decrease in conditions with sparseness. I discuss results for each factor varied in the simulations.

Item Loadings. I studied two levels of item loadings: 1.5 and 2.0. The impact of extreme thresholds varied by factor loading condition; with higher factor loadings the impact of extreme thresholds was minimized. Because marginal endorsement level and item loading are confounded (i.e., the same threshold yields different endorsement rates for different values of λ), this result is due in part to higher factor determinacy and in part to higher marginal endorsement rates. However, this general pattern of results is consistent with earlier work studying ML estimation for GLFA with categorical indicators in limited samples (Forero & Maydeu-Olivares, 2009; Moshagen & Musch, 2014). These results are also consistent with research for GLFA models with continuous indicators (Gagné & Hancock, 2006; Marsh et al., 1998), which shows that stronger factor loadings improve the quality of solutions in finite samples, in terms of convergence and parameter estimate efficiency.

Item Thresholds. The two levels of item thresholds I examined were ν=3.85 and ν=4.90, corresponding to expected frequencies of 25 and 10 when λ =1.5 and 37.5 and 17.5 when λ=2.0 for the moderate sample size of 500. Sparseness had very little effect on bias in parameter estimates using ML estimation. Under the conditions studied, ML estimation converged in a high proportion of replications, and convergence never fell below 90%. However, as expected, sparseness led to suspiciously large parameter estimates in a substantial proportion of replications. Effects of sparseness were minimal with ν=3.85, but substantial with ν=4.90 on a high proportion of items on a single factor or on both factors. Moshagen and Musch (2014) also reported suspicious ML estimates despite high convergence rates, and the present results support their finding that achieving convergence to proper ML solutions does not necessarily indicate that results are trustworthy. Besides decreased efficiency and the presence of extreme parameter estimates, empirical power to detect true effects decreased in conditions with substantial sparseness, especially with ν=4.90 or a lower item loading.

Considering the broader literature on GLFA models, the issue of very low endorsement for categorical items is analogous to continuous items with very low variance. Continuous items with low variance can cause estimation problems related to empirical under-identification (Bentler & Chou, 1987; Rindskopf, 1984). With item variances near zero, there is too little information available to perform estimation. While this research is not intended to identify exact frequencies or marginal probabilities where sparseness becomes an issue, the general principle is that sparse endorsement can lead to items with insufficient information to perform ML estimation. I note that ML estimation performed reasonably well in more mild sparseness conditions for the models studied. However, smaller sample size, lower item loadings, fewer items per factor, and increased model complexity would all be expected to worsen the performance of ML (Forero & Maydeu-Olivares, 2009; Gagné & Hancock, 2006; Marsh et al., 1998; Moshagen & Musch, 2014).

This study does not unambiguously disentangle the relationship between sample size, endorsement rates, and endorsement frequency, because sample size was held constant throughout the simulation. However, it is clear that frequencies play a more important role than endorsement rates; a 5% probability of endorsement with N=100 will be more problematic than 5% probability of endorsement with N=500.

Patterns of sparseness. I studied the effects of sparseness in models with three patterns of sparseness: 2/5 items sparse on both factors, 4/5 items sparse on only one factor, and 4/5 items sparse on both factors. Just as the impact of sparseness was more pronounced with a higher threshold (ν=3.85 versus ν=4.9), the impact of sparseness was also dependent on the pattern of sparse items. The presence of extreme values and parameter estimate efficiency worsened with a high proportion of sparse items on one or both factors. As with the level of sparseness, the effect of the number of sparse items will also depend on the overall determinacy of the model; fewer

sparse items may be problematic with a smaller sample, lower factor loadings, and based on the other relevant factors identified (e.g. Forero & Maydeu-Olivares, 2009; Gagné & Hancock, 2006).

Empirical power was lower in models with a majority of sparse items on one or both factors. In these models item loadings, thresholds for sparse items, and the correlation between factors were substantial, true effects. Of these, item loadings and the correlation between factors are particularly meaningful in practice. For models with a high number of sparse items, empirical power to detect significant item loadings for both sparse and non-sparse items was low. Non-significant factor loadings for indicators of a latent construct would be very troubling in practice; these items would typically be removed (e.g., Kline, 1994). Power to detect a significant correlation between factors or significant factor loadings also fell to about .50 in the most severe conditions studied. Decreased power was a result of the large increase of variability in the estimates and associated increase in standard errors for ML estimation of models with sparse indicators.

In sum, the general pattern of results in the first simulation was consistent with analytic theory and my hypotheses that ML solutions would perform poorly in conditions characterized by high sparseness, in terms of probability of endorsement and number of sparse items, even with a moderately large sample size and reasonably high item loadings.

Comparing Bayesian Estimation to ML for Sparse Items

In this study I examined Bayesian estimation as an alternative to ML estimation for GLFA models with sparse categorical indicators. First, I compared Bayesian and ML estimation for GLFA models with no sparse items. Second, I compared Bayesian estimation to ML estimation in two conditions where ML estimation failed.

Baseline Comparison. Results from the second simulation study also supported my hypotheses that Bayesian estimation would perform as well as ML estimation in baseline conditions where ML performs well. In a baseline comparison condition the performance of Bayesian estimation matched ML estimation using a variety of prior specifications. These findings are consistent with theory that Bayesian estimation and ML estimation are generally equivalent using flat priors, and also that prior information is inconsequential given sufficient information in the data (Gelman et al., 2013). These findings also demonstrate that there should be no disadvantage to choosing Bayesian estimation over ML for GLFA models, even if sparseness is not an issue. Although not studied here, Bayesian estimation may also be useful as an alternative estimator to ML for high-dimensional models and for assessing model fit (Béguin & Glas, 2001; Edwards, 2010).

Comparison with Sparse Items. Results comparing Bayesian and ML estimation in two conditions where ML estimation was poor showed that Bayesian estimation could provide improved efficiency and empirical power and eliminate extreme estimates; however this performance was dependent on a reasonably concentrated prior and resulted in an increase in bias for some parameters. I did not expect Bayesian estimation to lead to such high bias using a moderate prior specification. The moderate prior (i.e. $\pi(\lambda_i, \nu_i) \sim N(0, \sigma = 10)$) aided posterior simulation with sparse indicators in terms of sampling (compared to flat priors); however the prior information was not enough to limit relatively extreme estimates in the posterior distribution. The concentrated priors specifying high probability that item intercepts and loadings were less than 7 in magnitude, loadings were positive, and the correlation between factors was positive, contained a sufficient amount of information to limit extreme estimates.

The bias in some parameter estimates resulting from Bayesian estimation had a notable pattern. With concentrated priors, the correlation estimate was underestimated, loadings for items

that were not sparse were overestimated, and factor loadings for items that were sparse were slightly underestimated. This pattern attributes relatively higher weight to non-sparse items; it is also interesting because ML estimation was more likely to yield extreme estimates (loadings and thresholds) for sparse items. In Bayesian estimation, less emphasis is based on unbiasedness and more emphasis is based on variance (Gelman et al., 2013, Ch. 4.5). Despite the tradeoff in bias, the overall efficiency of these parameter estimates, empirical power, and absence of extreme values were all an improvement over ML estimation.

Posterior simulation was reasonably fast using at least moderately concentrated priors. Flat priors were problematic for posterior simulation in the conditions with high sparseness; this is not surprising because the likelihood in these conditions is not well-behaved (as we saw for ML estimation of these conditions) and posterior simulation using Bayesian estimation without any restriction on the prior distribution is computationally challenging. Note that because the simulation design drew from a population distribution with fixed values, the simulation was actually set up to be more consistent for a traditional estimation approach than a Bayesian approach. The fact that Bayesian estimation performed well despite the simulation design being more consistent with traditional methods demonstrates the satisfactory performance of Bayesian estimation even in theoretically sub-optimal conditions.

Unique Contribution

Sparse items commonly arise in psychological research due to limited sample sizes and rare behaviors such as substance use. Previous research has suggested that ML is the best estimation method available for GLFA estimation with sparse items (Forero & Maydeu-Olivarez, 2009), but that the ML estimation may fail under conditions of sparseness (Forero & Maydeu-Olivarez, 2009; Moshagen & Musch, 2013). This previous work was suggestive of the effects of sparseness, but the effects of sparseness in these studies were confounded by low item loadings

and few indicators per factor. To my knowledge, no previous studies have directly studied the effects of sparseness for ML estimation in well-determined GLFA models.

In this study I found that in properly-specified, well-determined models, moderate sparseness had a small impact. More pronounced sparseness on a larger proportion of items, especially with lower factor loadings, led to problems using ML. These findings add to the prior literature (Forero & Maydeu-Olivarez, 2009; Moshagen & Musch, 2013, Wirth & Edwards, 2007) that using currently available estimation methods, some models cannot be reliably tested using currently available estimation methods. This means that researchers may be forced to drop sparse items and that some research questions involving sparse items cannot be asked using currently available methods. Forero and Maydeu-Olivarez (2009) suggested that "future research should investigate if new estimators are able to yield adequate results in these conditions".

Bayesian estimation for GLFA models has been demonstrated previously (Albert & Chib, 1993; Béguin & Glas, 2001; Edwards, 2010; Patz & Junker, 1999, Song & Lee, 2002, 2012; Lee & Tang, 2006). My work here builds on the prior research in three ways. First, this study is the first to study Bayesian estimation for GLFA models with sparse indicators. Previous studies have motivated the use of Bayesian estimation for reasons such as estimating high-dimensional models (Edwards, 2010) or for advantages for testing hypotheses related to fit (Béguin & Glas, 2001). Second, previous research using Bayesian estimation for GLFA models used relatively flat prior distributions and did not incorporate prior information to stabilize parameter estimates as I do here. Béguin & Glas (2001) examine different prior distributions as a sensitivity analysis, and Edwards (2010) incorporated minimal prior information to aid convergence, but my study is the first to utilize prior information about the expected range for parameter estimates in psychology to stabilize estimates for MCMC. Third, previous research disseminating Bayesian estimation of GLFA models used a combination of Gibbs and Metropolis-Hastings MCMC

algorithms that were difficult to implement, requiring implementation using custom programming (e.g. Edwards, 2010, Patz & Junker, 1999) and offered less flexibility for prior specification. I demonstrate Bayesian estimation using Stan (Stan Development Team, 2015), which offers fast and efficient computation as well as flexible prior and model specification.

In my study, I found that using prior information for Bayesian estimation of GLFA models with sparse indicators helped stabilize estimates and improve power compared to ML. This provides a new tool for researchers to address the limitations of currently available estimation methods for a challenging problem that often arises in psychological research.

Recommendations for Applied Researchers

Bayesian estimation of GLFA models as I demonstrate here requires introductory knowledge of Bayesian statistics and careful oversight to ensure that sampling is done correctly. It is difficult to imagine using Bayesian estimation for these models without this oversight and introductory knowledge. There are many helpful resources available specifically for Bayesian analysis using Stan (e.g., Stan Development Team, 2015; Gelman et al., 2013), which also includes an active online users group.

Before constructing models and choosing an estimator, I recommend examining item-level frequencies for sparseness. For well-determined models in moderate to large samples with moderate sparseness, it may not be necessary to take a Bayesian approach. However my results and previous research suggest that Bayesian estimation should not give results inferior to ML estimation, if done correctly. If sparseness is not an issue, results should be comparable using either estimator. However if a research question requires modeling sparse data, a Bayesian estimation approach can be useful to stabilize estimates and increase statistical power by incorporating prior information. For some research questions (e.g. illicit drug use, early alcohol use), investing the time and effort to take a Bayesian estimation approach may maximize a

researcher's ability to draw inferences from data that is exceedingly difficult to collect. In practice it may be difficult to determine if ML estimates are "untrustworthy" – since extreme estimates may appear in converged solutions. However, if ML estimates appear unreasonable, this suggests that the researcher has prior information about parameter estimates (deeming estimates unreasonable requires knowledge about what is reasonable), which could be incorporated into a Bayesian specification.

My results show that a Bayesian approach will be most helpful if adequate prior information is incorporated. This can include the expected direction of factor loadings, correlations between factors, and ranges of parameter estimates. The concentrated prior I suggest here I believe is reasonable for a variety of applications, but ultimately this choice will depend on knowledge of the content area. Bayesian estimation with flat priors will offer no benefit over ML estimation for GLFA models with sparse data.

Applied researchers should also be aware of difficulties in MCMC estimation. Specifically, sign indeterminacy is an issue for the scaling I demonstrate here if prior information does not restrict the sign of the latent factors, and posterior inference required post-processing the solutions to an interpretable solution. It is important to monitor convergence diagnostics including plots, and statistics such as the effective sample size and potential scale reduction factor. The method of Bayesian estimation I demonstrate here can be adapted for many different models within the GLFA family with different types of indicators, numbers of items or factors, and by incorporating predictors, and could also be extended to more comprehensive structural equation models.

Limitations and Future Directions

As in any line of enquiry, the present work answers some questions while also raising new ones. There are several remaining questions to be addressed by future research.

Most notably, the models considered here were correctly specified. It is important to investigate if the use of priors to stabilize estimates could also have the unintended consequence of masking model misspecification. Related to this, an important line of research will be to investigate the utility of Bayesian methods for assessing model fit (e.g. posterior predictive checks, Bayesian model selection) in these models. Currently, methods for assessing model fit using ML estimation are limited. Limited-information methods for estimation provide tests for model fit, but are less appropriate for modeling sparse data (Wirth & Edwards, 2007; Rhemtulla et al., 2012).

In the simulation studies here, I studied a very small subset of all possible conditions. It is possible to predict how these results would extend to many different conditions based on theory. For example, with smaller factor loadings or fewer items, ML estimation would be expected to perform worse, and the benefit of Bayesian estimation may be greater, however this would still depend on the incorporation of sufficient prior information. I did not study different overall sample sizes, however I predict that the performance found here is tied in large part to the frequencies for sparse items rather than overall sample size. Extending these results to polytomous items with more response categories raises a number of interesting issues. Polytomous items contain more information than binary items; however they require estimating additional thresholds and the potential for varied mechanisms and patterns of sparseness raises additional complexity. Despite these unanswered questions, the present work is a unique contribution, providing an alternative to improve estimation for models with sparse endorsement.

APPENDIX A. EXAMPLE MPLUS PROGRAM FOR GLFA

```
DATA: FILE IS data1.dat;
VARIABLE: NAMES ARE dep1-dep5 drug1-drug5;
categorical = dep1-dep5 drug1-drug5;

model:
[dep1$1*0];
[dep2$1*0];
[dep3$1*0];
[dep4$1*0];
[dep5$1*0];

[drug1$1*0];
[drug2$1*0];
[drug3$1*0];
[drug4$1*0];
[drug5$1*0];

dep by dep1* dep2* dep3* dep4* dep5*;
dep@1;
[dep@0];
drug by drug1* drug2* drug3* drug4* drug5*;
drug@1;
[drug@0];
dep with drug;

analysis:
estimator=ml;
```

APPENDIX B. STAN PROGRAM FOR GLFA – CONCENTRATED PRIORS

```
# 2-factor GLFA for binary items
data {
  int<lower=0> N;  // number of ppl
  int<lower=0> K;  // number of items
  int y[N,K];      // Y matrix of P items for N ppl

}
transformed data {
  vector[2] mu;
  for (i in 1:2) mu[i] <- 0;
}
parameters {
  vector[2] eta[N];  // eta for each person
  real nu[K];        // int for item k
  real <lower=0> lambda[K];    // loading item k
  real <lower=0, upper=1> cov;
}
transformed parameters {
matrix[2,2] sigma;
sigma[1,1]<-1;      sigma[1,2]<-cov;
sigma[2,1]<-cov;    sigma[2,2]<-1;
}
model {
  nu~ normal(0,3.57);
  lambda~normal(0,3.57);
  eta ~ multi_normal(mu,sigma);
  for (n in 1:N){
    for (k in 1:5){
      y[n,k]~ bernoulli_logit(-nu[k]+lambda[k]*eta[n,1]);
              }
    for (k in 6:K){
      y[n,k]~ bernoulli_logit(-nu[k]+lambda[k]*eta[n,2]);
              }
    }
 }
```

REFERENCES

Agresti, A. (2012). *Categorical data analysis* (2nd ed.). Hoboken: Wiley.

Albert, J.H., & Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*(422), 669–679.

Anderson, J.C. & Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155-173.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Hoboken: Wiley.

Bauer, D., Howard, A., Baldasaro, R., Curran, P., Hussong, A., Chassin, L., & Zucker, R. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods, 18*(4), 475-493.

Bauer, D. J, & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14*(2): 101-125.

Bayarri, M.J., & Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association, 95*(452), 1127-1142.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541-561.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2): 238-246.

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*(1), 78-117.

Berger, J. O., & Bernardo, J. M. (1992). Ordered group reference priors with application to the multinomial problem. *Biometrika, 79*(1), 25. doi:10.2307/2337144

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Bollen, K. A. (1989). *Structural equations with latent variables* (1st ed.). US: Interscience.

Bollen, K. A., & Bauldry, S. (2010). Model identification and computer algebra. *Sociological Methods & Research, 39*(2), 127-156.

Bollen, K. A., & Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective*. Hoboken : Wiley.

Bollen, K. A., & Maydeu-Olivares, A. (2007). A polychoric instrumental variable (PIV) estimator for structural equation models with categorical variables. *Psychometrika, 72*(3), 309-326.

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612.

Cai, L. & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*, 245-276.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2015) . Stan: A probabilistic programming language. Manuscript submitted for publication.

Chassin, L., Presson, C., Il-Cho, Y., Lee,, M. and Macy, J. (2013) Developmental Factors in Addiction: Methodological Considerations, in The Wiley-Blackwell Handbook of Addiction Psychopharmacology (eds J. MacKillop and H. de Wit), Wiley-Blackwell, Oxford, UK.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Collins, L., Schafer, J., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81-100.

Curran, P. J., et al. (in preparation). Improving factor score estimation through the use of exogenous covariates.

Dempster, A., N. Laird, & D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B, 39*, 1-38.

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in bayesian growth mixture modeling. *Structural Equation Modeling, 21*(2), 239-252. doi:10.1080/10705511.2014.882686

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309–326.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B, 195*(2), 216-222.

Dunson, D. B., & Dinse, G. E. (2001). Bayesian incidence analysis of animal tumorigenicity data. *Journal of the Royal Statistical Society.Series C (Applied Statistics), 50*(2), 125-141.

Edwards, M.C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor

analysis. *Psychometrika, 75*, 474-497.

Forero, C.G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded models for rating data: Limited vs. full information methods.*Psychological Methods*, 14, 275-299.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16, 625– 641.

Gagné, P. E., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, 41, 65-83.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*(409), 398.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*(3), 515-534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D.B., Vehtari, A, Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall.

Gelman, A., Meng, X.L., Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 779-786.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457-472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*(6), 721-741.

Ghosh, J., & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis.*Journal of Computational and Graphical Statistics, 18*(2), 306-320.

Hallquist, M., & Wiley, J. (2014). MplusAutomation: Automating Mplus Model Estimation and Interpretation. R package version 0.6-3. https://CRAN.R-project.org/package=MplusAutomation

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika, 57*(1), 97-109.

Hoffman, M., & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Resaerch, 15*, 1593-1623.

Houts, C. R., & Cai, L. (2013). flexMIRT R user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics: Second edition.*

Hussong, A. M., Curran, P. J. & Bauer, D. J. (2013). Integrative Data Analysis in Clinical Psychology Research. *Annual Review of Clinical Psychology*, 9:61-89.

Hussong, A. M., Flora, D. B., Curran, P. J., Chassin, L. A., & Zucker, R. A. (2008). Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents. *Development and Psychopathology, 20*(1), 165-193.

Hussong, A. M., Huang, W., Serrano, D., Curran, P. J., & Chassin, L. (2012). Testing whether and when parent alcoholism uniquely affects various forms of adolescent substance use. *Journal of Abnormal Child Psychology, 40*(8), 1265-1276.

Johnston, L. D., O'Malley, P. M., Miech, R. A., Bachman, J. G., & Schulenberg, J. E. (2015). *Monitoring the Future national survey results on drug use: 1975-2014: Overview, key findings on adolescent drug use.* Ann Arbor: Institute for Social Research, The University of Michigan.

Joreskog, K. G., & Sorbom, D. (2001). *LISREL user's guide.* Chicago: Scientific Software International.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*(435), 1343-1370.

Kline, P. (1994). *An Easy Guide to Factor Analysis*. Routledge: NY.

Koehler, K., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association, 75*, 336–344.

Kolenikov, S., & Bollen, K. A. (2012). Testing negative error variances: Is a Heywood case a symptom of misspecification? *Sociological Methods & Research, 41*(1), 124-167.

Lee, S., & Song, X. (2012). *Basic and advanced bayesian structural equation modeling: With applications in the medical and behavioral sciences.* GB: Wiley.

Lee, S.Y., & Tang, N.S. (2006). Bayesian analysis of structural equation models with mixed exponential family and ordered categorical data. *British Journal of Mathematical and Statistical Psychology, 59*, 151–172.

Loken, E. (2005) Identifiability constraints and the shape of the likelihood in confirmatory factor models. *Structural Equation Modeling, 12*, 232-244.

MacCallum, R.C., Edwards, M.C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods, 17*, 340-345.

MacKinnon DP and Fairchild A (2009) Current directions in mediation analysis. *Current Directions in Psychological Science*, 18:16–20.

Marsh, H. W., Hau, K., Balla, J. R., & Grayson, D. (1998). Is more ever too much? the number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*(2), 181-220.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713–732.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21*(6), 1087-1092.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177-195.

Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 10*(2), 60-70.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods*. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (2010). MCMC using Hamiltonian dynamics, to appear in the *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (eds.), Chapman & Hall.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association, 103*(482), 681-686.

Patz, R.J., & Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Peddada, S. D., Dinse, G. E., & Kissling, G. E. (2007). Incorporating historical control data when comparing tumor incidence rates. *Journal of the American Statistical Association, 102*(480), 1212-1220.

Rhemtulla, M., Brosseau-Liard, P. É., & 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, 17*, 354-373.

Rindskopf, D. (1984). Structural equation models: Empirical identification, heywood cases, and related problems.*Sociological Methods & Research, 13*(1), 109-119.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations.*Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 253-273.

Skrondal, A. (2000). Design and analysis of monte carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35*(2), 137-167.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* GB: Chapman & Hall.

Song, X.Y. & Lee, S.Y. (2002). Analysis of structural equation model with ignorable missing continuous and polytomous data. *Psychometrika, 67*(2), 261-288.

Stan Development Team (2015). *Stan Modeling Language Users Guide and Reference Manual, Version 2.7.0.*

Steiger, J.H., & Lind, J.C. (1980). Statistically Based Tests for the Number of Common Factors. Paper presented at the annual meeting of the Psychometric Society, May, Iowa City, IA.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*(398), 528-540.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288.

Wasserman, L. (2005). *All of Statistics*. New York, NY: Springer Science+Business Media, Inc.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.

Yuan, Y., Johnson, V.E. (2012). Goodness-of-Fit Diagnostics for Bayesian Hierarchical Models. *Biometrics, 68*(1), 156-164.

Zhu, M., & Lu, A. Y. (2004). The counter-intuitive non-informative prior for the bernoulli family. *Journal of Statistics Education, 12*(2), 1-10.