GENERALIZED FIDUCIAL INFERENCE FOR GRADED RESPONSE MODELS

Yang Liu

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.

Chapel Hill
2015

Approved by:

David Thissen

Daniel J. Bauer

Patrick J. Curran

Jan Hannig

Andrea Hussong

ABSTRACT

YANG LIU: GENERALIZED FIDUCIAL INFERENCE FOR GRADED RESPONSE MODELS.
(Under the direction of David Thissen)

Generalized fiducial inference (GFI) has been proposed as an alternative inferential framework in the statistical literature. Inferences of various sorts, such as confidence regions for (possibly transformed) model parameters, making prediction about future observations, and goodness of fit evaluation, can be constructed from a fiducial distribution defined on the parameter space in a fashion similar to those used with a Bayesian posterior. However, no prior distribution needs to be specified. In this work, the general recipe of GFI is applied to the graded response models, which are widely used in psychological and educational studies for analyzing ordered categorical survey questionnaire data. Asymptotic optimality of GFI is established (Chapter 2), and a Markov chain Monte Carlo algorithm is developed for sampling from the resulting fiducial distribution (Chapter 3). The comparative performance of GFI, maximum likelihood and Bayesian approaches is evaluated via Monte Carlo simulations (Chapter 4). The use of GFI as a convenient and powerful tool to quantify sampling variability in various inferential procedures is illustrated by an empirical data analysis using the patient-reported emotional distress data (Chapter 5).

To my dearest wife Flora.

ACKNOWLEDGMENTS

I would like to show my gratitude to all the committee members, especially my academic advisors Drs. David Thissen and Jan Hannig. I could not have finished the work without their valuable feedbacks and advices. The project is sponsored by the Harold Gulliksen Psychometric Research Fellowship from the Educational Testing Service (ETS). I am grateful to my mentors Drs. Shelby Haberman and Yi-Hsuan Lee at ETS. It has been a great experience working under their supervision, and I have been greatly benefited from their expertise in statistics. In addition, I would like to offer my sincere appreciation to the help and support from the current and former members of the Thurstone Lab—especially Brooke, Jim, and Jolynn, and those from Hannig's fiducial lab—Jessi, Qing, Dimitris, Abhishek, Rosie, and Jenny. I also cannot express enough thanks to Drs. Li Cai, Ji Seung Yang, and Scott Monroe for their continued support and encouragement for my work.

Finally, to my caring, loving, and supportive wife, Flora: my heartfelt thanks. I deeply appreciate you taking care of most household activities in the past year; it was a great comfort and relief for me to focus on completing my dissertation. Your encouragement during the difficult and painful times are duly noted; I could not have gone through the frustration and depression without your company.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Overview

This research focuses on an application of generalized fiducial inference, an omnibus statistical inference framework, to a general family of graded response models which has been extensively used in psychological research for analyzing survey data with ordinal response categories. Chapter 1 provides a brief literature review of various issues related to the graded response model, and the rationale of generalized fiducial inference. Specifically, extant likelihood-based and Bayesian approaches for point estimation, interval estimation, test scoring, and goodness of fit testing are summarized. Derivation of a fiducial distribution for item parameters and related statistical properties are discussed in Chapter 2. The key result of this part is a Bernstein-von Mises theorem that guarantees the asymptotic optimality of the fiducial distribution. A Markov chain Monte Carlo algorithm is developed for sampling from the fiducial distribution, which is described in Chapter 3. A Fortran implementation of the sampler is available from the author upon request. In Chapter 4, a large-scale simulation study is conducted to evaluate the comparative behavior of the proposed procedure with existing likelihood-based and objective Bayesian approaches for various inferential problems. Finally, patient-reported emotional distress data from the Patient Report Outcomes Measurement Information System (PROMIS) study are analyzed in Chapter 5 using generalized fiducial inference. We emphasize that the uncertainty in parameter estimates due to sampling variability should be taken into account in subsequent analyses; to this end, the proposed framework serves as a theoretically sound but analytics-free tool. Detailed proofs of the theorems and propositions are given in appendices.

1.2 The graded response model

The graded response model (GRM; Samejima, 1969) has become a standard item response theory (IRT) model for analyzing Likert-type response scales which have polytomous categories coded in order (e.g., strongly disagree, disagree, neutral, agree, strongly agree). Survey questionnaires including Likert items have been designed to measure many psychological constructs including personality attributes, attitudes, health-related outcomes, etc. The GRM models responses to each item as an ordinal logistic regression (also known as a proportional odds model) on one or more latent variables representing the underlying constructs of interest. Heuristically, an item response is treated in the GRM as a discrete realization of a continuous but latent propensity that is related to individual differences in target constructs and also item characteristics. The relative position of a particular response category on the latent continuum is gauged by the adjacent item difficulty/intensity parameters that are transformations of the slope and intercept parameters in the regression. The GRM reduces to the two-parameter logistic (2PL; Birnbaum, 1968) model when there are only two response categories.

In the current work, we focus our attention on a family of multidimensional logistic GRM models including unidimensional, bifactor, and exploratory GRMs as special cases. Our notation is more consistent with the mixed-effect modeling convention than the default choice in the IRT literature. For a K_j-category ordinal item j and a single respondent i, define the item response function (IRF), denoted $f_j(\boldsymbol{\theta}_j, k|\mathbf{z}_i)$, as the probability of endorsing the kth category, i.e., $Y_{ij} = k$, $k = 1, \ldots, K_j$, conditional on this particular person's latent variable values $\mathbf{Z}_i = \mathbf{z}_i$:

$$\begin{aligned} f_j(\boldsymbol{\theta}_j, k|\mathbf{z}_i) &= P\{Y_{ij} = k|\mathbf{Z}_i = \mathbf{z}_i\} \\ &= \begin{cases} 1 - \dfrac{1}{1 + e^{\alpha_{j1} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i}}, & k = 0; \\ \dfrac{1}{1 + e^{\alpha_{j,K_j-1} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i}}, & k = K_j - 1; \\ \dfrac{1}{1 + e^{\alpha_{jk} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i}} - \dfrac{1}{1 + e^{\alpha_{j,k+1} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i}}, & \text{otherwise.} \end{cases} \end{aligned} \tag{1.1}$$

In (1.1), α_{jk}'s denote the intercept parameters and $\boldsymbol{\beta}_j$ the slopes. We assume that all the intercept parameters are freely estimated, while some slopes are fixed for model identification; let $\boldsymbol{\theta}_j$ be all free parameters that calibrate item j. The r-dimensional latent variables are assumed to be standard normal, $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r\times r})$. Inference for models with unknown covariance structure among latent dimensions (e.g., simple-structure models) is beyond the scope of the present work.

For a test comprising m graded items, we assume conditional dependence among item responses given the latent variable (e.g., McDonald, 1981), which implies the likelihood function $f(\boldsymbol{\theta}, \mathbf{y}_i)$ of an individual item response vector $\mathbf{y}_i = (y_{ij})_{j=1}^m$:

$$f(\boldsymbol{\theta}, \mathbf{y}_i) = \int_{\mathbb{R}^r} \prod_{j=1}^m f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}_i), \tag{1.2}$$

in which $\Phi(\cdot)$ denotes the probability measure of an r-dimensional standard normal distribution. Further assume the sample is composed of n independent and identically distributed (i.i.d.) item response vectors $\mathbf{y} = (\mathbf{y}_i)_{i=1}^n$; the corresponding sample likelihood function is

$$f_n(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^n f(\boldsymbol{\theta}, \mathbf{y}_i). \tag{1.3}$$

1.2.1 Point estimation

In the frequentist framework, the gold standard maximum likelihood (ML) estimates of item parameters are typically obtained by numerically maximizing Equation 1.3 via either Newton-type (Bock and Lieberman, 1970; Haberman, 1988) or Expectation-Maximization (EM; Dempster, Laird, and Rubin, 1977; Bock and Aitkin, 1981) algorithms. When the latent dimensionality r is not high, the intractable expression in Equation 1.2 can be efficiently approximated by numerical integration; various types of quadrature systems have been used—e.g., rectangular, Gauss-Hermite (Ralston, 1965, pp. 93-97; Bock and Aitkin, 1981), or adaptive (Schilling and Bock, 2005; Haberman, 2006) quadrature. When the latent dimensionality is high, however, approximation of the likelihood function based on tensor-product quadrature suffers from the well-known "curse of dimensionality" that the number of quadrature points grows exponentially fast. One solution given by Meng and Schilling (1996) is to incorporate a Gibbs sampler for the E-step computation, resulting in a Monte Carlo EM algorithm. Alternatively, Cai (2010a; 2010b) proposed approximation of the gradient of the log-likelihood function by a Metropolis-Hasting sampler and locating its zero by a Robbins-Monro-type search.

On the other hand, Bayesian methods based on stochastic approximations of the posterior distribution (e.g., Albert, 1992; Patz and Junker, 1999; Bradlow, Wainer, and Wang, 1999; Curtis, 2010; also see Edwards (2010) for its application in the GRM) are not affected as much by increasing latent dimensionality. Such a complex sampling problem is usually addressed by Markov chain Monte Carlo (MCMC) methods. However, it is commonly agreed that Bayesian methods are less user-friendly than ML: Statistical expertise is required to specify prior distributions and tune sampling algorithms. Even though the asymptotic optimality of Bayesian posteriors can be guaranteed by the celebrated Bernstein-von Mises theorem (e.g., Le Cam and Yang, 1986), erroneous results may be seen in finite-sample applications resulting from improperly chosen prior distributions or ill-behaved samplers.

1.2.2 Confidence interval/set

Associated with ML estimation, confidence intervals (CIs) are most often constructed by inverting the Wald test, with standard errors produced by suitable supplemental procedures (e.g., Cai, 2008; Yuan, Cheng, and Patton, 2014). However, caveats on Wald-type intervals have been raised in the statistical literature (e.g., Neale and Miller, 1997) because they rest on a quadratic approximation to the log-likelihood: They are not invariant under nonlinear transformations of parameters (the delta method is often used to obtain standard errors for a reparameterized model; its performance depends on how quadratic the log-likelihood function is with respect to the new parameterization). They may also cover values beyond the boundary of the parameter space, and may have unsatisfying small sample behaviors.

As a better alternative, CIs obtained by inverting the likelihood-ratio test have not yet been available in the IRT literature; the procedure itself is computationally intensive, and might not be suitable for multidimensional GRMs. Quantification of uncertainty for more complex transformations of parameters, e.g., the problem of drawing simultaneous confidence band on the item response functions (Thissen and Wainer, 1990) or information curves, are typically handled by less rigorous methods such as the bootstrap.

Bayesian methods based on posterior sampling are extremely flexible in terms of making inferences about arbitrary transformations of model parameters: One can apply the desired transformation to each Monte Carlo draw from the posterior distribution, leading naturally to a Monte Carlo sample of the transformed posterior. Moreover, we can save the random draws and pass them to subsequent analyses via, e.g., multiple imputation; compared to plugging in the point estimates, this accounts for the sampling variability in the initial model estimation stage. But the arbitrariness in prior selection may exert unpredictable influence on the finite-sample performance of Bayesian confidence sets, and thus they should be used with caution.

1.2.3 Goodness of fit testing

The item response data matrix $\mathbf{y} = (y_{ij})_{i=1}^{n}{}_{j=1}^{m}$ can be reorganized as an m-way contingency table, in which each dimension corresponds to the K_j response categories of an item and each cell of the table corresponds to a response pattern $\mathbf{y}_i = (y_{ij})_{j=1}^{m}$. Therefore, it is natural to assess GRM model fit by means of residuals in contingency table cells, testing whether the observed proportions are identical to the model-implied response pattern probabilities. For a general discussion on GOF testing for contingency table data, see Rao (1973, pp. 391-394) and more recently Haberman and Sinharay (2013). One salient feature of the item response data is sparseness, i.e., very small expected proportions for some cells, as a consequence of the number of cells, $\prod_{j=1}^{m} K_j$, increasing exponentially with the test length. It is well-known that the asymptotic theory of residuals works poorly in sparse tables (e.g., Cochran, 1952). A simple workaround is to collapse the table in a systematic fashion such that the asymptotics are well suited to the resulting table of a smaller size; this has been termed the limited information approach by some authors. Existing GOF testing procedures in the IRT literature are mostly based on this rationale (e.g., Glas, 1988; Reiser, 1996; Maydeu-Olivares and Joe, 2005, 2006; Cai, Maydeu-Olivares, Coffman, and Thissen, 2006; Joe and Maydeu-Olivares, 2010; Cai and Hansen, 2013).

It is as important to identify the source of model misfit as to test the overall GOF of item response models. In practice, model modifications and/or item level fine-tuning should be performed before inferences can be safely drawn from the fitted GRM. When no *a priori* information about the misfitting pattern is available, the source of misfit can be investigated by exploring the fit of the IRT model in low-dimensional (typically two-way or three-way) marginal subtables. For a comprehensive description of this approach, see Liu and Maydeu-Olivares (2014). If there is information about a potential violation of an assumption (e.g., conditional dependence, differential item functioning, etc.), researchers may specify a less restrictive model that reduces to the original model after imposing constraints, which turns the examination of model fit into a nested model comparison problem.

On the Bayesian side, posterior predictive checking (PPC; Guttman, 1967; Rubin, 1984) serves as a straightforward approach for detecting model misfit. When the model is correctly specified and the sample size is large enough, the value of a test statistic $T(\mathbf{y})$ computed with the observed data set should be close to the same statistic computed from a predictive data set conditional on the posterior distribution of IRT model parameters. Here, the distribution of predictive data is a composite of the posterior distribution and the data-generating model, and thus can be approximated stochastically by drawing model parameters from the posterior and then simulating data conditional on these draws. For applications of PPC in IRT models, see Sinharay (2005), Sinharay, Johnson, and Stern (2006), and Levy, Mislevy, and Sinharay (2009).

To cater to the increasing prevalence of multidimensional GRMs, it is desirable to develop a comprehensive estimation and inference framework that is able to: a) deal with high-dimensional latent traits, b) facilitate the assessment of uncertainty for various kinds of inference based on the model, and c) avoid as much subjectivity and ambiguity as possible in application. In the current research, generalized fiducial inference (GFI; Hannig, 2009, 2013) for a general class of multidimensional GRMs is proposed as an alternative to the existing full information methods. GFI satisfies most of the aforementioned desiderata. This recent variant of Fisher's (1930, 1932, 1935) fiducial inference is an interesting theoretical middle ground between frequentist and Bayesian methods. Inferential procedures are based on a probability distribution supported on the parameter space, namely a *fiducial distribution*, which is derived using only the information contained in the data. Consequently, it inherits all the flexibility of Bayesian methods, but requires no prior knowledge of model parameters.

1.3 Generalized fiducial inference

Fisher (1930) put forward the notion of fiducial probability in response to the method of inverse probability (i.e., Bayesian inference, especially with a uniform prior), which in his view was "fundamentally false and devoid of foundation". His concerns were conveyed in the following excerpt:

> *The peculiar feature of the inverse argument proper is to say something equivalent to "We do not know the function* Ψ *specifying the super-population* [i.e., the prior distribution of model parameters θ]*, but in view of our ignorance of the actual values of* θ *we may take* Ψ *to be constant."... but however we might disguise it, the choice of this particular a priori distribution for the* θ*'s is just as arbitrary as any other could be.*

He continued to point out that the claimed objectivity of the inverse probability cannot be translated under reparameterization:

> *If we were, for example, to replace our* θ*'s by an equal number of functions of them,* θ'_1, θ'_2, $\theta'_3, \dots,$ *... all objective statements could be translated from the one notation to the other, but the simple assumption* $\Psi(\theta_1, \theta_2, \theta_3, \dots) =$ *constant may translate into*

> *a most complicated frequency function for* θ_1', θ_2', $\theta_3', \ldots$

He also summarized the major reason why inverse probability was popular—that is, quantifying uncertainty with probability is intuitive and handy, and Bayes' rule seemed to be the only available tool at that time:

> *The underlying mental cause is…in the fact that we learn by experience, that science has its inductive processes so that it is naturally thought that such inductions, being uncertain, must be expressible in terms of probability. …The assumption was almost a necessary one seeing that no other mathematical apparatus existed for dealing with uncertainties. …The introduction of quantitative variates* [representing model parameters]*, having continuous variation in place of simple frequencies as the observational basis, makes also a remarkable difference to the kind of inference which can be drawn. …Inverse probability has, I believe, survived so long in spite of its unsatisfactory basis, because its critics have until recent times put forward nothing to replace it as a rational theory of learning by experience.*

Fisher in his 1930's article provided a template fiducial argument for a one-parameter model $Y \sim F_\theta$, in which θ is the parameter and F_θ is the distribution function monotonically decreasing in θ, based on a single observation $Y = y$. By the probability integral transformation, $F_\theta(Y) \sim \text{Uniform}(0, 1)$, which is, in modern terminology, a pivotal quantity. He transferred the pivotal distribution to the parameter space through function F_θ that is considered a function of θ, equivalent to the operation of "de-pivoting" for the purpose of obtaining CIs with an exact coverage (see e.g., Casella and Berger, 2002). The resulting distribution, having density $-dF_\theta/d\theta$, determines what Fisher called fiducial probability, which in this case is the same as the correct coverage probability accumulated from repeated samples. Later, Fisher illustrated this approach again with a Gaussian variance example (Fisher, 1933), and generalized it to multidimensional parameters (Fisher, 1935).

In the 1935's paper, Fisher claimed that his fiducial solution to the Behrens-Fisher problem is exact, which was later challenged by Bartlett; see Bartlett (1965) and Zabell (1992) for summaries of the dispute. Moreover, Fisher's interpretation of fiducial probability had been by no means cohesive: In his earlier work (Fisher 1930, 1933, and 1935), the fiducial probability was largely treated as the synonym of frequent coverage as in the Neyman-Pearson repeated-sampling scheme (Neyman, 1934); however, a more epistemic view conceded to the Bayesian camp (Fisher, 1945, 1955) was adopted, at about the same time when he threw a heated polemic against Neyman. The confusion, according to Zabell (1992), was likely to be traced to Fisher's mixed understanding of the nature of probability: In his own writings, probability is both "a frequency in an infinite hypothetical population" (Fisher, 1922) and "a numerical measure of rational belief" (Fisher, 1930). As a consequence, fiducial inference has been largely renounced by mainstream statisticians; it has been viewed as Fisher's "one great failure" (Zabell, 1992) or "the biggest blunder" (Efron, 1998).

Many attempts have been made to revitalize the idea of fiducial inference (see Hannig, 2009, for a historical review), among which there are two streams of research that are particularly relevant to the current dissertation. One is the notion of confidence distribution (CD; e.g., Efron, 1998; Schwider and Hjort, 2002; Xie and Singh, 2013). Tying back to Fisher's first definition of fiducial probability, a CD for a single parameter θ (with or without the presence of nuisance parameters) is defined such that its upper α quantile is the upper limit of a one-sided $100(1-\alpha)\%$ CI under the true model. In many practical problems, however, only asymptotic CDs can be found. Various existing inference tools are special cases asymptotic CDs, such as Bayesian posterior distributions and bootstrap distributions, provided suitable regularity conditions are satisfied. The other branch of studies are rooted in Fraser's (1968) structural inference and the Dempster-Schafer calculus (e.g., Dempster, 1968, 2008; Shafer, 1976), including generalized fiducial inference (GFI; Hannig, 2009; 2013) and inferential models (Martin, Zhang, and Liu, 2010; Martin and Liu, 2013). These works retained Fisher's idea of "finding solutions" of model parameters from the mathematical formula that links data, parameters, and pivotal quantities, but incorporated additional adjustment to achieve desirable statistical properties. Specifically, Hannig (2009) provided a simple recipe to construct an asymptotic CD, namely the generalized fiducial distribution, which is adaptable to a broad collection of statistical models.

In brief, the goal of GFI is to find a fiducial distribution on the parameter space capturing all the information that the observed data conveys about model parameters. It is achieved by a role-switching between data and parameters similar to that involved in the definition of a likelihood function. Hannig's (2009) fiducial argument operates on the data generating equation (also known as the structural equation):

$$\mathbf{Y} = \mathbf{g}(\boldsymbol{\theta}, \mathbf{U}), \tag{1.4}$$

which describes the data $\mathbf{Y}$ as a function of the parameters $\boldsymbol{\theta} \in \Theta$ and random components $\mathbf{U}$ having parameter-free distributions (i.e., pivotal quantities). For observed data $\mathbf{Y} = \mathbf{y}$, the data generating equation can be considered as an implicit function relating $\boldsymbol{\theta}$ to $\mathbf{U}$. Properly solving for $\boldsymbol{\theta}$ from Equation 1.4, i.e., writing the parameters as a function of the data and random components, transfers the known distribution of $\mathbf{U}$ to the parameter space and produces a fiducial distribution.

From now on, lowercase letters are routinely used for realizations of random variables. Let

$$Q(\mathbf{y}, \mathbf{u}) = \{\boldsymbol{\theta} : \mathbf{g}(\boldsymbol{\theta}, \mathbf{u}) = \mathbf{y}\} \tag{1.5}$$

be a set inverse of Equation 1.4 containing solutions of $\boldsymbol{\theta}$ to Equation 1.4 for fixed $\mathbf{y}$ and $\mathbf{u}$. In general, Equation 1.5 may contain more than one element for some values of $\mathbf{u}$, and may be empty for others; in terms of finding a solution $\boldsymbol{\theta}$, they correspond to under-identified and over-identified systems, respectively. Here, drawing analogy to solving a system of linear equations might be helpful. When Equation 1.5 consists of multiple elements, it resembles a linear system that has fewer equations than variables, and thus more than one solution is admissible. In this case, preference for one value of $\boldsymbol{\theta}$ or another cannot be decided from the values of $\mathbf{y}$ and $\mathbf{u}$ *per se*. A general solution for such an under-identified system can be denoted $\mathbf{v}(Q(\mathbf{y}, \mathbf{u}))$, in which $\mathbf{v}(\cdot)$ is some user-defined selection rule that chooses a point from the closure of Equation 1.5. On the other hand, when the set determined by Equation 1.5 is empty, it is similar to a linear system that has more equations than variables, in which case conflict may arise and no solution can be found. This implies that no feasible parameter value is able to recover $\mathbf{y}$ combined with the particular $\mathbf{u}$. Because we assume the model is correctly specified, and thus at least the true parameter values should be contained in the set inverse, intuitively it means that this $\mathbf{u}$ value is not helpful to the inference of $\boldsymbol{\theta}$ and should be discarded. Therefore, we should always prevent this from happening, and one natural workaround is to concentrate on the set of $\mathbf{u}$ such that Equation 1.5 is non-empty. Following these heuristics, a fiducial distribution can be defined as

$$\mathbf{v}(Q(\mathbf{y}, \mathbf{U}^\star)) \mid \{Q(\mathbf{y}, \mathbf{U}^\star) \neq \emptyset\}, \tag{1.6}$$

in which $\mathbf{U}^\star$ is an independent and identically distributed (i.i.d.) copy of the data generating $\mathbf{U}$. A (possibly vector-valued) random variable having the distribution determined by Equation 1.6 is referred to as a generalized fiducial quantity (GFQ), denoted $\mathbf{R}$.

Next, we discuss an illustrative example, namely, the binomial proportion problem (see Dempster, 1966; Hannig, 2009). GFQ for the binomial proportion parameter is derived following the generic recipe; the derivation is in many aspects similar to that of our main problem described in the next chapter.

Example: Binomial proportion. Suppose $Y_1, \ldots, Y_n$ are independent and identically distributed (i.i.d.) Bernoulli(π) random variables with success probability π. The data generating equation for each Y_i is

$$Y_i = \mathbb{I}\{U_i \leq \pi\},\ U_i \sim \text{Uniform}(0, 1), \tag{1.7}$$

in which $\mathbb{I}(\cdot)$ denotes the indicator function. To make inference about π, we consider the set inverse of Equation 1.7:

$$Q_i(y_i, u_i) = \begin{cases} [u_i, 1], & \text{if } y_i = 1; \\ [0, u_i), & \text{if } y_i = 0. \end{cases} \tag{1.8}$$

Equation 1.8 is one of the two segments of interval $[0, 1]$ divided by the value of u_i; see the top panel of Figure 1.1 for a visualization.

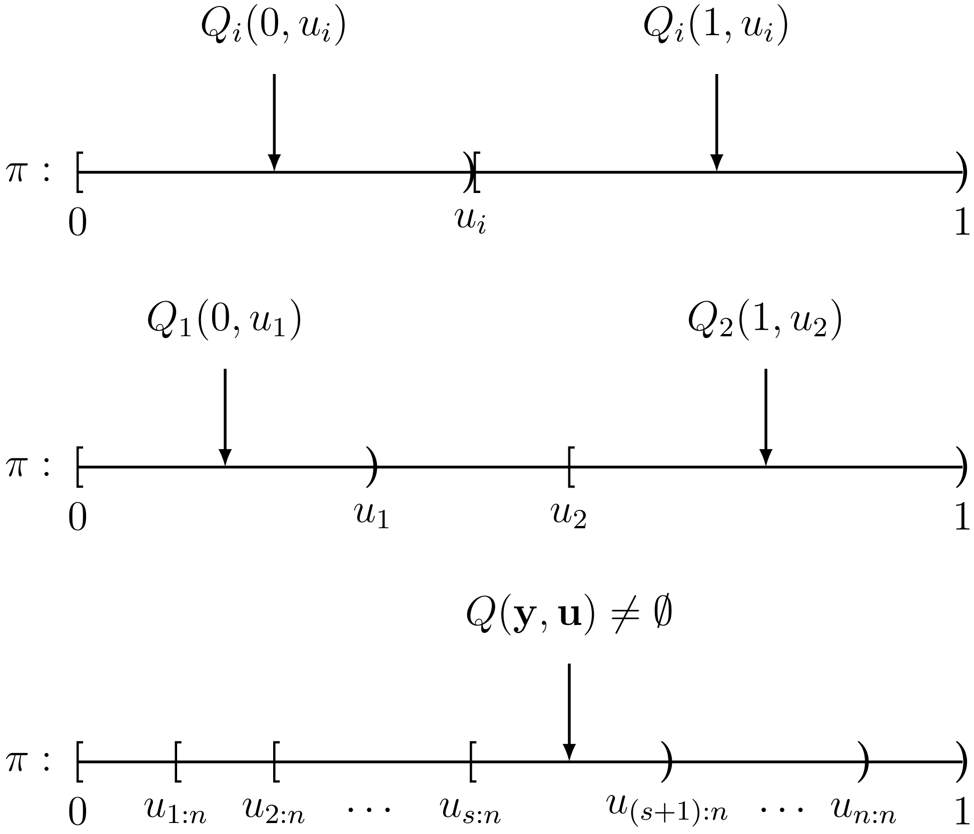

Figure 1.1: The binomial proportion example. The top panel shows the individual set inverse function $Q_i(y_i, u_i)$ for $y_i = 0$ and 1, respectively. The middle panel gives an example of empty $Q(\mathbf{y}, \mathbf{u}) = Q_1(y_1, u_1) \cap Q_2(y_2, u_2)$, in which $y_1 = 0$, $y_2 = 1$, and $u_1 < u_2$. The bottom panel displays a non-empty $Q(\mathbf{y}, \mathbf{u})$, in which the $s = \sum_{i=1}^n y_i$ smallest u_i's, denoted $u_{1:n}, \ldots, u_{s:n}$, correspond to successes, and the rest correspond to failures.

Let $\mathbf{Y} = (Y_i)_{i=1}^n$, and $S = \sum_{i=1}^n Y_i \sim \text{Binomial}(n, \pi)$. The set inverse function for $\mathbf{Y}$, denoted $Q(\mathbf{y}, \mathbf{u})$ in which $\mathbf{u} = (u_i)_{i=1}^n$, can be obtained by intersecting all individual set inverse functions (Equation 1.8), because by definition the set inverse includes all values of π that are consistent with all individual data generating equations (Equation 1.7). Formally, it can be written as

$$Q(\mathbf{y},\mathbf{u}) = \bigcap_{i=1}^{n} Q_i(y_i,u_i) = [\max_{i:y_i=1} u_i,\ \min_{i:y_i=0} u_i). \tag{1.9}$$

The set defined by Equation 1.9 can be empty; an example is given in the middle panel of Figure 1.1. To obtain a non-empty intersection, we need $\max_{i:y_i=1} u_i < \min_{i:y_i=0} u_i$, which is illustrated in the bottom panel of Figure 1.1. Also let $v(\cdot)$ be a selection rule that yields an element of Equation 1.9. Thereby we define the generalized fiducial distribution of π following the generic recipe (Equation 1.6):

$$R \overset{d}{=} v([\max_{i:y_i=1} U_i^\star,\ \min_{i:y_i=0} U_i^\star)) \mid \{\max_{i:y_i=1} U_i^\star < \min_{i:y_i=0} U_i^\star\}, \tag{1.10}$$

in which $U_i^\star$'s are i.i.d. copies of U_i's as usual. The GFD defined by Equation 1.10 satisfies the stochastic ordering: $U_{s:n} \preceq R \preceq U_{(s+1):n}$, in which $\preceq$ means "stochastically smaller than or equal to". More detailed discussion of this example, including the choice of selection rules $v(\cdot)$, can be found in Hannig (2009).

A GFQ serves as a prospective probabilistic quantification for the plausibility of model parameters after observing data, in contrast to the deterministic quantification given by the likelihood function, and also to the posterior distribution obtained by updating the prior knowledge of model parameters with the observed data. The fiducial probability of event $\{\mathbf{R} \in A \subset \Theta\}$ corresponds to the long-run proportion that parameter values in A would be needed in order to reproduce the observed data $\mathbf{y}$, over repeated data generation from the model (i.e., generate $\mathbf{U}^\star$ from its parameter-free distribution). Here we ignore for the sake of a simpler elucidation the fact that $Q(\mathbf{y},\mathbf{u})$ is possibly set-valued, and this can be taken as approximately correct in large samples (Hannig, 2013). The construction and interpretation do not require any prior knowledge on the model parameters, which is a marked logical difference from the Bayesian approach. Mathematically, however, GFQ is closely attached to empirical Bayeisan inference: The fiducial density can be written in the form of a Bayesian posterior with a data-dependent prior (see Hannig, 2009, Section 4.2, for a general discussion). A different connection between the two is established later for our problem involving a marginal likelihood, similar to that given by Liu and Hannig (2014, Remark 3).

GFQs defined by Equation 1.6 suffers from three major sources of non-uniqueness (Hannig, 2009): a) the choice of data generating equations, b) the choice of selection rules, and c) conditioning on a set with probability 0 (i.e., the Borel paradox, see Proschan and Presnell (1998) for detailed discussions). In the application of GFI to the GRM, we use a simple and natural data generating equation that parallels the way graded item response data are typically simulated in Monte Carlo studies, and that is shown to lead to a fiducial distribution that satisfies a Bernstein-von Mises theorem (and consequently an asymptotic CD); therefore, we do not feel pressed to explore other possible data generating equations. In addition, c) does not apply to categorical data, so it will not be discussed either. For b), it can be shown that the diameter of the set given by Equation 1.5 in our problem shrinks to 0 at the rate $1/n$, faster than the rate $1/\sqrt{n}$ at which GFQ approaches its normal limit as dictated by the Bernstein-von Mises theorem. Non-informative and data independent selection rules are recommended by Hannig (2009, Section 7) for finite sample applications.

In practice, Monte Carlo methods are frequently used when GFI is applied to complex parameteric models, due to fact that exact computation of functionals of the fiducial distribution, e.g., median and quantiles, are often intractable. The target distribution (Equation 1.6) can be approximated by simulating $\mathbf{U}^\star$ subject to the constraint $Q(\mathbf{y}, \mathbf{U}^\star) \neq \emptyset$ and constructing the implied set inverse $Q(\mathbf{y}, \mathbf{U}^\star)$ from each Monte Carlo draw of $\mathbf{U}^\star$, which is typically a quite involved truncated sampling problem in essence. Markov chain Monte Carlo (MCMC; e.g., E, Hannig, and Iyer, 2009; Hannig and Lee, 2009) or sequential Monte Carlo (SMC; e.g., Cisewski, and Hannig, 2012) approaches have been invoked to generate samples from the target fiducial density; see Hannig, Lai, and Lee (2014) for a detailed discussion on various computational issues of GFI.

It has been demonstrated in applications that GFI not only offers asymptotically optimal inference but outperforms ML and Bayesian approaches in small samples as well (e.g., Hannig, 2009; Cisewski and Hannig, 2012; Liu and Hannig, 2014). In the current work, we show that GFI, again, when applied to the GRM, delivers added value over conventional likelihood-based and Bayesian methods: We derive a Bernstein-von Mises theorem and some other important properties that guarantees the asymptotic correctness of GFI; in the simulation study, we continue to see that GFI is well-behaved even in very extreme conditions (small sample and skewed item parameters) where both ML and Bayesian approaches fail. In the next chapter, we first look at some large sample properties of GFI in the family of GRMs (Equation 1.1).

CHAPTER 2: THEORY

In this chapter, we derive a generalized fiducial distribution of item parameters under the family of multidimensional graded response models (GRMs; characterized by Equation 1.1). A Bernstein-von Mises theorem is established to justify the asymptotic correctness of generalized fiducial inference (GFI) for making inferences about item parameters and their transformations. We also discuss the consistency of the fiducial predictive distribution for sample statistics whose distributions depend on the item parameters, which is applicable to constructing predictive intervals for response pattern scores. We conclude this section by discussing an easy-to-implement goodness of fit testing procedure, the fiducial predictive check (FPC), analogous to the posterior predictive check (Guttman, 1967; Rubin, 1984) in the Bayesian literature.

2.1 A generalized fiducial distribution for item parameters

Following the general recipe introduced in Chapter 1, we derive a generalized fiducial distribution for item intercepts and slopes given independent and identically distributed (i.i.d.) responses to a collection of graded items. We start from the data generating equation of a person's response to an item under the GRM, and find the set inverse function of item parameters corresponding to this particular data entry. Combining all individual set inverse functions, we arrive at the set inverse for the entire item response data set, based on which a fiducial distribution can be defined in the form Equation 1.6.

Conditional on the latent variable $\mathbf{Z}_i$, person i's response to item j, i.e., Y_{ij}, follows a multinomial distribution with probabilities $P\{Y_{ij} = k|\mathbf{Z}_i\} = f_j(\boldsymbol{\theta}_j, k|\mathbf{Z}_i)$, $k = 1, \dots, K_j$, given by Equation 1.1 in the previous chapter, in which the free item parameters $\alpha_{j1}, \dots, \alpha_{j,K_j-1}$

and $\boldsymbol{\beta}_j$ are collected in $\boldsymbol{\theta}_j$. Similar to the binomial proportion example discussed in the previous chapter, the data generating equation (e.g., Hannig, 2009, Example 5) of the ordinal Y_{ij} can be written as

$$Y_{ij} = \sum_{k=1}^{K_j-1} \mathbb{I}\{U_{ij} \le f_j(\boldsymbol{\theta}_j, k|\mathbf{Z}_i)\} = \sum_{k=1}^{K_j-1} \mathbb{I}\{A_{ij} \le \alpha_{jk} + \boldsymbol{\beta}_j{}^\top \mathbf{Z}_i\}, \tag{2.1}$$

in which $U_{ij} \sim \text{Uniform}(0,1)$, $A_{ij} = \text{logit}(U_{ij}) \sim \text{Logistic}(0,1)$, and $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r\times r})$. Here, A_{ij} and $\mathbf{Z}_i$ can be identified as the pivotal component $\mathbf{U}$ in the general formulae (equations 1.4 to 1.6). Assume r_j slopes are free ($r_j \le r$), and thus the dimension of $\boldsymbol{\theta}_j$ is $q_j = r_j + K_j - 1$; in the sequel, we only consider the case in which the fixed slopes are zero for simplicity[1]. The set inverse function of Equation 2.1 is the following subset of the q_j-dimensional parameter space:

$$\begin{aligned} Q_{ij}(y_{ij}, a_{ij}, \mathbf{z}_i) = \{\boldsymbol{\theta}_j \in \mathbb{R}^{q_j} : {} & a_{ij} > \alpha_{j1} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i, \quad \text{if } y_{ij} = 0; \\ & a_{ij} \le \alpha_{j,K-1} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i, \quad \text{if } y_{ij} = K; \\ & \alpha_{j,k+1} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i < a_{ij} \le \alpha_{jk} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i, \ \text{otherwise.}\} \end{aligned} \tag{2.2}$$

Geometrically, Equation 2.2 corresponds to the intersection of two half-spaces if k is a middle category, and a single half-space if $k = 0$ or $K_j - 1$. A graphical illustration of Equation 2.2 for a three-category item is given in the left panel of Figure 2.1, in which the parameter space is three dimensional (two intercepts and one slope).

The set inverse function for n i.i.d. responses to item j, denoted $\mathbf{Y}_{(j)} = (Y_{ij})_{i=1}^n$, is given

---

[1] In practice, slopes might be fixed at values other than zero. The theoretical properties discussed in the current work still apply after subtracting the inner product of those fixed slopes and the corresponding normal variates from A_{ij}'s and substituting its distribution function for the standard logistic cumulative distribution function (cdf).

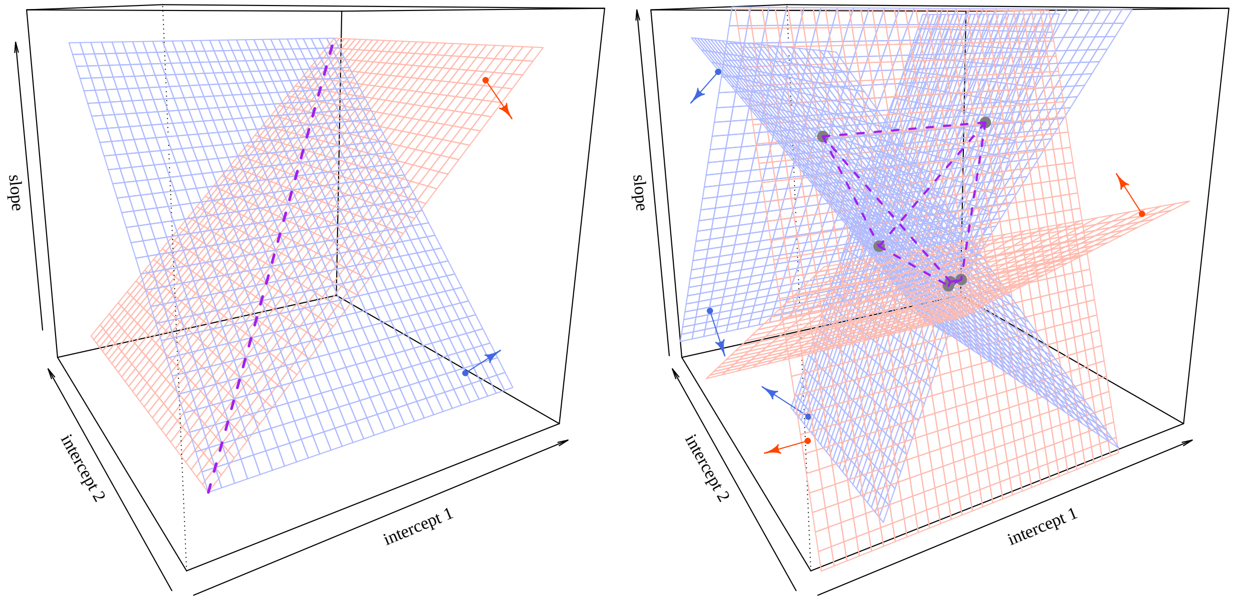

Figure 2.1: Set inverse functions for a single response entry Y_{ij} (left) and five i.i.d. responses (right) for a 3-category graded item. Colors of the wireframes indicate directions of inequality signs, and arrows point into the corresponding half-spaces. On the left, the purple-colored dashed line gives a boundary of the parameter space $\alpha_{j1} = \alpha_{j2}$. On the right, the intersection of all half-spaces is shown as the polytope surrounded by its purple-colored edges and highlighted vertices

by the intersection of all individual set inverse functions (Equation 2.2):

$$Q_j(\mathbf{y}_{(j)}, \mathbf{a}_{(j)}, \mathbf{z}) = \bigcap_{j=1}^{m} Q_{ij}(y_{ij}, a_{ij}, \mathbf{z}) \tag{2.3}$$

in which $\mathbf{a}_{(j)} = (a_{ij})_{i=1}^{n}$ and $\mathbf{z} = (\mathbf{z}_i)_{i=1}^{n}$ are realizations of the logistic and normal random variables. We take the intersection for a reason similar to that discussed in the binomial proportion example: Because the same intercept and slope parameters appear in the data generating equations of all n responses $(Y_{ij})_{i=1}^{n}$, the set inverse should contain values of those item parameters that are consistent with all the equations. The right panel of Figure 2.1 depicts the set inverse for five responses to the same three-category item: A three-dimensional closed polyhedron is obtained as the intersection of the corresponding half-spaces.

Finally, consider a sample of i.i.d. responses $\mathbf{Y} = (\mathbf{Y}_i)_{i=1}^{n} = (Y_{ij})_{i=1}^{n}{}_{j=1}^{m}$ to a test of

m graded items. Because we assume that items do not share parameters, the set inverse function for the entire set of item response data

$$Q(\mathbf{y}, \mathbf{a}, \mathbf{z}) = \bigtimes_{j=1}^{m} Q_j(\mathbf{y}_{(j)}, \mathbf{a}_{(j)}, \mathbf{z}) \tag{2.4}$$

is a subset of the entire parameter space $\Theta \subset \mathbb{R}^{q_1} \times \cdots \times \mathbb{R}^{q_m}$, in which $\times$ denotes the Cartesian product. A generalized fiducial distribution can be constructed following the general recipe (Equation 1.6):

$$\mathbf{v}(Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)) \mid \{Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}, \tag{2.5}$$

in which $\mathbf{v}(\cdot)$ denotes some user-defined rule that selects one point from each item's polyhedron. $\mathbf{A}^\star$ and $\mathbf{Z}^\star$ are i.i.d. copies of $\mathbf{A}$ and $\mathbf{Z}$, respectively; again, the asterisks are used to distinguish them from their data-generating counterparts. Note that both $\mathbf{A}^\star$ and $\mathbf{Z}^\star$ are continuous random variables, and thus we do not differentiate $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)$ from its closure: i.e., the polyhedrons with attained boundaries. In the sequel, call a random variable following the distribution given by Equation 2.5 a generalized fiducial quantity (GFQ), denoted $\mathbf{R}$.

In finite samples, the polyhedron implied by Equation 2.3 is not necessarily bounded; for example, when $n \leq q_j$, it is certainly unbounded, because a bounded polytope on the q_j-dimensional space has at least $q_j + 1$ faces. We require a finite point being returned by the selection rule: i.e., $|\mathbf{v}(Q(\mathbf{y}, \mathbf{a}, \mathbf{z}))| < \infty$ for all $\mathbf{y}$, $\mathbf{a}$ and $\mathbf{z}$ such that $Q(\mathbf{y}, \mathbf{a}, \mathbf{z})$ is non-empty; hence, infinity is not included in the support of the resulting fiducial distribution (Equation 2.5) in unbounded cases. Eventually, the polyhedrons become bounded as the sample size tends to infinity; it is in fact a corollary of Theorem 2 which states a stronger property that the diameter of the set inverse shrinks to zero at a fast rate.

In addition, it is plausible that the set inverse (Equation 2.4) touches the boundary of the parameter space imposed by the ordering of item intercepts, i.e., $\partial\Theta = \{\boldsymbol{\theta} : \alpha_{jk} = \alpha_{j,k+1}$ for some j and $k\}$. In large samples, however, this almost never happens. In fact, if there exists more than one endorsement to a response category k of item j, e.g., $y_{1j} = y_{2j} = k$, then $\alpha_{j,k+1} < \min\{A^\star_{1j} - \boldsymbol{\beta}_j{}^\top \mathbf{Z}^\star_1, A^\star_{2j} - \boldsymbol{\beta}_j{}^\top \mathbf{Z}^\star_2\} \leq \max\{A^\star_{1j} - \boldsymbol{\beta}_j{}^\top \mathbf{Z}^\star_1, A^\star_{2j} - \boldsymbol{\beta}_j{}^\top \mathbf{Z}^\star_2\} \leq \alpha_{jk}$, with a strict inequality attained almost surely by the continuous nature of the logistic and normal variates. As long as the data-generating values of the item parameters are in the interior of the parameter space, all the response patterns happen with a positive probability, and thus the set inverse is eventually bounded away from $\partial\Theta$ with probability one.

Some extra notation is introduced. Let $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i) = \alpha_{jk} + \boldsymbol{\beta}_j{}^\top \mathbf{z}_i$ be the linear regression on the latent variable. Also set $\tau_{j0}(\cdot,\cdot) = \infty$ and $\tau_{jK_j}(\cdot,\cdot) = -\infty$ by convention; with the help of these notations, we simplify the IRF (Equation 1.1) to

$$f_j(\boldsymbol{\theta}_j, k|\mathbf{z}_i) = \frac{1}{1 + e^{-\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i)}} - \frac{1}{1 + e^{-\tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}_i)}}, \tag{2.6}$$

and the set inverse function (Equation 2.2) to

$$Q_{ij}(y_{ij}, a_{ij}, z_i) = \{\boldsymbol{\theta}_j \in \mathbb{R}^{q_j} : \tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}_i) < a_{ij} \leq \tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i)\}. \tag{2.7}$$

Consider $\mathbf{y}$ fixed for now. Each vertex of the possibly unbounded $\mathbb{R}^{q_j}$-polyhedron $Q_j(\mathbf{y}_{(j)}, \mathbf{a}_{(j)}, \mathbf{z})$ residing in the interior of Θ is the solution of a set of q_j linear equations of form $a_{ij} = \tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i)$, contributed from q_j observations and some suitable choices of left/right bounds depending on the responses of those selected observations[2]. Notationally, let I_j be a size-q_j sub-sample of observations. Also let $\mathbf{k}_{I_j} = (k_{ij})_{i \in I_j}$ be an index tuple of length q_j, each element of which $k_{ij} \in \{y_{ij}, y_{ij} + 1\}$ indicates whether the right half-space $a_{ij} \leq \tau_{jy_{ij}}(\boldsymbol{\theta}_j, \mathbf{z}_i)$ or the left half-space $\tau_{j,y_{ij}+1}(\boldsymbol{\theta}_j, \mathbf{z}_i) < a_{ij}$ is selected for each $i \in I_j$[3]. Only a small fraction of $(I_j, \mathbf{k}_{I_j})$ pairs are needed to determine a vertex: It only happens when the q_j boundary hyperplanes of the selected half-spaces are finite and produce a non-singular linear system.

---

[2] If an observation contributes two equations, then the resulting vertex is on the boundary of Θ. This happens with probability zero for sufficiently large n.

[3] Here, I_j is treated as an unordered set, while $\mathbf{k}_{I_j}$ is a q_j-tuple, each element of which, i.e., k_{ij}, uniquely maps onto an element $i \in I_j$.

Suppose a fixed non-empty set inverse function $Q_j(\mathbf{y}_{(j)}, \mathbf{a}_{(j)}, \mathbf{z})$ has v_j vertices; they can be indexed by a collection of properly selected pairs $\mathcal{P}_j = \{(I_j^{(l)}, \mathbf{k}_{I_j^{(l)}})\}_{l=1}^{v_j}$. Pooling across all m items in the test, write $I = \bigcup_{j=1}^m I_j$, $\mathbf{k}_I = (\mathbf{k}_{I_j})_{j=1}^m$, and $\mathcal{P} = \bigtimes_{j=1}^m \mathcal{P}_j$; $\mathcal{P}$ indexes all extremal points of the set inverse $Q(\mathbf{y}, \mathbf{a}, \mathbf{z})$ for the entire set of item response data.

We first consider selection rules that yield interior extremal points of $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)$ and are independent of $\mathbf{A}^\star$ and $\mathbf{Z}^\star$, for which the resulting fiducial density has a closed-form expression. Write the selected point $\mathbf{V} = (\mathbf{V}_j)_{j=1}^m = \mathbf{v}(Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star))$, in which $\mathbf{V}_j$ is the selected vertex of polyhedron $Q_j(\mathbf{y}_{(j)}, \mathbf{A}^\star_{(j)}, \mathbf{Z}^\star)$. For each item j, let $\mathbf{V}_{I_j,\mathbf{k}_{I_j}}$ be the solution determined by a particular sub-sample I_j together with a particular combination of left/right bounds $\mathbf{k}_{I_j}$, and $E_{I_j,\mathbf{k}_{I_j}}$ be the event that $\mathbf{V}_{I_j,\mathbf{k}_{I_j}}$ gives an interior vertex of $Q_j(\mathbf{y}_{(j)}, \mathbf{A}^\star_{(j)}, \mathbf{Z}^\star)$. Also let $\mathbf{V}_{I,\mathbf{k}_I} = (\mathbf{V}_{I_j,\mathbf{k}_{I_j}})_{j=1}^m$, $E_{I,\mathbf{k}_I}$ denote the event that $\mathbf{V}_{I_j,\mathbf{k}_{I_j}}$ determines an extremal point of $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)$, and $E_{\mathcal{P}}$ denote the event that all the extremal points are indexed by $\mathcal{P}$. The generic fiducial quantity (Equation 2.5) associated with a selected extremal point of the set inverse function (Equation 2.4) is given in the following lemma; see Appendix A for the derivation, which is similar in spirit to the proof of Lemma A.1 and B.1 in Hannig (2013), the first part of Theorem 1 in E et al. (2009), and Lemma 1 in Liu and Hannig (2014).

Lemma 1. *Consider m graded items each of which is characterized by equation 2.6. Let $\Theta \subset \mathbb{R}^{q_1} \times \cdots \times \mathbb{R}^{q_m}$ be the parameter space of item parameters $\boldsymbol{\theta}$, comprising all free item intercepts and slopes. We observe i.i.d. ordinal item response data $\mathbf{y} = (\mathbf{y}_i)_{i=1}^n$, in which each response category of each item has more than one endorsement. Then, the density of the GFQ corresponding to a selected extremal point, i.e., $\mathbf{V} \mid \{Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}$, can be* *written as*[4]

$$\begin{aligned} g_n(\boldsymbol{\theta}|\mathbf{y}) \propto & \sum_{I} \sum_{\mathbf{k}_I} w_{I,\mathbf{k}_I}(\mathbf{y}) \\ & \cdot \int_{\mathbb{R}^{nr}} \prod_{j=1}^{m} \left\{ d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}_j, \mathbf{z}_{I_j}) \cdot \prod_{i \in I_j} \frac{e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j, \mathbf{z}_i)}}{[1 + e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j, \mathbf{z}_i)}]^2} \right. \\ & \left. \cdot \prod_{i \in I_j^c} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) \right\} d\Phi(\mathbf{z}). \end{aligned} \tag{2.8}$$

In Equation 2.8, $d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}_j, \mathbf{z}_i) = \left|\det(\partial\tau_{jk_{ij}}(\boldsymbol{\theta}_j, \mathbf{z}_{I_j})/\partial\boldsymbol{\theta}_j)_{i\in I_j}\right|$, $\mathbf{z}_{I_j} = (\mathbf{z}_i)_{i\in I_j}$, $\Phi(\cdot)$ *denotes a standard normal probability measure*[5]*, and*

$$\begin{aligned} w_{I,\mathbf{k}_I}(\mathbf{y}) &= P\{\mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} | E_{I,\mathbf{k}_I}\} \\ &\propto \sum_{\mathcal{P} \ni (I,\mathbf{k}_I)} P\{\mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} | E_{\mathcal{P}}\} \cdot P\{E_{\mathcal{P}}\} \end{aligned} \tag{2.9}$$

is contingent upon $P\{\mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} | E_{\mathcal{P}}\}$*, i.e., the specific selection rule being used.*

Remark 1. The connection between GFI and Bayesian inference can be seen from Equation 2.8. As a general notation, we put index set in subscript to denote the corresponding

---

[4] The sum on the right-hand side of Equation 2.8 is taken over all combinations of I and $\mathbf{k}_I$. Some of $(I_j, \mathbf{k}_{I_j})$ pairs are not able to produce a vertex; in those cases, the Jacobian determinant $d_{I_j,\mathbf{k}_{I_j}}$ is zero, and thus the corresponding summand in Equation 2.8 vanishes.

[5] Here, the dimensionality of the random variable is suppressed for succinctness.

elements: e.g., $\mathbf{z}_I = (\mathbf{z}_i)_{i\in I}$. Rewrite Equation 2.8 by splitting the integral into two parts—one for $\mathbf{z}_I$, and the other for $\mathbf{z}_{I^c}$:

$$\begin{aligned} g_n(\boldsymbol{\theta}|\mathbf{y}) \propto & \sum_I \sum_{\mathbf{k}_I} w_{I,\mathbf{k}_I}(\mathbf{y}) \int \prod_{j=1}^{m} d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}_j, \mathbf{z}_i) \\ & \cdot \prod_{i\in I} \left\{ \prod_{j\in J(i)} \frac{e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}}{[1+e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}]^2} \prod_{j\notin J(i)} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) \right\} d\Phi(\mathbf{z}_I) \\ & \cdot \int \prod_{i\in I^c} \prod_{j=1}^{m} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}_{I^c}), \end{aligned} \tag{2.10}$$

in which $J(i) = \{j : i \in I_j\}$ for $i \in I$. Note that the last line of Equation 2.10 is the marginal likelihood function of the remaining observations I^c. We can multiply and divide the right-hand side of Equation 2.10 by the likelihood of the vertex-determining observations I, and then simplify it to

$$g_n(\boldsymbol{\theta}|\mathbf{y}) \propto b_n(\boldsymbol{\theta}, \mathbf{y}) f_n(\boldsymbol{\theta}, \mathbf{y}). \tag{2.11}$$

In Equation 2.11,

$$f_n(\boldsymbol{\theta}, \mathbf{y}) = \int \prod_{i=1}^{n} \prod_{j=1}^{m} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}), \tag{2.12}$$

denotes the complete sample likelihood, and

$$\begin{aligned} b_n(\boldsymbol{\theta}, \mathbf{y}) = & \sum_I \sum_{\mathbf{k}_I} w_{I,\mathbf{k}_I}(\mathbf{y}) \int \prod_{j=1}^{m} d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}_j, \mathbf{z}_{I_j}) \\ & \prod_{i\in I} \left\{ \prod_{j\in J(i)} \frac{e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}}{[1+e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}]^2} \prod_{j\notin J(i)} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) \right\} d\Phi(\mathbf{z}_I) \\ & \Big/ \int \prod_{i\in I} \prod_{j=1}^{m} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}_I). \end{aligned} \tag{2.13}$$

is a function of both the item parameters and data. Therefore, our fiducial distribution for parameters, although obtained from a seemingly unrelated argument, can be conceived as the (empirical) Bayesian posterior computed from the data-dependent prior proportional to Equation 2.13.

To simplify the proof of our main theorem (Theorem 1), we impose a further restriction on the selection rules: Whenever $I \neq I'$ but $\mathbf{y}_I = \mathbf{y}_{I'}$ and $\mathbf{k}_I = \mathbf{k}_{I'}$, it is required that $w_{I,\mathbf{k}_I}(\mathbf{y}) = w_{I',\mathbf{k}_{I'}}(\mathbf{y})$; the common function value is denoted by $w_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{y})$. It implies that the number of distinct values of $w_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{y})$ does not grow with the sample size, because $\mathbf{y}_I$ and $\mathbf{k}_I$ have only finitely many patterns. A simple selection rule that satisfies such condition, which is also recommended in actual computation, is to select with equal probability among the interior vertices of each polyhedron $Q_j(\mathbf{y}_{(j)}, \mathbf{A}^\star_{(j)}, \mathbf{Z}^\star)$.

The next result implies that the fiducial density (Equation 2.8) is invariant under smooth transformations, similar to the invariance of the Bayesian posterior derived from the Jeffreys prior. This is a desirable property because inference about item parameters remains the same when the model is re-parameterized. For example, researchers may be interested in the alternative slope-difficulty or the standardized loading-threshold parameterizations of the GRM model, and inference can be safely drawn about those transformed parameters using the correspondingly transformed generalized fiducial distributions. Definitions of those specific transformations are provided later in Chapter 4.

Proposition 1 (Invariance). *Let* $\boldsymbol{\theta} = \mathbf{q}(\boldsymbol{\xi})$ *be a one-to-one and continuously differentiable function onto the parameter space* Θ. *Denote the data generating equation corresponding to Equation 2.8 by* $\mathbf{Y} = \mathbf{g}(\boldsymbol{\theta}, \mathbf{A}, \mathbf{Z})$, *and write* $\tilde{g}_n(\boldsymbol{\xi}|\mathbf{y})$ *as the generalized fiducial distribution computed from the data generating equation* $\mathbf{Y} = \mathbf{g}(\mathbf{q}(\boldsymbol{\xi}), \mathbf{A}, \mathbf{Z})$. *Then for any measurable set* $B \subset \Theta$,

$$\int_B g_n(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int_{\mathbf{q}^{-1}(B)} \tilde{g}_n(\boldsymbol{\xi}|\mathbf{y}) d\boldsymbol{\xi}. \tag{2.14}$$

2.2 A fiducial Bernstein-von Mises theorem

Now we are ready to expound our major theoretical result, a fiducial Bernstein-von Mises theorem, which describes the asymptotic optimality and normality of the fiducial distribution and further implies the large-sample correctness of fiducial interval estimators in the frequentist sense. In this section, we start with the introduction of some notation and a heuristic description of the Bernstein-von Mises phenomenon. Then, we provide the formal statement of the theorem (Theorem 1) based on the fiducial density (Equation 2.8) that has been derived in Lemma 1. We conclude that the result is applicable regardless of the selection rule being used, due to the fact that the diameter of the set inverse function is a higher order term (Theorem 2).

Some standard notation is needed for the asymptotic theory. The (marginal) multinomial likelihood for each response pattern $\mathbf{y}_i$ is expressed as Equation 1.2. Let $\mathbf{s}(\boldsymbol{\theta}, \mathbf{y}_i) = \partial \log f(\boldsymbol{\theta}, \mathbf{y}_i)/\partial\boldsymbol{\theta}$ be the single-observation score vector, and $\mathbf{H}(\boldsymbol{\theta}, \mathbf{y}_i) = \partial^2 \log f(\boldsymbol{\theta}, \mathbf{y}_i)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ be the single-observation Hessian matrix. Also define $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}\left[\mathbf{s}(\boldsymbol{\theta}, \mathbf{Y}_i)\right]$ which is usually referred to as the Fisher information matrix. It can be verified by direct calculation that

$$\begin{aligned} &E_{\boldsymbol{\theta}}\left[\mathbf{s}(\boldsymbol{\theta}, \mathbf{Y}_i)\right] = \mathbf{0}, \\ &\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\mathbf{s}(\boldsymbol{\theta}, \mathbf{Y}_i)\mathbf{s}(\boldsymbol{\theta}, \mathbf{Y}_i)^\top\right] = -E_{\boldsymbol{\theta}}\left[\mathbf{H}(\boldsymbol{\theta}, \mathbf{Y}_i)\right]. \end{aligned} \tag{2.15}$$

Also let $\boldsymbol{\theta}_0$ be the data-generating parameter values, $\boldsymbol{\mathcal{I}}_0$ be a shorthand notation for $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)$.

Loosely speaking, the Bernstein-von Mises phenomenon refers to the fact that in large samples the random probability measure corresponding to the properly centered GFQ, $\sqrt{n}(\mathbf{R} - \boldsymbol{\theta}_0)$, "converges" to a normal distribution, $\mathcal{N}(\mathbf{X}, \boldsymbol{\mathcal{I}}_0^{-1})$, whose mean $\mathbf{X}$ is a random quantity following a normal distribution with zero mean and the efficient covariance matrix $\boldsymbol{\mathcal{I}}_0^{-1}$. It can be inferred that a proper central tendency measure (e.g., the median) of the fiducial distribution is asymptotically equivalent to the ML estimator, and that CIs constructed from the fiducial distribution have the correct frequentist coverage asymptotically.

The appropriate mode of convergence involved in the foregoing heuristics is that the total variation distance between the density of $\sqrt{n}(\mathbf{R} - \boldsymbol{\theta}_0)$ and $\mathcal{N}(\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{S}_n, \boldsymbol{\mathcal{I}}_0^{-1})$ converges to zero in $P_{\boldsymbol{\theta}_0}$-probability, in which the sample score function $\mathbf{S}_n$ satisfies

$$\mathbf{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\boldsymbol{\theta}_0, \mathbf{Y}_i)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_0) \tag{2.16}$$

by the Central Limit Theorem. $\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{S}_n$ serves as a finite-sample "centering sequence" for $\sqrt{n}(\mathbf{R} - \boldsymbol{\theta}_0)$ in place of the limiting version $\mathbf{X}$ in the heuristics. See Appendix B for the proof of the theorem, which is similar in spirit to Ghosh and Ramamoorthi's (2003, Theorem 1.4.2) proof of a Bayesian Bernstein-von Mises theorem.

Theorem 1 (Bernstein-von Mises)**.** *Suppose that item response data* $\mathbf{Y} = (\mathbf{Y}_i)_{i=1}^n$ *are i.i.d. with probability mass function* $f(\boldsymbol{\theta}_0, \mathbf{y}_i)$*. Let* $\Theta \subset \mathbb{R}^q$ *be the parameter space as usual. Assume that*

(i) $m \geq r + 1$*;*

(ii) For all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ *such that* $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, $f_{\boldsymbol{\theta}} \neq f_{\boldsymbol{\theta}'}$ *for some response pattern;*

(iii) $\boldsymbol{\theta}_0$ *is at the interior of* Θ*;*

(iv) The Fisher information matrix $\boldsymbol{\mathcal{I}}_0$ *is positive definite.*

Let $\bar{g}_n(\mathbf{h}|\mathbf{y}) = g_n(\boldsymbol{\theta}_0 + \mathbf{h}/\sqrt{n}|\mathbf{y})/\sqrt{n}$ *be the fiducial density of* $\sqrt{n}(\mathbf{R} - \boldsymbol{\theta}_0)$*,* H_n *be the correspondingly rescaled parameter space, and* $\phi_{\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{S}_n, \boldsymbol{\mathcal{I}}_0^{-1}}$ *be the density of* $\mathcal{N}(\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{S}_n, \boldsymbol{\mathcal{I}}_0^{-1})$*. Then,*

$$\int_{H_n} \left|\bar{g}_n(\mathbf{h}|\mathbf{Y}) - \phi_{\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{S}_n, \boldsymbol{\mathcal{I}}_0^{-1}}(\mathbf{h})\right| d\mathbf{h} \xrightarrow{P_{\boldsymbol{\theta}_0}} 0, \tag{2.17}$$

Remark 2. Assumptions (ii) to (iv) are standard regularity conditions for establishing the asymptotic optimality of the ML estimator. (i) ensures the existence of some neighborhood of $\boldsymbol{\theta}_0$ such that for $\boldsymbol{\theta}$ outside the likelihood ratio statistic $f_n(\boldsymbol{\theta}, \mathbf{Y})/f_n(\boldsymbol{\theta}_0, \mathbf{Y})$ uniformly goes to zero; this is similar to Assumption (v) in Ghosh and Ramamoorthi (2003). Also, for some choices of K_j and r, (i) is implied by (ii).

Remark 3. As remarked in van der Vaart (2000, Section 10.2), the alternative "centering sequence" $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, in which $\hat{\boldsymbol{\theta}}$ is the ML estimator, can be used in place of $\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{S}_n$ in

Equation 2.17, because the the latter is a local linear approximation of the former at the true parameter values $\boldsymbol{\theta}_0$ and the two are asymptotically equivalent.

The next theorem dictates that the diameter of the set inverse $Q(\mathbf{Y}, \mathbf{A}^\star, \mathbf{Z}^\star)$ goes to 0 at the rate $1/n$, higher than the rate $1/\sqrt{n}$ at which the fiducial distribution approaches its normal limit. Consequently, different selection rules tend to give converged inference about model parameters when the sample size is large enough.

Theorem 2. *Suppose that assumptions (i)–(iv) of Theorem 1 hold. For any* $K > 0$*, define*

$$\rho_K(\mathbf{y}) = P^\star\{\mathrm{diam} Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) > K/n \mid Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}, \tag{2.18}$$

in which $P^\star$ *denotes the probability measure of the generated variates* $\mathbf{A}^\star$ *and* $\mathbf{Z}^\star$*. Then, for each* $\varepsilon > 0$*,*

$$P_{\boldsymbol{\theta}_0}\{\exists K, N > 0 : \ \rho_K(\mathbf{Y}) < \varepsilon, \ \forall n > N\} \to 1, \tag{2.19}$$

in which $P_{\boldsymbol{\theta}_0}$ *denotes the probability measure of* $\mathbf{Y}$ *under the true parameter values* $\boldsymbol{\theta}_0$*.*

Remark 4. We only establish the proof for unidimensional GRMs (i.e., $r = 1$), which is relegated to Appendix C. We conjecture that a similar proof using more sophisticated geometric argument can be established for multidimensional models.

2.3 Fiducial predictive inference

In this section, we discuss predictive inference using the derived fiducial distribution for item parameters. In the extant literature, prediction is typically defined as making inferences about future observations, or statistics computed from future observations, based on their distributions conditional on the values already observed; such distributions are likely to depend on unknown parameters that need to be estimated from the observed data (e.g., Aitchison and Dunsmore, 1975; Geisser, 1993). In the current discussion, we focus on constructing prediction intervals (PIs) for the target future data/statistics for the purpose of quantifying prediction errors. We provide a general consistency theorem for predictive densities computed from any consistent distributions of model parameters under mild smoothness requirements for the target statistics. As a corollary of the fiducial Bernstein-von Mises theorem, the GFQ is consistent; therefore, its use in predictive inference is justified. An important application of fiducial predictive inference is to obtain PIs for latent variable scores, which quantify the precision of the measurement for the substantive construct(s) of interest. With GFI, a Monte Carlo sample of the predictive distribution for each respondent's response pattern score is available as a by-product of the sampling algorithm described in Chapter 3.

2.3.1 Consistency

The following proposition is applicable to the prediction for any test statistics $\mathbf{T}$ whose density function $h(\mathbf{t}, \boldsymbol{\theta}_0)$ with respect to some dominating measure μ depends on the data-generating parameter values $\boldsymbol{\theta}_0$. We claim that the *predictive density*:

$$h_n(\mathbf{t}|\mathbf{y}) = \int_{\Theta} h(\mathbf{t}, \boldsymbol{\theta}) g_n(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \tag{2.20}$$

derived from a consistent distribution defined on the parameter space with density $g_n(\boldsymbol{\theta}|\mathbf{y})$ converges in total variation to the target density $h(\mathbf{t}, \boldsymbol{\theta}_0)$ in $P_{\boldsymbol{\theta}_0}$-probability, provided in some small neighborhood of $\boldsymbol{\theta}_0$ the density function $h(\mathbf{t}, \boldsymbol{\theta})$ is continuous and dominated by an integrable function.

Proposition 2 (Predictive consistency)**.** *Let* $\{g_n(\boldsymbol{\theta}|\mathbf{y})\}$ *be a consistent sequence of density functions at* $\boldsymbol{\theta}_0$ *in the sense that as* $n \to \infty$*,*

$$\int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta} g_n(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} \to 1 \textit{ in } P_{\boldsymbol{\theta}_0}\textit{-probability.} \tag{2.21}$$

for all $\delta > 0$*. Let* $\mathbf{T}$ *be a statistic having density* $h(\mathbf{t}, \boldsymbol{\theta}_0)$ *with respect to some dominating measure* μ*. Assume that there exists some neighborhood of* $\boldsymbol{\theta}_0$*, denoted* N_0*, such that* $h(\mathbf{t}, \boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$ *for each fixed* $\mathbf{t}$*, and that there also exists a measurable function* $e(\mathbf{t})$ *such that*

$$\sup_{\boldsymbol{\theta}\in N_0} h(\mathbf{t}, \boldsymbol{\theta}) \le e(\mathbf{t}), \textit{ and } \int e(\mathbf{t})\mu(d\mathbf{t}) < \infty. \tag{2.22}$$

Then the predictive density, i.e., Equation 2.20, satisfies

$$\int |h_n(\mathbf{t}|\mathbf{Y}) - h(\mathbf{t}, \boldsymbol{\theta}_0)|\mu(d\mathbf{t}) \to 0 \textit{ in } P_{\boldsymbol{\theta}_0}\textit{-probability.} \tag{2.23}$$

Remark 5. As a consequence of Theorem 1 and 2, the GFQ defined by Equation 2.5 with any selection rule $\mathbf{v}(\cdot)$ satisfies Equation 2.21, and thus can be used for making predictions about statistic $\mathbf{T}$. The proposition can be considered as an extension to Theorem 1 in Wang, Hannig, and Iyer (2012).

Remark 6. We need the mild continuity requirement and Equation 2.22 to apply the Dominated Convergence Theorem in the proof which is rather straightforward and can be found in Appendix D. In applications of the GRM, those conditions are typically easy to check.

Remark 7. Proposition 2 can be used to check the compatibility of the calibrated GRM in independent cross-validation samples. Statistics that are linear combinations of response patterns, e.g., marginal response patterns, are often used for this purpose. In those cases, the continuity of the target density is guaranteed by the continuity of the response pattern likelihood function (Equation 1.2), and Equation 2.22 is trivially satisfied because this type of statistics take only finitely many possible values. However, Proposition 2 cannot be applied to probing the fit to the current data set; intuitively it is because the current data set has already been used to obtain the predictive distribution, and thus its reuse in fit checking leads to unresolved dependencies. More involved techniques, i.e., the fiducial predictive check which is introduced in the next section, must be invoked in this scenario.

In practice, the predictive distribution (Equation 2.20) can be conveniently approximated by Monte Carlo simulations, especially when a sample from the consistent distribution $g_n(\boldsymbol{\theta}|\mathbf{y})$ is available and the test statistic $\mathbf{T}$ is some simple function of the data $\mathbf{t}(\mathbf{Y})$. For easy reference, we provide the generic pseudo-code, i.e., Algorithm 1, for constructing Monte Carlo percentile PIs.

Algorithm 1 Monte Carlo PIs

```
generate $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(S)}$ from distribution $g_n(\boldsymbol{\theta}|\mathbf{y})$.
for all $s = 1, \dots, S$ do
    if $\mathbf{T} = \mathbf{t}(\mathbf{Y})$ then
        generate $\mathbf{A}^\star = \mathbf{a}^{(s)}$, $\mathbf{Z}^\star = \mathbf{z}^{(s)}$.
        compute $\mathbf{y}^{(s)} = \mathbf{g}(\boldsymbol{\theta}^{(s)}, \mathbf{a}^{(s)}, \mathbf{z}^{(s)})$, $\mathbf{t}^{(s)} = \mathbf{t}(\mathbf{y}^{(s)})$.
    else
        generate $\mathbf{T} = \mathbf{t}^{(s)}$ from $h(\mathbf{t}, \boldsymbol{\theta}^{(s)})$
    end if
end for
construct PIs with empirical percentiles of $\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(S)}$.
```

2.3.2 Example: Response pattern scoring

Next, we discuss the use of fiducial predictive inference in the interval estimation of response pattern scores. When the items are already calibrated and the parameters are considered known, inference about individual pattern scores $\mathbf{z}_i$ is a Bayesian problem and can be obtained from the posterior density of $\mathbf{Z}_i$ given $\mathbf{y}_i$:

$$p(\mathbf{z}_i, \boldsymbol{\theta}|\mathbf{y}_i) \propto \phi(\mathbf{z}_i) \prod_{j=1}^{m} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) \tag{2.24}$$

evaluated at $\boldsymbol{\theta}_0$, in which the standard normal density $\phi(\cdot)$ serves as the prior density of this Bayesian problem. Based on Equation 2.24 the posterior mean is typically used as a point estimate of $\mathbf{z}_i$ and often referred to as the expected *a posteriori* (EAP) score; interval estimates of $\mathbf{z}_i$ can be constructed by numerically computing the quantiles of the posterior or by a normal approximation using the EAP score and the posterior standard deviation. The resulting posterior intervals are asymptotically normal and efficient as a consequence of the Bayesian Bernstein-von Mises theorem (e.g., Le Cam & Yang, 1986; van der Vaart, 2000), given the true item parameter values satisfying some mild conditions. In the situation when item parameters need to be simultaneously estimated from the data, we resort to making predictive inference about the posterior (Equation 2.24) by substituting $p(\mathbf{z}_i, \boldsymbol{\theta}|\mathbf{y}_i)$ for $h(\mathbf{t}, \boldsymbol{\theta})$ in Equation 2.20.

In order to use Proposition 2, the local behavior of the posterior density $p(\mathbf{z}_i, \boldsymbol{\theta}|\mathbf{y}_i)$ in the neighborhood of $\boldsymbol{\theta}_0$ must be checked. The continuity of the posterior density with respect to $\boldsymbol{\theta}$ is obvious from expression 2.24. Condition 2.22 is also satisfied due to the facts: a) $f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) \leq 1$, b) the standard normal density is integrable, and c) there are only finitely many patterns of $\mathbf{y}_i$ for a fixed-length test, and thus the likelihood function values in a neighborhood of $\boldsymbol{\theta}_0$ are bounded from below. Then Proposition 2 guarantees the asymptotic correctness of the corresponding predictive inference using Equation 2.20.

Next, we claim that the conditional distribution $\mathbf{Z}_i^\star \mid \{Q(\mathbf{Y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}$ is asymptotically equivalent to the fiducial predictive distribution (Equation 2.20), given fixed $\mathbf{Y}_i = \mathbf{y}_i$. The posterior distribution (Equation 2.24) can be alternatively interpreted as the marginal distribution of $\mathbf{Z}_i^\star$ when $\mathbf{Z}_i^\star$ and $\mathbf{A}_i^\star$ are generated such that $\boldsymbol{\theta}_j{}^\top \tilde{\mathbf{Z}}_{ij,k+1}^\star \leq A_{ij}^\star \leq \boldsymbol{\theta}_j{}^\top \tilde{\mathbf{Z}}_{ijk}^\star$ for all j, in which $\tilde{\mathbf{Z}}_{ijk}^\star$ is the random version of $\tilde{\mathbf{z}}_{ijk}$ as appeared in Equation 2.8. Denote by a subscript $-i$ the components corresponding to all but the ith observation. Then $\mathbf{Z}_i^\star \mid \{Q(\mathbf{Y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}$ is in fact the marginal distribution of $\mathbf{Z}_i^\star$ conditional on the event that there exists some $\boldsymbol{\theta} = (\boldsymbol{\theta}_j)_{j=1}^m \in Q(\mathbf{Y}_{-i}, \mathbf{A}_{-i}^\star, \mathbf{Z}_{-i}^\star)$ such that $\boldsymbol{\theta}_j{}^\top \tilde{\mathbf{Z}}_{ij,k+1}^\star \leq A_{ij}^\star \leq \boldsymbol{\theta}_j{}^\top \tilde{\mathbf{Z}}_{ijk}^\star$ for each j. The asymptotic equivalence follows as a corollary of Theorem 1 and 2. As a result, for observed response patterns, PIs for the corresponding scores can be obtained along with sampling from the fiducial distribution of item parameters (the detailed algorithm is relegated to Chapter 3). For patterns not present in the calibration sample, however, extra Monte Carlo simulations using Algorithm 1 are necessary.

The joint consistency or asymptotic normality of the fiducial distribution in estimating $\boldsymbol{\theta}$ and $\mathbf{z}$ as both n and m tend to infinity is beyond the scope of the current work. We conjecture that a more general fiducial Bernstein-von Mises theorem for item parameters can be established as the dimension of the parameter space grows to infinity at a slow enough rate, and that consequently interval estimators for both item parameters and response pattern scores are both asymptotically correct.

2.4 Goodness of fit testing with a fiducial predictive check (FPC)

As mentioned in the previous section, we are tempted to check the compatibility of the fitted GRM to the observed item response data using a suitable test statistic $T = t(\mathbf{Y})$ via predictive simulations. In particular, we are interested in approximating the *predictive p-value*:

$$p(\mathbf{y}) = \int P_{\boldsymbol{\theta}}\{t(\mathbf{Y}) > t(\mathbf{y})\} g_n(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \tag{2.25}$$

after observing $\mathbf{Y} = \mathbf{y}$, in which $P_{\boldsymbol{\theta}}\{\cdot\}$ highlights the probability calculation under parameter values $\boldsymbol{\theta}$. We call the resulting procedure a *fiducial predictive check* (FPC), inspired by the posterior predictive check in Bayesian statistics. The pseudo-code is provided as Algorithm 2, the structure of which closely resembles Algorithm 1. Note that the formula for a one-tailed empirical p-value is given in Line 7 of Algorithm 1; whenever desired, two-tailed p-values can be obtained in a similar manner.

Algorithm 2 Fiducial predictive check (FPC)

```
generate $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(S)}$ from distribution $g_n(\boldsymbol{\theta}|\mathbf{y})$.
for all $s = 1, \dots, S$ do
    generate $\mathbf{A}^\star = \mathbf{a}^{(s)}$, $\mathbf{Z}^\star = \mathbf{z}^{(s)}$.
    compute $\mathbf{y}^{(s)} = \mathbf{g}(\boldsymbol{\theta}^{(s)}, \mathbf{a}^{(s)}, \mathbf{z}^{(s)})$, $t^{(s)} = t(\mathbf{y}^{(s)})$.
end for
compute the observed statistic $t = t(\mathbf{y})$
compute the empirical $p$-value $\hat{p}(\mathbf{y}) = S^{-1} \sum_{s=1}^{S} \mathbb{I}\{t^{(s)} > t\}$
```

As pointed out by Bayarri and Berger (2000) and Robins, van der Vaart, and Ventura (2000) in the context of Bayesian posterior predictive check, the p-value calculated from Equation 2.25 is not always asymptotically uniform, because the observed data $\mathbf{y}$ are effectively used in both computing the statistic and obtaining its predictive distribution. There have been philosophical disputes among Bayesian statisticians about the necessity for a posterior p-value to be uniform (see e.g., Bayarri and Berger, 1999; Gelman, 2013); however, FPC is treated as a freqentist method in the current work, so we consider asymptotically uniform p-values desirable. Two tweaks for Algorithm 2 are introduced next, both of which are based on the theoretical discussion traced to Robins et al. (2000).

2.4.1 The centering approach

Robins et al. considered a family of asymptotically normal test statistics $t(\mathbf{Y})$, and concluded that the predictive p-value (Equation 2.25) is uniform if and only if a) the density $g_n(\boldsymbol{\theta}|\mathbf{y})$ satisfies a Bernstein-von Mises theorem, b) the asymptotic mean of $t(\mathbf{Y})$ under correctly specified model, denoted $\nu(\boldsymbol{\theta})$, is constant in $\boldsymbol{\theta}$, and c) $\partial\nu(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ gives the asymptotic covariance of $\sqrt{n}[t(\mathbf{Y}) - \nu(\boldsymbol{\theta}_0)]$ and the sample score function $\mathbf{S}_n$. For checking the fit of the GRM, a) follows from Theorem 1 and 2. When $\nu(\boldsymbol{\theta})$ is non-constant but continuous in $\boldsymbol{\theta}$, a simple workaround to fulfill b) is to use the centered statistic $\hat{t}(\mathbf{Y}) = t(\mathbf{Y}) - \nu(\tilde{\boldsymbol{\theta}}(\mathbf{Y}))$ in which $\mathbf{Y}$ is generated by parameter values $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}(\mathbf{Y})$ is some asymptotically normal and consistent estimator of $\boldsymbol{\theta}$ computed from $\mathbf{Y}$. As for practical choices of point estimators, Robins et al. (2000) suggests the ML estimator or a one-step Newton-Raphson approximation starting from the point estimates of the observed data. For the GRM, one could alternatively use the computationally less demanding weighted least square estimators (e.g., Muthén, 1978; Gunsjö, 1994; Maydeu-Olivares, 2006). Finally, c) is guaranteed for our choices of test statistics, which is derived in Appendix E.

2.4.2 The partial predictive approach

This approach is based on Bayarri and Berger's (2000) partial posterior predictive p-value, which removes the dependency caused by a double-use of the observed data by replacing $g_n(\boldsymbol{\theta}|\mathbf{y})$ in Equation 2.25 by a conditional version:

$$g_n(\boldsymbol{\theta}|t, \mathbf{y}) \propto \frac{g_n(\boldsymbol{\theta}|\mathbf{y})}{f_T(\boldsymbol{\theta}, t)}, \tag{2.26}$$

in which $f_T(\boldsymbol{\theta}, t)$ is the density/likelihood function of the test statistic T evaluated at its observed value t. Intuitively, the use of Equation 2.26 partials out the information of $T = t$ when constructing the predictive distribution, as though the analyses were based upon the corresponding conditional model. Meanwhile, conditional on $T = t$ the resulting partial predictive p-value is subject to a usual predictive interpretation as discussed in the previous section. Robins et al. (2000) established the asymptotic uniformity of the partial predictive p-value, assuming a Bernstein-von Mises theorem holds for the conditional model. In the current work, we leave the rigorous proof of the asymptotic normality of the conditional model as a topic for future investigation, and only discuss approximating the partial predictive p-value using Monte Carlo simulations.

Taking advantage of Equation 2.26, i.e., the relationship between $g_n(\boldsymbol{\theta}|t, \mathbf{y})$ and $g_n(\boldsymbol{\theta}|\mathbf{y})$, we could modify Algorithm 2 to adopt the technique of importance sampling (Bayarri and Berger, 2000, Section 2.3), instead of implementing separately a direct Monte Carlo computation from $g_n(\boldsymbol{\theta}|t, \mathbf{y})$. By the importance sampling identity, the partial predictive p-value, denoted $p^\star(\mathbf{y})$, can be re-written as

$$\begin{aligned} p^\star(\mathbf{y}) &= \int P_{\boldsymbol{\theta}}\{t(\mathbf{Y}) > t(\mathbf{y})\} g_n(\boldsymbol{\theta}|t, \mathbf{y}) d\boldsymbol{\theta} \\ &= \int P_{\boldsymbol{\theta}}\{t(\mathbf{Y}) > t(\mathbf{y})\} \frac{g_n(\boldsymbol{\theta}|t, \mathbf{y})}{g_n(\boldsymbol{\theta}|\mathbf{y})} g_n(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &\propto \int \frac{P_{\boldsymbol{\theta}}\{t(\mathbf{Y}) > t(\mathbf{y})\}}{f_T(\boldsymbol{\theta}, t)} g_n(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \end{aligned} \tag{2.27}$$

Let $w(\boldsymbol{\theta}, t) = 1/f_T(\boldsymbol{\theta}, t)$ be the sampling weight. We can modify accordingly the empirical p-value calculation, i.e., Line 7 of Algorithm 2, i.e., to:

$$\hat{p}^\star(\mathbf{y}) = \frac{\sum_{s=1}^{S} w(\boldsymbol{\theta}^{(s)}, t) \mathbb{I}\{t^{(s)} > t\}}{\sum_{s=1}^{S} w(\boldsymbol{\theta}^{(s)}, t)}. \tag{2.28}$$

This importance sampling scheme is favored for the reason that the original Monte Carlo sample from the fiducial distribution can be re-used for every test statistic T of interest; however, it has two significant drawbacks. First, it requires evaluating the density function $f_T(\boldsymbol{\theta}, t)$, which can be challenging or even numerically impossible for certain choices of T. This is the case for our choice of bivariate fit diagnostics, and we resort to a normal approximation of the density function as a workaround. Second, it is well-known that when the proposal density ($g_n(\boldsymbol{\theta}|\mathbf{y})$) is dissimilar to the target one ($g_n(\boldsymbol{\theta}|t, \mathbf{y})$), a few draws may vastly outweigh the rest (i.e., possess a much larger weight $w(\boldsymbol{\theta}, t)$), which consequently results in a sharp increase in approximation error. The degeneracy of sampling weights can be monitored by the effective sample size (ESS):

$$S_e = \frac{[\sum_{s=1}^{S} w(\boldsymbol{\theta}^{(s)}, t)]^2}{\sum_{s=1}^{S} w(\boldsymbol{\theta}^{(s)}, t)^2}. \tag{2.29}$$

The reasoning behind Equation 2.29 is that the variance of an unweighted average of S_e i.i.d. random variables is identical to the weighted sum of S of them.

2.4.3 Choice of test statistics

We now discuss the choice of test statistics for testing sum-score-profile and bivariate fit for the GRM.

Fit to the sum-score profile Following Sinharay et al. (2006), we consider assessing model fit to the observed sum-score distribution (see also Ferrando and Lorenzo-seva, 2001; Hambleton and Han, 2004; and Haberman and Sinharay, 2013). At each sum-score level l, $l = 0, \ldots, \sum_{j=1}^{m} K_j - m$, the observed proportion of this particular level is used as the test statistic:

$$T_l = t_l(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\mathbf{1}^\top \mathbf{y}_i = l\}. \tag{2.30}$$

When the model is correctly specified, the mean of T_l, denoted $\nu_l(\boldsymbol{\theta})$ is the model-implied proportion at the sum-score level l:

$$\nu_l(\boldsymbol{\theta}) = \sum_{\mathbf{y}_i} \mathbb{I}\{\mathbf{1}^\top \mathbf{y}_i = l\} f(\boldsymbol{\theta}, \mathbf{y}_i). \tag{2.31}$$

Expression 2.31 is directly involved in the centering approach, and also in the calculation of the likelihood of T_l:

$$f_{T_l}(\boldsymbol{\theta}, t) \propto \nu_l(\boldsymbol{\theta})^{nt} [1 - \nu_l(\boldsymbol{\theta})]^{n-nt}, \tag{2.32}$$

the reciprocal of which serves as the sampling weight in the partial predictive approach. In practice, $\nu_l(\boldsymbol{\theta})$ can be efficiently computed using the Lord-Wingersky recursive algorithm

(Lord and Wingersky, 1984; Thissen, Pommerich, Billeaud, and Williams, 1995).

When the sample size is small and/or the number of items is large, examining the entire sum-score profile may not be feasible. In this case, we resort to conveniently constructed score groups (e.g., equally spaced across the entire range), and compute observed proportions in each group. The corresponding mean and likelihood of the test statistics have expressions similar to equations 2.31 and 2.32.

Fit to bivariate margins For a pair of items j and k, the marginal lack of fit of the GRM can be revealed by the bivariate cross-product statistic (e.g., Liu and Maydeu-Olivares, 2014):

$$T_{jk} = t_{jk}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y_{ij}y_{ik}, \tag{2.33}$$

which is grounded in a similar rationale as calculating the Spearman correlation for ranked bivariate data. The mean of T_{jk} under correctly specified model can be written as

$$\nu_{jk} = \sum_{\mathbf{y}_i} y_{ij}y_{ik}f(\boldsymbol{\theta}, \mathbf{y}_i). \tag{2.34}$$

The likelihood function of T_{jk} is not easily computed in practice. A normal approximation to the likelihood is derived in Appendix E using the standard asymptotic normality result for i.i.d. multinomial random variables.

CHAPTER 3: COMPUTATION

In line with the general discussion in Chapter 1, sampling from the fiducial distribution (Equation 2.5) is isomorphic to truncated sampling of the random components $\mathbf{A}^\star$ and $\mathbf{Z}^\star$. A Gibbs sampler is developed in this chapter to produce a Markov chain that, by the general theory of Gibbs sampling, approaches the equilibrium given by the target fiducial distribution. We first introduce the general structure of the algorithm, followed by computational details involved at each stage. Some tuning aspects, i.e., choosing starting values and avoiding heavy-tailedness, are discussed next. The computational time needed for various combinations of sample sizes and test lengths is summarized in the end.

3.1 General structure

Throughout this chapter, consider the data $\mathbf{y}$ fixed. Recall that the generalized fiducial distribution (Equation 2.5) is determined by the distribution of $\mathbf{A}^\star$ and $\mathbf{Z}^\star$ truncated to the set $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset$. Algorithm 3 defines a Gibbs sampler for the truncated sampling of $\mathbf{A}^\star$ and $\mathbf{Z}^\star$. Starting from $\mathbf{A}^\star = \mathbf{a}^{(0)}$, $\mathbf{Z}^\star = \mathbf{z}^{(0)}$, and a large bounding box on the parameter space (see the later discussion of starting the algorithm), the algorithm at each cycle updates sequentially each component of $\mathbf{A}^\star$ and $\mathbf{Z}^\star$ conditional on the current values of all other variates and the key restriction $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset$. The representation of the implied set inverse must be updated after each conditional sampling step, in order to yield the desirable truncation at the next conditional sampling step. As the number of cycles tends to infinity, the generated Markov chain is stable around the joint distribution of $\mathbf{A}^\star, \mathbf{Z}^\star \mid Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)$. At the end of each cycle, an extremal point of the updated set inverse is selected, which is regarded as (approximately) an instance of the generalized fiducial having distribution.

Algorithm 3 A Gibbs sampler

```
 1: Starting values: $\mathbf{A}^\star = \mathbf{a}^{(0)}$ and $\mathbf{Z}^\star = \mathbf{z}^{(0)}$ (Algorithm 7)
 2: for cycles $s = 1, \ldots, S$ do
 3:     for observations $i = 1, \ldots, n$ do
 4:         for items $j = 1, \ldots, m$ do
 5:             Unlink observation $i$ from the interior polytope $Q_j(\mathbf{y}_{(j)}, \mathbf{a}_{(j)}^{(s-1)}, \mathbf{z}^{(s-1)})$
 6:         end for
 7:         for dimensions $d = 1, \ldots, r$ do
 8:             Update $Z_{id}^\star = z_{id}^{(s)}$ (Algorithm 5)
 9:         end for
10:         for items $j = 1, \ldots, m$ do
11:             Update $A_{ij}^\star = a_{ij}^{(s)}$ (Algorithm 4)
12:             Update the $j$th polytope (Algorithm 6)
13:         end for
14:     end for
15:     for items $j = 1, \ldots, m$ do
16:         Select with equal probability a vertex of $Q_j(\mathbf{y}_{(j)}, \mathbf{a}_{(j)}^{(s)}, \mathbf{z}^{(s)})$
17:     end for
18: end for
```

Remark 8. For each i, we need the operation of Line 5 in Algorithm 3 prior to executing any updating steps about this particular observation. When neither half-space given by observation i is interior, no extra computation is needed there. When i constitutes the interior polytope, however, the unlinking step is computationally challenging. Currently, Line 5 is achieved by intersecting the initial bounding box with the half-spaces for all but the ith observations (i.e., repeatedly running Algorithm 6). Fortunately, we only need to run the unlinking once for each combination of i and j.

3.2 Conditional sampling steps

3.2.1 Conditional sampling of $A_{ij}^\star$

Fix observation i and item j. The goal of this step is to obtain an updated $A_{ij}^\star$ such that the implied half-spaces have a non-empty intersection with the interior polytope determined by all but the ith observations evaluated at the current values of the corresponding random components; the latter is readily available from Line 5 of Algorithm 3. Here, we only describe the case when a middle category on the response scale is selected: i.e., $0 < y_{ij} < K_j - 1$. The workaround we implement to reduce the impact of a heavy-tailed fiducial distribution (see the later discussion) amounts to augmenting the actual response scale with two "phantom" extreme categories that have no endorsement in the observed data. This effectively renders every actual response option a middle category, and thus the discussion here suffices.

For notational convenience, we use a superscript 0 to highlight the dependency solely on the current values of random components at this particular sampling step, and a superscript 1 to denote the involvement of the updated one. Let $\mathbf{a}^0_{-i(j)}$ and $\mathbf{z}^0_{-i}$ be the current values logistic and normal variates without the ith observation, $\mathbf{a}^1_{(j)}$ be the logistic variates including the updated ith component, and $\mathbf{y}_{-i(j)} = (y_{i'j})_{i' \neq i}$ be the corresponding item responses to the same item. The updated value a^1_{ij} should yield a non-empty polytope: i.e.,

$$Q_j(\mathbf{y}_{(j)}, \mathbf{a}^1_{(j)}, \mathbf{z}^0) = Q_j(\mathbf{y}_{-i(j)}, \mathbf{a}^0_{-i(j)}, \mathbf{z}^0) \cap Q_{ij}(\mathbf{y}_{ij}, a^1_{ij}, \mathbf{z}^0) \neq \emptyset. \tag{3.1}$$

Denoted V^0_{-ij} the collection of vertices of $Q_j(\mathbf{y}_{-i(j)}, \mathbf{a}^0_{-i(j)}, \mathbf{z}^0)$. Due to convexity, Equation 3.1 is further identical to the existence of at least one element of V^0_{-ij} being consistent with each of the two updated half-spaces: i.e.,

$$a^1_{ij} > \tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}^0_i) \tag{3.2}$$

for some $\boldsymbol{\theta}_j \in V^0_{-ij}$, and

$$a^1_{ij} \leq \tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}^0_i), \tag{3.3}$$

for some $\boldsymbol{\theta}_j \in V^0_{-ij}$ as well. It follows that $A^\star_{ij} = a^1_{ij}$ should be generated from a standard logistic distribution truncated to $[\min_{\boldsymbol{\theta}_j \in V^0_{-ij}} \tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}^0_i), \max_{\boldsymbol{\theta}_j \in V^0_{-ij}} \tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}^0_i)]$. An implementation of the updating step is described in Algorithm 4.

Algorithm 4 Updating $A_{ij}^{\star}$

```
1:  set $m = \infty$ and $M = -\infty$
2:  for $\boldsymbol{\theta}_j \in V_{-ij}^0$ do
3:      compute $m_1 = \tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}_i^0)$
4:      if $m_1 < m$ then
5:          $m = m_1$
6:      end if
7:      compute $m_2 = \tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i^0)$
8:      if $m_2 > M$ then
9:          $M = m_2$
10:     end if
11: end for
12: generate $A_{ij}^{\star} = a_{ij}^1$ from the logistic distribution truncated to $[m, M]$
```

Remark 9. Samples from truncated logistic distributions (Line 12) are obtained by an implementation of a slice sampler (Neal, 2003), which is by itself an MCMC algorithm; five cycles are performed for each call of the sampler, which appears to behave well in a pilot study. We also found that slice sampling outperforms the inverse cumulative distribution function (cdf) approach when the truncation bounds are extreme.

3.2.2 Conditional sampling of $Z_{id}^{\star}$

The conditional sampling of $Z_{id}^{\star}$ is slightly more involved than that of $A_{ij}^{\star}$, because a single $Z_{id}^{\star}$ is linked to multiple polytopes belonging to the items loading on latent dimension d. $Z_{id}^{\star} = z_{id}^1$ should be sampled from a suitably truncated standard normal distribution ensuring for each associated item that the updated interior polytope is not empty.

Fix i and d. Let $\mathbf{z}_{i,-d}^0 = (z_{ie}^0)_{e \neq d}$ be the current values of all but the dth dimension of the normal variates, and $\boldsymbol{\theta}_{j,-d}$ be the item parameters without the dth slope. Also write

$$\tau_{jk}^d(\boldsymbol{\theta}_{j,-d}, \mathbf{z}_{i,-d}^0) = \alpha_{jk} + \sum_{e \neq d} \beta_{je} z_{ie}^0. \tag{3.4}$$

For all items j loading on dimension d, the updated value z_{id}^1 should satisfy

$$\beta_{jd} z_{id}^1 < a_{ij}^0 - \tau_{j,k+1}^d(\boldsymbol{\theta}_{j,-d}, \mathbf{z}_{i,-d}^0) \tag{3.5}$$

for some $\boldsymbol{\theta}_j = (\beta_{jd}\ \boldsymbol{\theta}_{j,-d}{}^\top)^\top \in V^0_{-ij}$, and

$$\beta_{jd} z^1_{id} \geq a^0_{ij} - \tau^d_{jk}(\boldsymbol{\theta}_{j,-d}, \mathbf{z}^0_{i,-d}) \tag{3.6}$$

for some $\boldsymbol{\theta}_j = (\beta_{jd}\ \boldsymbol{\theta}_{j,-d}{}^\top)^\top \in V^0_{-ij}$ as well. Let J_d be the collection of items that are associated with $Z^\star_{id}$; equations 3.5 and 3.6 together yield the desirable truncation:

$$\begin{aligned} Z^\star_{id} \in \bigcap_{j \in J_d} \Bigg[& \left(\bigcup_{\boldsymbol{\theta}_j \in V^0_{-ij}} \{ z^1_{id} : \beta_{jd} z^1_{id} < a^0_{ij} - \tau^d_{j,k+1}(\boldsymbol{\theta}_{j,-d}, \mathbf{z}^0_{i,-d}) \} \right) \\ & \cap \left(\bigcup_{\boldsymbol{\theta}_j \in V^0_{-ij}} \{ z^1_{id} : \beta_{jd} z^1_{id} \geq a^0_{ij} - \tau^d_{jk}(\boldsymbol{\theta}_{j,-d}, \mathbf{z}^0_{i,-d}) \} \right) \Bigg]. \end{aligned} \tag{3.7}$$

Both equations 3.5 and 3.6 define one-sided intervals for z^1_{id}, the direction of which is contingent upon the sign of β_{jd} for each vertex in V^0_{-ij}. As a consequence, Equation 3.7 might be an interval or a union of disjoint intervals. The foregoing updating mechanism is summarized as Algorithm 5.

Algorithm 5 Updating $Z_{id}^{\star}$

```
1: set $T = (-\infty, \infty)$
2: for items $j = 1, \dots, m$ do
3:     if $\beta_{jd}$ is fixed to 0 then
4:         cycle the item loop
5:     else
6:         set $T_j = \emptyset$
7:         for $\boldsymbol{\theta}_j \in V_{-ij}^0$ do
8:             if $\beta_{jd} = 0$ then
9:                 cycle the vertex loop
10:            else
11:                compute $m_1 = [a_{ij}^0 - \tau_{jk}^d(\boldsymbol{\theta}_{j,-d}, \mathbf{z}_{i,-d}^0)]/\beta_{jd}$
12:                compute $m_2 = [a_{ij}^0 - \tau_{j,k+1}^d(\boldsymbol{\theta}_{j,-d}, \mathbf{z}_{i,-d}^0)]/\beta_{jd}$
13:                if $\beta_{jd} > 0$ then
14:                    update $T_j = T_j \cup [m_1, m_2]$
15:                else
16:                    update $T_j = T_j \cup [m_2, m_1]$
17:                end if
18:            end if
19:        end for
20:    end if
21:    update $T = T \cap T_j$
22: end for
23: generate $Z_{id}^{\star} = z_{id}^1$ from the standard normal distribution truncated to $T$
```

Remark 10. Again, the technique of slice sampling is used in Line 23 of Algorithm 5. As mentioned earlier, the truncation T can be either a bounded interval, or a disjoint union of bounded intervals. In the latter case, the sampling is done in three steps: a) computing probabilities of the intervals under a standard normal distribution and normalizing to a total sum of one; b) randomly selecting an interval with probabilities computed in step a); c) slice sampling on the selected interval.

3.3 Updating interior polytopes

Inside the observation loop of Algorithm 3, all interior polytopes need to be renewed after the logistic and normal variates are updated. This is geometrically a polytope-cutting problem: Cutting the old polytope formed by the rest of the observations, i.e., $Q_j(\mathbf{y}_{-i(j)}, \mathbf{a}_{-i(j)}^0, \mathbf{z}^0)$, by the two new half-spaces $\tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}_i^1) < a_{ij}^1$ and $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i^1) \geq a_{ij}^1$; the resulting intersection is certainly non-empty due to the truncation enforced for $A_{ij}^{\star}$'s and $Z_{id}^{\star}$'s.

The updating algorithm requires an effective representation of the $\mathbb{R}^{q_j}$-polytope $Q_j(\mathbf{y}_{-i(j)}, \mathbf{a}^0_{-i(j)}, \mathbf{z}^0)$ for each item j. It is well-known that a convex polytope is uniquely determined by its vertices; for our problem, it suffices to record V^0_{-ij}. With a slight abuse of notation, we now consider V^0_{-ij} as a set of doublets $\mathcal{V} = (\boldsymbol{\theta}_j, I)$, in which I indexes the observations that are used to solve for $\boldsymbol{\theta}_j$. If a half-space, say $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}^1_i) \geq a^1_{ij}$, is known to cut the polytope, vertices in V^0_{-ij} can be partitioned into two groups by whether or not they are consistent with the cutting half-space: Those satisfying $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}^1_i) \geq a^1_{ij}$ are still *feasible*, and the rest become infeasible. In addition to vertices, we also track the edges of the old polytope, denoted E^0_{-ij}. Each edge connects a pair of vertices sharing $q_j - 1$ observations, i.e., $\mathcal{E} = (\mathcal{V}, \mathcal{V}')$, in which $\mathcal{V} = (\boldsymbol{\theta}_j, I)$, $\mathcal{V}' = (\boldsymbol{\theta}'_j, I')$, and $|I \cap I'| = q_j - 1$; in other words, the shared observations determines this particular edge. One advantage of keeping the edges is that new vertices introduced by the cutting half-space can be easily obtained: An edge together with the cutting hyperplane $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}^1_i) = a^1_{ij}$ produce a vertex if and only if the edge connects a feasible-infeasible pair of vertices, provided the resulting linear system is non-singular. In addition, the vertex and edge lists need to be updated; entries that are no longer feasible should be removed, and the new ones produced by the cutting half-space should be appended. A pseudo-code of the polytope-cutting procedure is provided as Algorithm 6; in Line 12 of Algorithm 3, two executions of Algorithm 6 are needed for the left and right half-spaces corresponding to a single observation, respectively.

Algorithm 6 Cutting $Q_j(\mathbf{y}_{-i(j)}, \mathbf{a}^0_{-i(j)}, \mathbf{z}^0)$ by $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}^1_i) \geq a^1_{ij}$

```
1: for $\mathcal{V} = (\boldsymbol{\theta}_j, I) \in V^0_{-ij}$ do                ▷ check feasibility
2:     if $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}^1_i) \geq a^1_{ij}$ then
3:         cycle the vertex loop ($\mathcal{V}$ feasible)
4:     else
5:         remove $\mathcal{V}$ ($\mathcal{V}$ infeasible)
6:     end if
7: end for
8: if all vertices are feasible then
9:     terminate the program
10: end if
11: create empty vertex list $V^1_{ij}$ and edge list $E^1_{ij}$
12: for $\mathcal{E} = (\mathcal{V}, \mathcal{V}') \in E^0_{-ij}$ do               ▷ obtain new vertices
13:     if both $\mathcal{V}$ and $\mathcal{V}'$ are feasible then cycle the edge loop
14:     else if neither $\mathcal{V}$ nor $\mathcal{V}'$ is feasible then remove $\mathcal{E}$
15:     else
16:         set $I'' = (I \cap I') \cup \{i\}$
17:         calculate the new vertex determined by $I''$, denoted $\boldsymbol{\theta}''_j$
18:         append $\mathcal{V}'' = (\boldsymbol{\theta}''_j, I'')$ to $V^1_{ij}$
19:         in $\mathcal{E}$, replace the infeasible vertex by $\mathcal{V}''$
20:     end if
21: end for
22: for $\mathcal{V}$, $\mathcal{V}' \in V^1_{ij}$ do                          ▷ obtain new edges
23:     if $|I \cap I'| = q_j - 1$ then
24:         add $(\mathcal{V}, \mathcal{V}')$ to $E^1_{ij}$
25:     end if
26: end for
27: append $V^1_{ij}$ to $V^0_{-ij}$
28: append $E^1_{ij}$ to $E^0_{-ij}$
```

Remark 11. In terms of data structure, we recommend the use of linked lists (i.e., adjacent units are concatenated via pointers) instead of arrays (i.e., adjacent units are stored in consecutive memory locations) as containers of vertex and edge lists, for the reason that the former eases removal and addition of elements to arbitrary locations in the list, which appears frequently in Algorithm 6.

Remark 12. Recording edges facilitates finding new vertices, i.e., Line 16-19 of Algorithm 6; however, the algorithm may fail whenever the linear system determined by observations $(I \cap I') \cup \{i\}$ (Line 16) is singular. When $K_j > 2$, it could happen occasionally; it corresponds to the case that the new half-space cuts the polytope exactly along the edge. In theory, this loophole can be redressed by treating all vertices satisfying $\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i^1) - a_{ij}^1 = 0$ as infeasible; it follows that the edge being cut along is first removed in Line 14, and then added back in Line 24 with the new observation included in the index set. In the presence of numerical errors, a slacking constant $\varepsilon > 0$ should be used instead of the exact zero in practice; in the current implementation of the algorithm, ε is chosen to be 10^{-10}.

3.4 Starting values

The proposed sampler requires initial values of the logistic and normal variates, i.e., $\mathbf{a}^{(0)}$ and $\mathbf{z}^{(0)}$, which imply a non-empty and bounded polytope $Q_j(\mathbf{y}_{(j)}, \mathbf{a}_{(j)}^{(0)}, \mathbf{z}^{(0)})$ for each j. There is certainly more than one way to achieve this. Our algorithm, described in this section, requires user-input of starting values $\boldsymbol{\theta}^{(0)}$ and factor score estimates $\mathbf{z}^{(0)}$. The logistic variates $\mathbf{a}^{(0)}$ are subsequently generated using Algorithm 4, in which each interior polytope comprises only one vertex $\boldsymbol{\theta}_j^{(0)}$; the non-emptiness and boundedness of polytopes are ensured by the truncated sampling.

The boundedness requirement is unnecessary in theory; for each fixed $\mathbf{y}$ the polytope can be unbounded with positive probability. However, the sampling algorithm, especially the polytope-updating part (Algorithm 6), only applies to bounded cases. As a result, an arbitrarily specified initial bounding box is needed (similar configurations can be found in Cisewski and Hannig, 2012, and Liu and Hannig, 2014). For the GRM, we define the following bounding box for $\boldsymbol{\theta}_j$:

$$\begin{aligned} &\alpha_{j1} \geq -M, \alpha_{j,K-1} \leq M \\ &\alpha_{jk} \geq \alpha_{j,k+1}, \text{ for all } k = 1, \ldots, K-2 \\ &-M \leq \beta_{jd} \leq M, \text{ for all } d = 1, \ldots, r. \end{aligned} \tag{3.8}$$

The parameter bound M is an important tuning parameter of the sampling algorithm; the discussion about how to choose M in practice is deferred to Chapter 4 and 5. Based on the foregoing discussion, we outline the starting value program as Algorithm 7.

Algorithm 7 Starting values

```
for items $j = 1, \dots, m$ do
    set $V^0_{-ij}$ and $E^0_{-ij}$ to represent the initial bounding box
    for observations $i = 1, \dots, n$ do
        generate $A^\star_{ij} = a^{(0)}_{ij}$ from Logistic$(0, 1)$ truncated to $[\tau_{j,k+1}(\boldsymbol{\theta}^{(0)}_j, \mathbf{z}^{(0)}_i), \tau_{jk}(\boldsymbol{\theta}^{(0)}_j, \mathbf{z}^{(0)}_i)]$
        Update the $j$th polytope (Algorithm 6)
    end for
end for
```

Remark 13. In practice, $\boldsymbol{\theta}^{(0)}$ can be provided by computationally economical limited information estimators, such as various weighted least square methods based on polychoric correlations (e.g., Muthén, 1978; Gunsjö, 1994). Alternatively, one could use naive starting values such as ordered constants for intercepts and 1 for slopes. $\mathbf{z}^{(0)}$ can be generated from the conditional distribution of the latent variables given $\mathbf{y}$ and $\boldsymbol{\theta}^{(0)}$ (i.e., the posterior distribution (Equation 2.24) evaluated at $\boldsymbol{\theta}^{(0)}$ for each observed response pattern), or point estimates (e.g., EAP) derived from such distribution. $\mathbf{z}^{(0)}$ can also be generated from a standard normal distribution unconditionally. The naive starting values are indeed nowhere near the true item parameters and factor scores, nor the center of the fiducial distribution, but they work reasonably well in our Monte Carlo experiments. From our experience, the generated Markov chain appears stationary after about a thousand iterations, and the final results are not significantly affected by the choice of initial status.

3.5 Heavy-tailedness and a workaround

The generalized fiducial distribution defined by Equation 2.5 is heavy-tailed. Although Theorem 2 guarantees the boundedness of the polytopes when the sample size tends to infinity, in finite samples, however, unbounded polytopes emerge with positive probability. In fact, if the selection rule $\mathbf{v}(\cdot)$ allows infinity, then the distribution does not even have a finite mean. This peculiar feature of the fiducial distribution produces undesirable behavior of the sampling algorithm in practice: The simulated Markov chain of item parameters may sometimes hit the bounding box, which leaves spikes on the trace plot (see the left panel of

Figure 3.1 for an illustration).

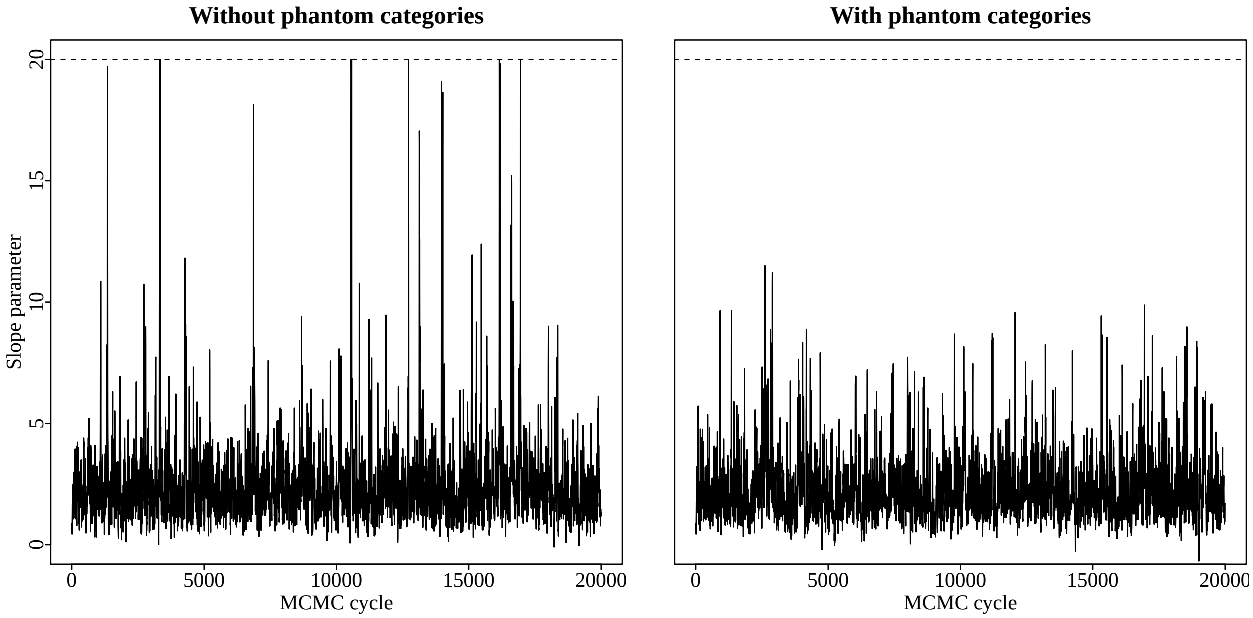

Figure 3.1: Trace plot for a slope parameter before (left) and after (right) implementing the workaround. The data set used for illustration is composed of 50 observations and five 3-category items. The arbitrary bound M is set to 20, which is highlighted by the horizontal dashed lines.

Liu and Hannig (2014) discussed a workaround that enforces a minimal number of lower and upper bounds for the slope parameters implied by the constituting half-spaces, based on the observation that a small number of such bounds is likely to induce unbounded polytopes. In the current work, we propose a more efficient fix, which is naturally adapted to the graded model. The idea is to introduce two "phantom" response categories outside the actual response scale (from 0 to $K_j - 1$), coded as $y_{ij} = -1$ and K_j, in company with two additional intercept parameters α_{j0} and α_{jK}. This extra configuration converts the actual extremal responses 0 and $K_j - 1$ into middle categories; therefore, the set inverse (Equation 2.2) involves two-sided inequalities for all observable responses, and it follows that each observation provides both lower and upper bounds for each slope parameter. No endorsement of the phantom categories can be found in the observed data, so estimates of the extra intercepts are not meaningful. Moreover, freely estimating α_{j0} and α_{jK} increases the dimension of the parameter space, and results in longer computation time. Therefore, we fix $\alpha_{j0} = M$ and $\alpha_{jK} = -M$ in the current implementation, which has proved to perform well in our pilot investigation.

3.6 Computational time

Factors affecting the computational time of the sampling Algorithm 3 are the sample size n, test length m, and number of dimensions $q_j = r_j + K_j - 1$ of each polytope. In Table 3.1, the average CPU time consumed by a single MCMC cycle (averaged across 1000 cycles) is tabulated for different model sizes; the computations were carried out on a laptop computer equipped with a quad-core 2.1GHz Intel Core i7-3687U processor and 8GB of RAM.

Table 3.1 shows that, as expected, the computational time increases approximately linearly as the sample size (n) or test length (m) increases, whereas the time needed grows at a faster than linear rate as the number of response categories (K) or the latent dimensionality (r) increases. n and m affect the dimensionality of the generated variates $\mathbf{A}^\star$ and $\mathbf{Z}^\star$, and thus we expect a linear complexity as the number of sampling and updating computations conducted is in proportion to the dimension of the generated variables. K and r determine the (maximum) dimension of the parameter space for each item, which is associated with the size of the vertex and edge lists involved in the polytope cutting algorithm (Algorithm 6). For a p-dimensional simplex, i.e., a closed polytope with $p + 1$ vertices, it is well-known that the number of edges is $\binom{p+1}{2} = p(p + 1)/2$; therefore, the computational time may grow as a quadratic function of the dimension of the polytope. Indeed, the polytopes generated in the sampling algorithm often have more vertices than simplexes. As a result, this intuitive interpretation is not exact, but it does explain the observed super-linear complexity.

Table 3.1: The average CPU time (in seconds) consumed by a single MCMC iteration under different combinations of sample size n, test length m, latent dimensionality r (exploratory model, minimally constrained), and number of categories K ($K_j = K$ for all j)

n	m	$r = 1$			$r = 2$			$r = 3$		
		$K = 2$	$K = 3$	$K = 5$	$K = 2$	$K = 3$	$K = 5$	$K = 2$	$K = 3$	$K = 5$
100	5	<0.01	<0.01	0.06	<0.01	0.02	0.15	0.02	0.06	0.43
100	10	<0.01	0.02	0.12	0.02	0.05	0.34	0.05	0.14	1.04
100	20	0.02	0.04	0.24	0.04	0.11	0.72	0.11	0.30	2.37
200	5	<0.01	0.02	0.10	0.02	0.04	0.25	0.04	0.09	0.70
200	10	0.01	0.04	0.20	0.04	0.09	0.57	0.09	0.22	1.61
200	20	0.03	0.07	0.41	0.08	0.19	1.19	0.20	0.48	3.40
500	5	0.02	0.04	0.25	0.04	0.10	0.54	0.09	0.21	1.31
500	10	0.04	0.08	0.45	0.09	0.21	1.20	0.21	0.48	2.92
500	20	0.07	0.17	0.91	0.18	0.43	2.48	0.50	1.05	6.50

CHAPTER 4: MONTE CARLO SIMULATIONS

In this chapter, we report a large-scale Monte Carlo simulation study to evaluate the finite sample behavior of the previously discussed implementation of GFI applied to unidimensional and bifactor GRMs. Comparisons of GFI with existing likelihood-based and Bayesian approaches are the foci of our attention; in particular, we are interested in their performance in three types of inference: parameter recovery, test scoring, and checking goodness of fit. For unidimensional GRMs, the simulation design is described in Section 4.1, followed by displays and discussions of the results for the three types of inference (Section 4.2 to 4.4). The simulation design for bifactor models is introduced in Section 4.5, and the parameter recovery results are presented in Section 4.6. A majority of the computations involved in the simulation study were completed on the parallel computing cluster KillDevil located at the University of North Carolina at Chapel Hill.

4.1 Unidimensional models: Simulation design

Graded response data were generated from unidimensional GRMs under a fully factorial design involving three sample size levels, $n = 100$, 200, and 500, and two test length levels $m = 9$ and 18. All items had five ordered response categories ($K_j = 5$ for all j). Results were accumulated across 500 simulated data sets in each condition.

For $m = 9$, the true item parameters were determined by two factors: communality and skewness. In the factor analysis literature, communality for each item measures the proportion of variance explained by the latent variables. Under the logit parameterization of the unidimensional GRM (Equation 1.1), the variance explained is calculated as approximately the squared value of the standardized factor loading parameter (see e.g., Wirth and Edwards, 2007):

$$\lambda_j = \frac{\beta_j/1.7}{\sqrt{1+(\beta_j/1.7)^2}}, \tag{4.1}$$

in which β_j is the unidimensional slope parameter for item j, and 1.7 is the constant to match the standardized logistic and normal cdfs. Values 0.1, 0.5, and 0.9 were selected to represent low, medium, and high levels of communality, respectively. Skewness refers to the degree to which the intercept parameters α_{jk}'s are centered around zero. Here, we also manipulated a standardized version of the intercept, namely, the threshold parameter:

$$\tau_{jk} = \frac{\alpha_{jk}/1.7}{\sqrt{1+(\beta_j/1.7)^2}},\ k = 1,2,3,4. \tag{4.2}$$

Values $(-0.75\ \ -0.55\ \ -0.05\ 0.75)^\top$, $(-0.25\ \ -0.05\ 0.45\ 1.25)^\top$, and $(0.25\ 0.45\ 0.95\ 1.75)^\top$ were used as symmetric, moderately skewed, and skewed threshold conditions, respectively. The nine combinations of three communality and three threshold levels yielded the data generating parameter values for all items in the test. The true parameter values under both standardized loading-threshold and slope-intercept parameterizations are tabulated in Table 4.1. For $m = 18$, the first half of the items had the same parameter values as listed in Table 4.1; the second half had the same factor loading parameters, and threshold parameters with the same absolute values but with reversed signs and ordering. We remark that the parameter values considered here are more extreme than those used in many simulation studies (e.g., Forero and Maydeu-Olivares, 2009). Highly skewed or highly discriminating items are by no means rare in practice, especially in health-related surveys; an example would be an item about suicidal attempts in a scale measuring depressive symptoms.

A Fortran program that implements the proposed Gibbs sampler was used to obtain Monte Carlo samples from the fiducial distribution of item parameters. We used $(1.5\ 0.5\ -0.5\ -1.5)^\top$ as starting values for intercepts, and 1 for slopes. The starting values for the normal variates $\mathbf{z}^0$ were generated from the standard normal distribution unconditionally, and those for the logistic variates $\mathbf{a}^0$ were generated by Algorithm 7 described in the previous

Table 4.1: Data-generating parameter values for the unidimensional GRM ($m = 9$)

Item	Communality	Skewness	Loading λ_j	Thresholds			
				τ_{j1}	τ_{j2}	τ_{j3}	τ_{j4}
1	low	symmetric	0.32	-0.75	-0.55	-0.05	0.75
2	low	moderate	0.32	-0.25	-0.05	0.45	1.25
3	low	skewed	0.32	0.25	0.45	0.95	1.75
4	medium	symmetric	0.71	-0.75	-0.55	-0.05	0.75
5	medium	moderate	0.71	-0.25	-0.05	0.45	1.25
6	medium	skewed	0.71	0.25	0.45	0.95	1.75
7	high	symmetric	0.95	-0.75	-0.55	-0.05	0.75
8	high	moderate	0.95	-0.25	-0.05	0.45	1.25
9	high	skewed	0.95	0.25	0.45	0.95	1.75

Item	Slope β_j	Intercepts				Difficulties			
		α_{j1}	α_{j2}	α_{j2}	α_{j4}	δ_{j1}	δ_{j2}	δ_{j2}	δ_{j4}
1	0.57	1.34	0.99	0.09	-1.34	-2.37	-1.74	-0.16	2.37
2	0.57	0.45	0.09	-0.81	-2.24	-0.79	-0.16	1.42	3.95
3	0.57	-0.45	-0.81	-1.70	-3.14	0.79	1.42	3.00	5.53
4	1.70	1.80	1.32	0.12	-1.80	-1.06	-0.78	-0.07	1.06
5	1.70	0.60	0.12	-1.08	-3.01	-0.35	-0.07	0.64	1.77
6	1.70	-0.60	-1.08	-2.28	-4.21	0.35	0.64	1.34	2.47
7	5.10	4.03	2.96	0.27	-4.03	-0.79	-0.58	-0.05	0.79
8	5.10	1.34	0.27	-2.42	-6.72	-0.26	-0.05	0.47	1.32
9	5.10	-1.34	-2.42	-5.11	-9.41	0.26	0.47	1.00	1.84

chapter. Item parameters were restricted to the bounding box defined by Equation 3.8 with $M = 20$. Small M relative to the magnitude of the true item parameters may reduce the coverage of the resulting CIs; meanwhile, large M may increase the length of the CIs and thus reduce the efficiency. The value $M = 20$ was selected after a pilot study, to strike a balance between those two concerns.

In each replication of the simulation, the sampler was run for 60000 cycles. We visually examined in a pilot study the resulting trace plots, and concluded that 60000 cycles are sufficient for the generated Markov chain to attain stationarity. In addition, we burned in the first 10000 to remove the influence of starting status, and used a thinning interval of 10 to reduce the auto-correlation of the generated Markov chain, as well as the use of computer storage space. As a result, inference was based on 5000 draws in each replication.

ML estimates were computed via the Bock-Aitkin EM algorithm using M*plus* 7.0 (Muthén and Muthén, 2012). The integral in the likelihood function was approximated using 49 equally spaced rectangular quadrature points on interval $[-5, 5]$. We adopted the default convergence criteria and the maximum number of iterations provided by the software, and the same starting values for item intercepts and slopes as in fiducial estimation. In small sample conditions ($n = 100$ and 200), response data with unfilled item response categories emerged occasionally, in which case the ML estimate for the associated intercept parameter does not exist. In order to make the results comparable across all methods, we simulated 500 data sets with no missing response category for all items.

An objective Bayesian method was also considered for comparison. Motivated by Gelman's (2006) recommendation on less informative priors for variance components in the context of linear mixed effects modeling, we specified a standard Cauchy prior for each slope parameter β_j. A slightly more involved prior configuration was needed for intercept parameters due to the order constraints. The four-dimensional prior distribution for $(\alpha_{j1} \ \cdots \ \alpha_{j4})^\top$ was given by the joint distribution of all order statistics of four i.i.d. Cauchy random variables; a similar order-statistic approach was recommended by Curtis (2010). Again, a Cauchy prior was used to reflect our lack of *a priori* knowledge of those intercept parameters in a Bayesian sense[1]. In a pilot study, we also considered a wide uniform prior Uniform$(-20, 20)$ and a diffuse Gaussian prior $\mathcal{N}(0, 100)$; however, they both had substantially lower coverages for large item slopes and extreme intercept parameters compared to the Cauchy, and thus were dropped from the final simulation study. There might be other reasonable choices of prior distributions; the topic of finding the best objective Bayesian approach deserves a careful treatment on its own and should be addressed by future research. JAGS (Plummer, 2013) and its R interface package `rjags` (Plummer, 2013b) were used for Bayesian estimation. The

---

[1] In the extant literature, the prior specification for intercept parameters have been treated in various ways: For example, Edwards (2010) specified a unidimensional prior for the first intercept, and then put non-negative priors on the consecutive intercept differences. A comprehensive comparison of different prior setup is beyond the scope of the current work.

JAGS code for fitting a unidimensional GRM can be found in Curtis (2010, Section 5). A single Markov chain of 60000 cycles was generated with a burn-in period of 10000 and a thinning interval of 10, which paralleled the sampler setup for fiducial estimation.

4.2 Unidimensional models: Parameter recovery

In this section, we describe the comparative behavior of fiducial, likelihood-based, and Bayesian interval estimators in recovering the data generating values of item parameters. Five types of parameters are of interest: namely, item slope β_j, intercept α_{jk}, loading λ_j given by Equation 4.1), threshold τ_{jk} given by equation 4.2) and difficulty δ_{jk}:

$$\delta_{jk} = -\alpha_{jk}/\beta_j. \tag{4.3}$$

The threshold and loading parameters are on a normalized scale and pertain to the notion of explained variance (communality); they serve as the preferred metric in the literature of item factor analysis. The kth difficulty parameters, $k = 1, \ldots, K_j - 1$, is the latent variable value at which response categories $\{0, \ldots, k-1\}$ and $\{k, \ldots, K_j - 1\}$ are equally likely to be endorsed. In the scenario of assigning partial credit in an educational test, δ_{jk} gauges the difficulty to obtain an item score higher than or equal to k. The true values of these parameters can be found in Table 4.1.

Two key evaluation criteria for interval estimators are coverage and length. Ideally, we prefer intervals having coverage probabilities greater than or equal to the nominal level (95% in the current study), and shorter in length. In practice, however, a trade-off between coverage and length is typically observed. We always prioritize coverage over length when comparing the performance of different intervals.

For fiducial and Bayesian methods, we constructed equi-tailed percentile CIs from the fiducial or posterior distribution of model parameters: For a 95% nominal level, the lower and upper confidence bounds for a particular item parameter are set to the 2.5 and 97.5 empirical percentiles of a random sample from the corresponding marginal fiducial or posterior distribution. As discussed earlier, fiducial or posterior distributions for transformed parameters (i.e., threshold, loading, and difficulty parameters) were approximated by transforming the Monte Carlo samples drawn from the original fiducial (posterior) distribution under the slope-intercept parameterization.

For ML estimation, Wald-type CIs for item slopes and intercepts were obtained from two types of standard errors resulting from two commonly used sample estimates of the Fisher information matrix: i.e., the cross-product form (in M*plus*, set `estimator = MLF`) and the Hessian form (set `estimator = ML`). For transformed parameters, the Delta method is used, which is the default method of M*plus*. Other likelihood-based interval estimators such as the profile-likelihood method, as well as resampling-based heuristic methods such as bootstrapping, are not considered here.

Simulation results are summarized in Figure 4.1 to 4.6.

We observe that the fiducial percentile CIs always exhibit on-target coverage for all five types of parameters of interest in all six sample size and test length conditions. Moreover, they are at least as short as other interval estimators in most scenarios; for the difficulty parameters of the low-communality items (items 1 to 3), the fiducial CIs are less efficient than the cross-product-form Wald intervals when $n = 100$ and $m = 9$, and slightly less efficient than the Bayesian intervals when $n = 100$ and $m = 18$. Consequently, we conclude that GFI is the most reliable approach in recovering item parameters among the four candidates being considered in the current work.

The Hessian-form Wald CI, having been regarded as the gold-standard interval estimator associated with the ML estimation, is the most comparable alternative to the fiducial CI. However, it is liberal (i.e., having significantly lower coverage than the nominal level) when applied to the loading parameters of the high-communality items (item 7 to 9) in small samples ($n = 100$ and 200). In those cases, the true parameter value (0.95) is very close to the boundary of the parameter space (1), and the quadratic approximation to the log-likelihood fails. Similar reasoning applies to the under-coverage for the difficulty parameters of the

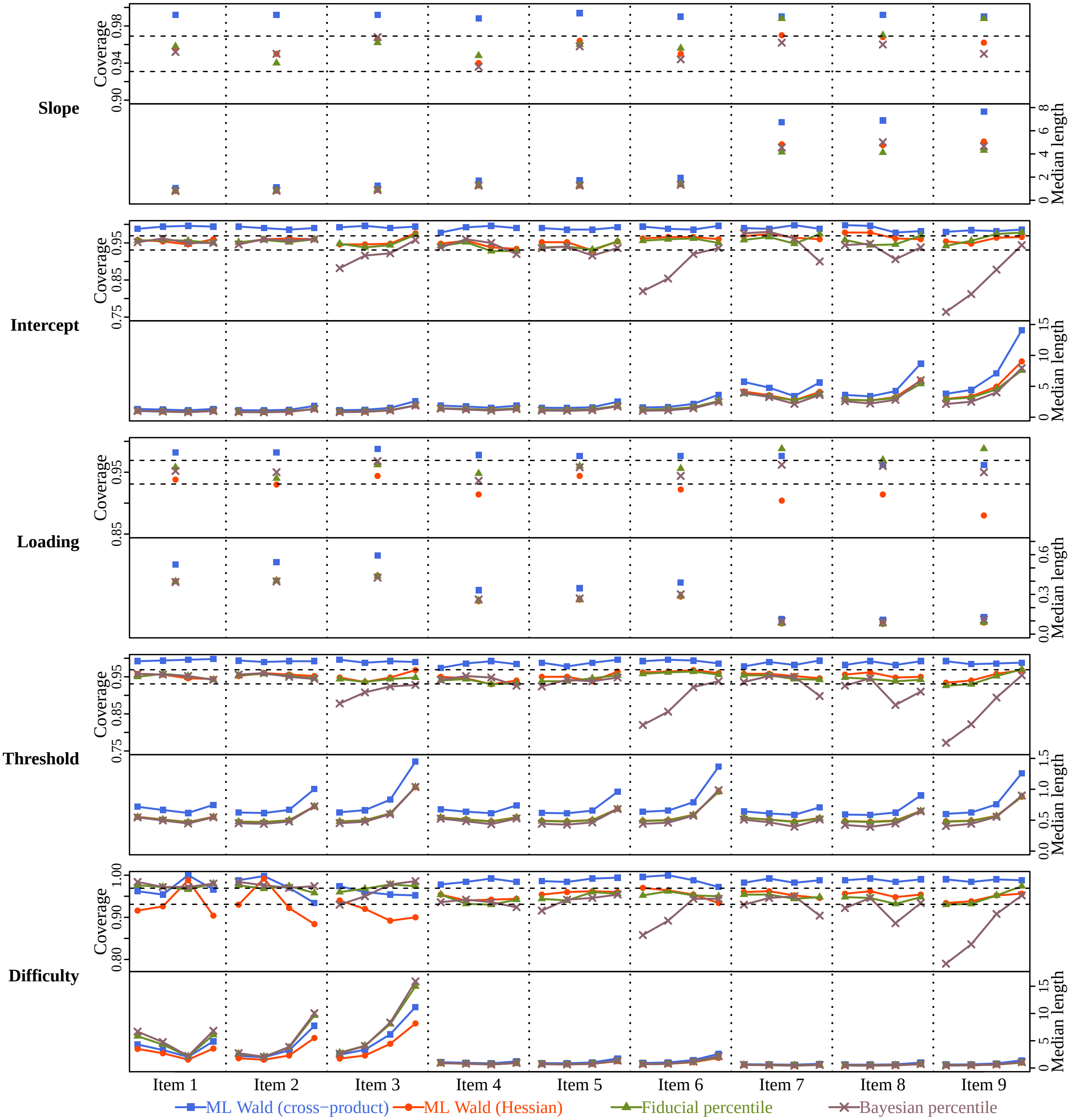

Figure 4.1: Empirical coverages and median lengths of the four types of interval estimators (shown in different colors). Here, the sample size $n = 100$ and the number of items $m = 9$. Each row corresponds to one type of parameter, in which coverage is plotted in the upper panel and median length in the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

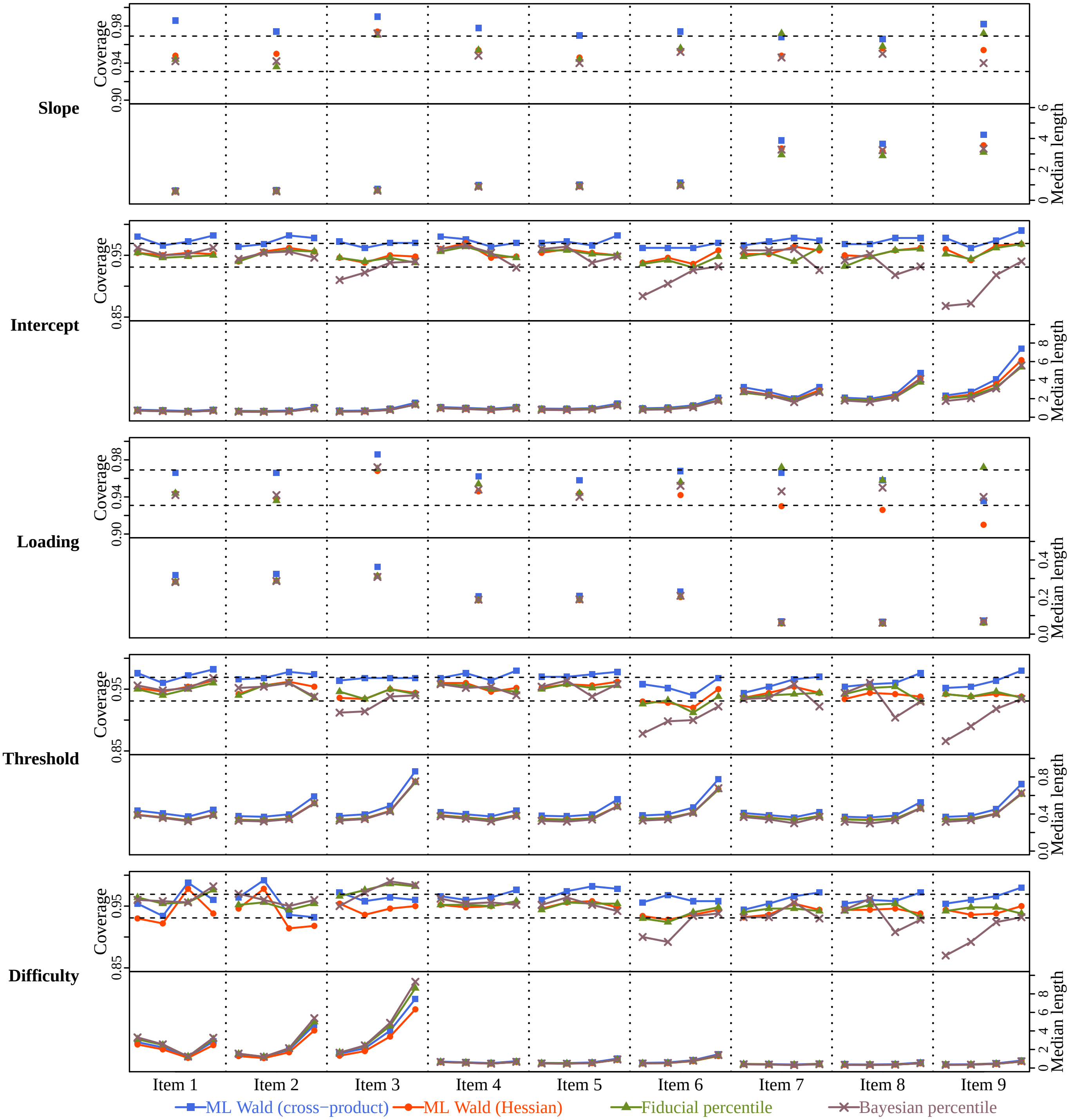

Figure 4.2: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 200$ and the number of items $m = 9$. Each row corresponds to one type of parameter, in which coverage is plotted in the upper panel and median length in the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

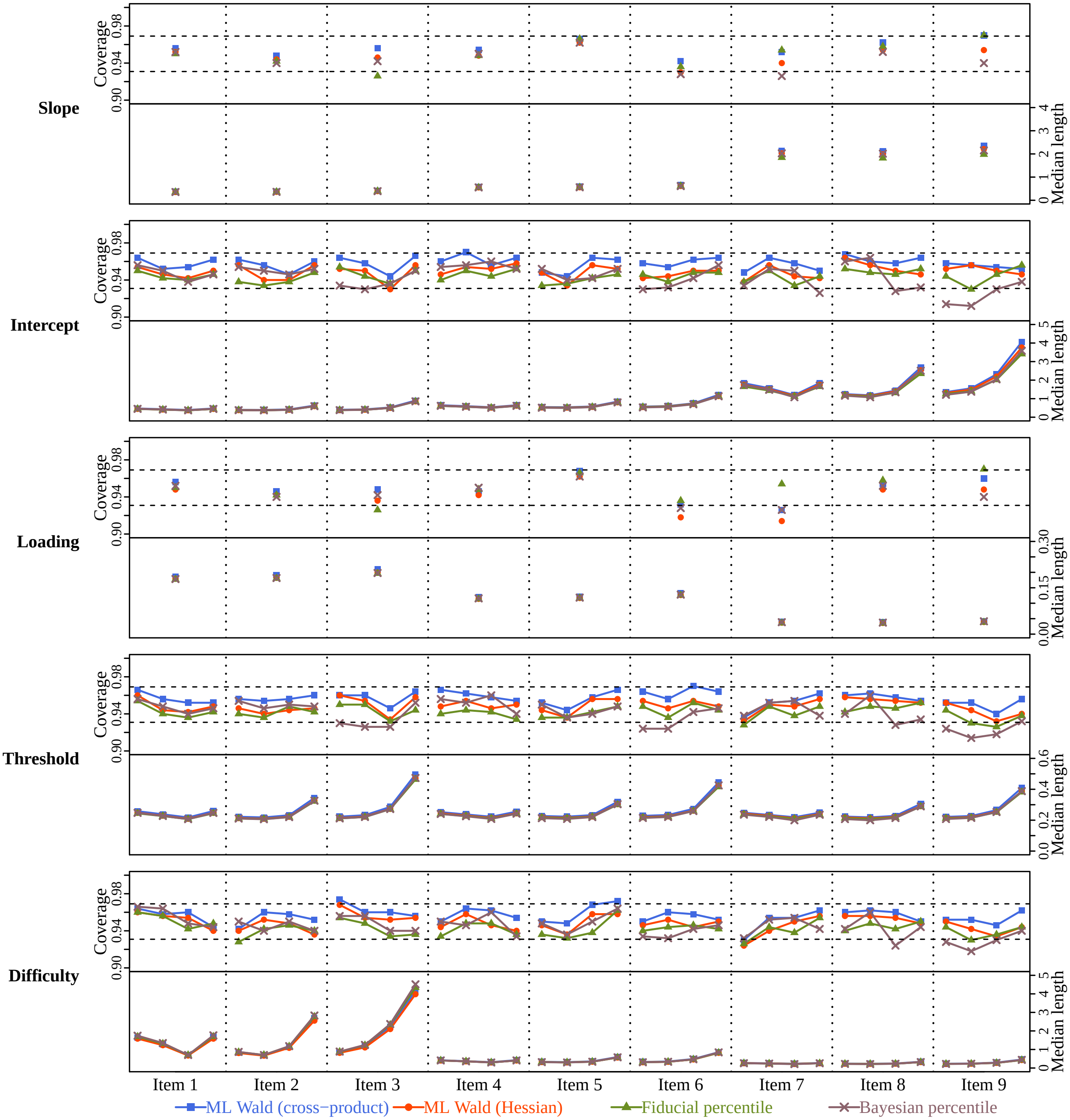

Figure 4.3: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 500$ and the number of items $m = 9$. Each row corresponds to one type of parameter, in which coverage is plotted in the upper panel and median length in the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

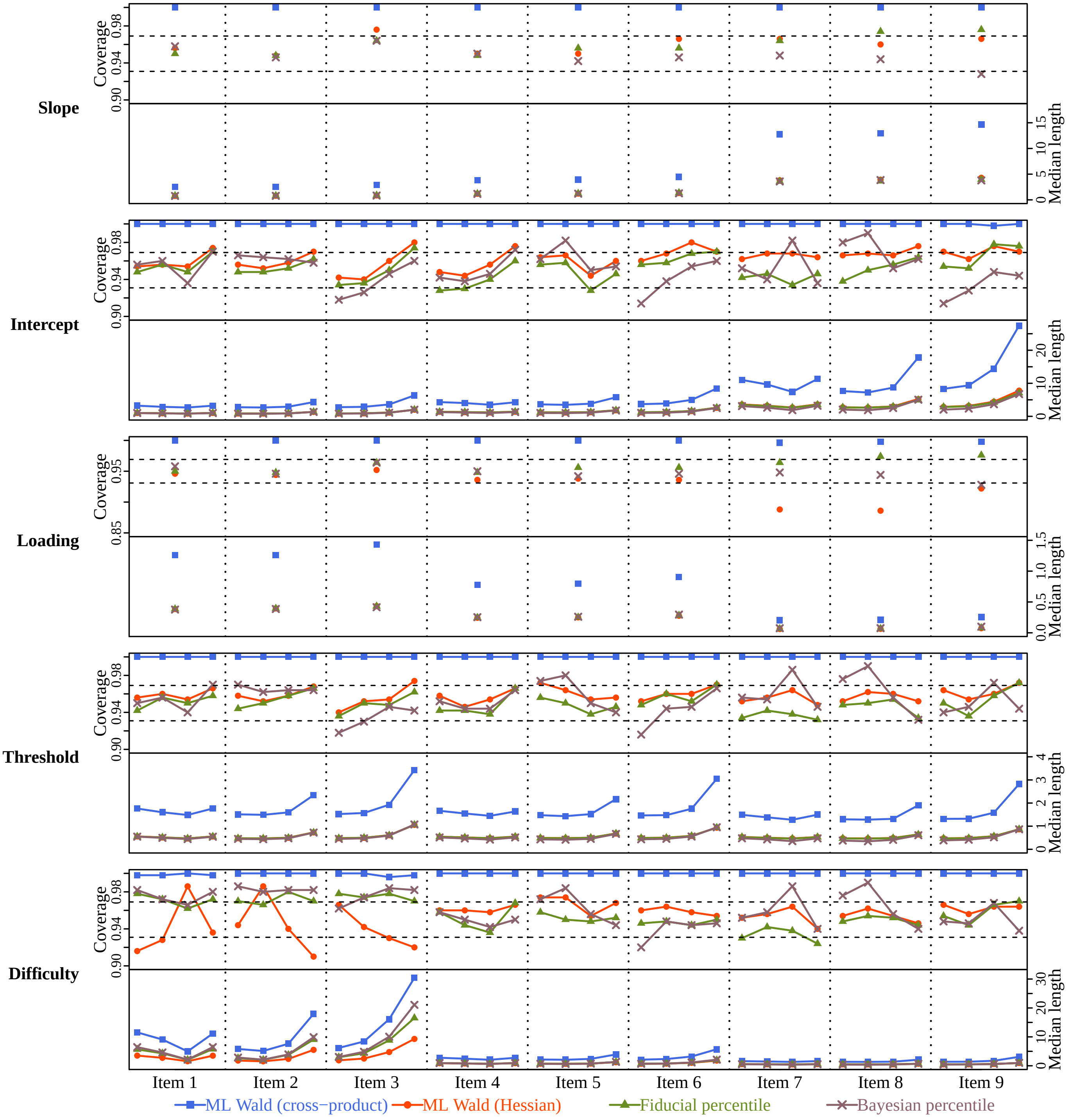

Figure 4.4: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 100$ and the number of items $m = 18$. Each row corresponds to one type of parameter, in which coverage is plotted in the upper panel and median length in the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

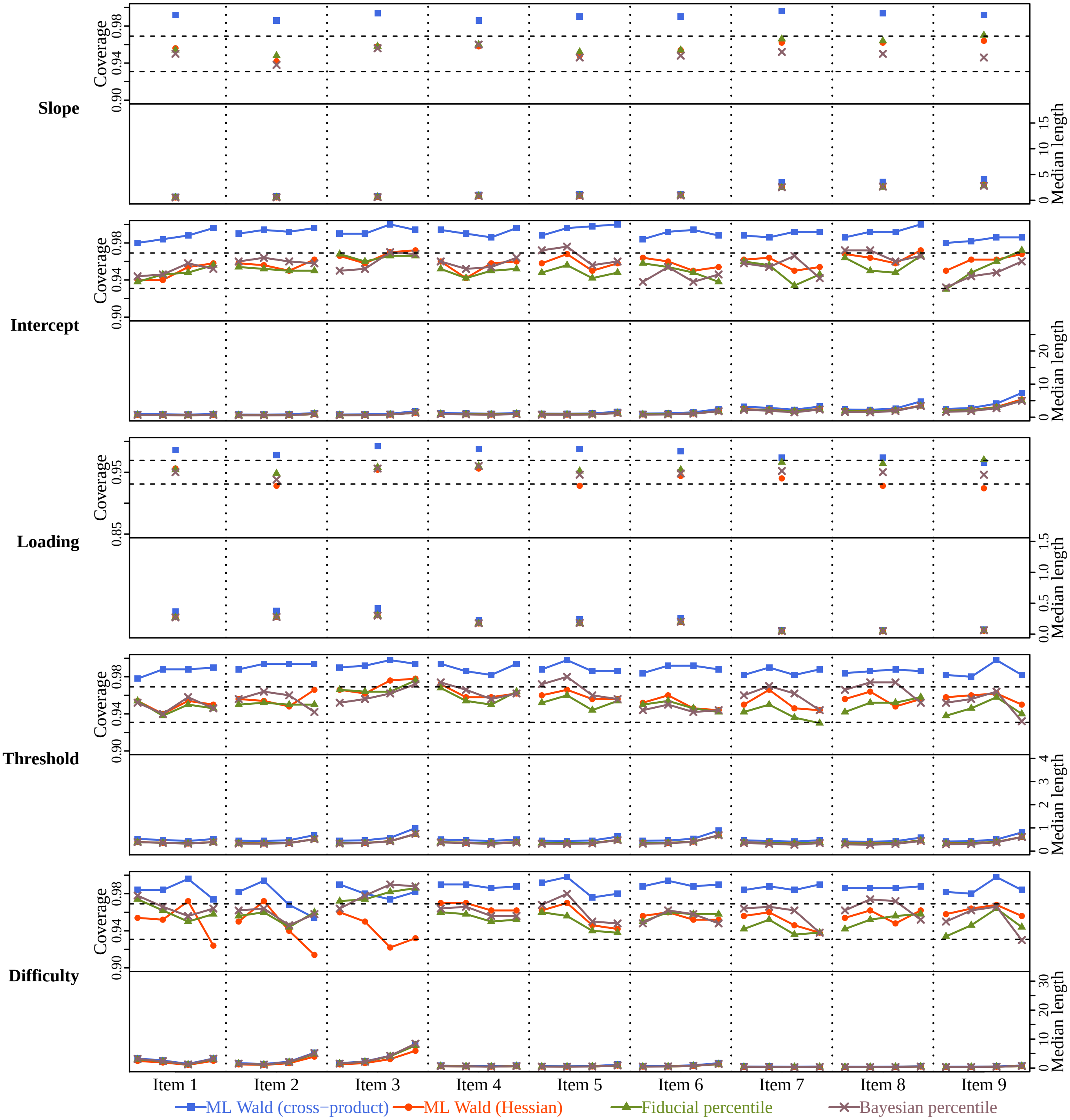

Figure 4.5: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 200$ and the number of items $m = 18$. Each row corresponds to one type of parameter, in which coverage is plotted in the upper panel and median length in the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

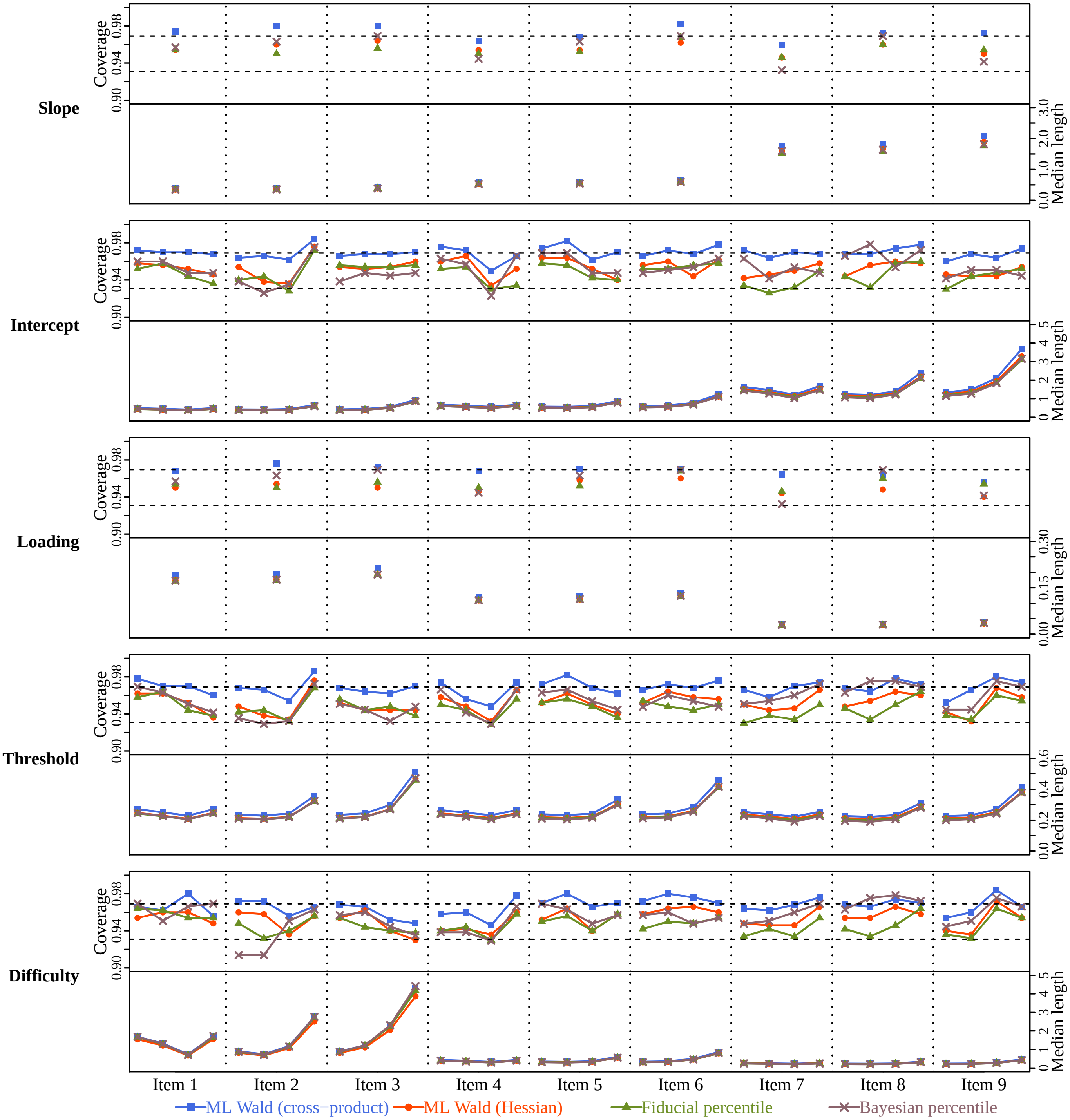

Figure 4.6: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 500$ and the number of items $m = 18$. Each row corresponds to one type of parameter, in which coverage is plotted in the upper panel and median length in the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

low-communality items (item 1 to 3). When a small slope co-exists with a large intercept, the resulting difficulty parameters tend to be large; the largest difficulty parameter involved here is greater than 5, which is almost indistinguishable from infinity, i.e., the "bound" of the parameter space, given that item difficulty is defined on the scale of a standard normal latent variable. Furthermore, the Hessian-form Wald CI also yields slightly longer intervals than the fiducial approach for extreme slope and intercept values when the sample size is small.

The cross-product form Wald CIs, on the other hand, are too conservative (i.e., they have significantly higher coverage than the nominal level) in small sample conditions, resulting in substantially wider intervals than other candidate methods. The pattern is the most salient when $n = 100$ and $m = 18$ (Figure 4.4), in which case the coverages are almost always 1 and the intervals are on average three times as wide as its competitors. Note that the unidimensional GRM is almost unidentified in this extreme condition: The number of parameters is $18 \times 5 = 90$, which is likely to cause numerical difficulty in inverting the cross-product information matrix; that may underlie the excessively conservative results.

The specific type of Bayesian CIs considered in the simulation study can be very liberal for extreme intercept, threshold, and difficulty parameters when the test is short ($m = 9$) and the sample size is small ($n = 100$): The coverage can be as low as 0.75 under the nominal level 0.95. Although they improve as the sample size increases, coverage can still be less than the nominal level when $n = 500$. The problem is significantly alleviated in longer tests ($m = 18$). For other parameterizations of interest, however, the Bayesian method behaves similarly to the fiducial and Hessian-form Wald methods. The observed inferior performance in short tests and small samples might be traced to the particular prior distribution we used; future research is encouraged to explore alternative prior configurations for improvement.

4.3 Unidimensional models: Response pattern scoring

Next, we compare the finite-sample behaviors of various types of asymptotically correct prediction intervals (PIs) for latent variable scores.

Recall that in a unidimensional GRM the latent variable $Z_i \sim \mathcal{N}(0, 1)$ is a random effect; therefore, we often resort to the conditional distribution $Z_i|\mathbf{y}_i$, i.e., Equation 2.24, which has been referred to as the posterior distribution of Z_i in Chapter 2, for making predictions for response pattern scores. Ideally, all item parameters in the scoring model are known; this can be regarded as approximately true when the parameters have been calibrated with high accuracy and precision using a huge sample (e.g., in large-scale educational tests). Under such circumstances, a natural 95% PI for each response pattern score is given by the interval bounded between the 2.5 and 97.5 percentiles of the corresponding true posterior distribution. Since no closed-form expression is available for those percentiles, approximations via numerical quadrature or Monte Carlo sampling are necessary; in the current experiment, the quadrature approach was used. Alternatively, a Wald approach, i.e., treating the posterior distribution as approximately normal, is also applicable: A symmetric PI is given by the posterior mean (i.e., the EAP score) plus or minus the posterior standard deviation multiplied by the normal quantile matching the nominal coverage probability. In the simulation study, we only computed the percentile PIs under the true posterior, for the purpose of setting a gold-standard reference for comparison.

We are interested primarily in the situation in which item parameters need to be simultaneously calibrated from the scoring sample. Constrained by the possibly rare population being sampled from (e.g., people diagnosed with a certain type of psychological disorder), available data for item calibration and scoring are often less than 500 in many psychological applications of the GRM. In those cases, the plug-in method has been widely used: That is, we first obtain some point estimates (usually the ML estimates) of the item parameters, and subsequently evaluate the posterior distribution (Equation 2.24) at those estimated values. We can construct percentile- or Wald-type PIs based on the plug-in posterior in a fashion similar to those obtained from the true posterior. Both types of PIs were computed in the simulation study; nevertheless, we note that the Wald approach is more popular in practice.

As discussed in Chapter 2, the fiducial distribution $Z_i^\star \mid \{Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}$ can be directly used to construct PIs for observed pattern scores. Monte Carlo samples from such fiducial distributions for all individuals in the data set are byproducts of the Gibbs sampler, from which percentile-based PIs can be easily computed. Predictive inference of response pattern scores can likewise be obtained in the Bayesian framework: Monte Carlo samples of Z_i's are readily available from the sampler, drawn from the posterior distribution with all item parameters integrated out under the prior measure, and can be subsequently used for constructing PIs.

Because the true latent variable values were generated from a continuous probability distribution, we are not able to make comparisons of PIs at any particular score level. Therefore, we assign those true values to 10 score groups separated by 9 equally-spaced cutoff values from -2 to 2, and compute empirical coverages and median lengths by groups for the first 50000 simulated observations—they constitutes all 500 simulated data sets when $n = 100$, the first 250 when $n = 200$, and the first 100 when $n = 500$. Results are visualized in Figures 4.7 to 4.12 for different sample size and test length conditions, respectively.

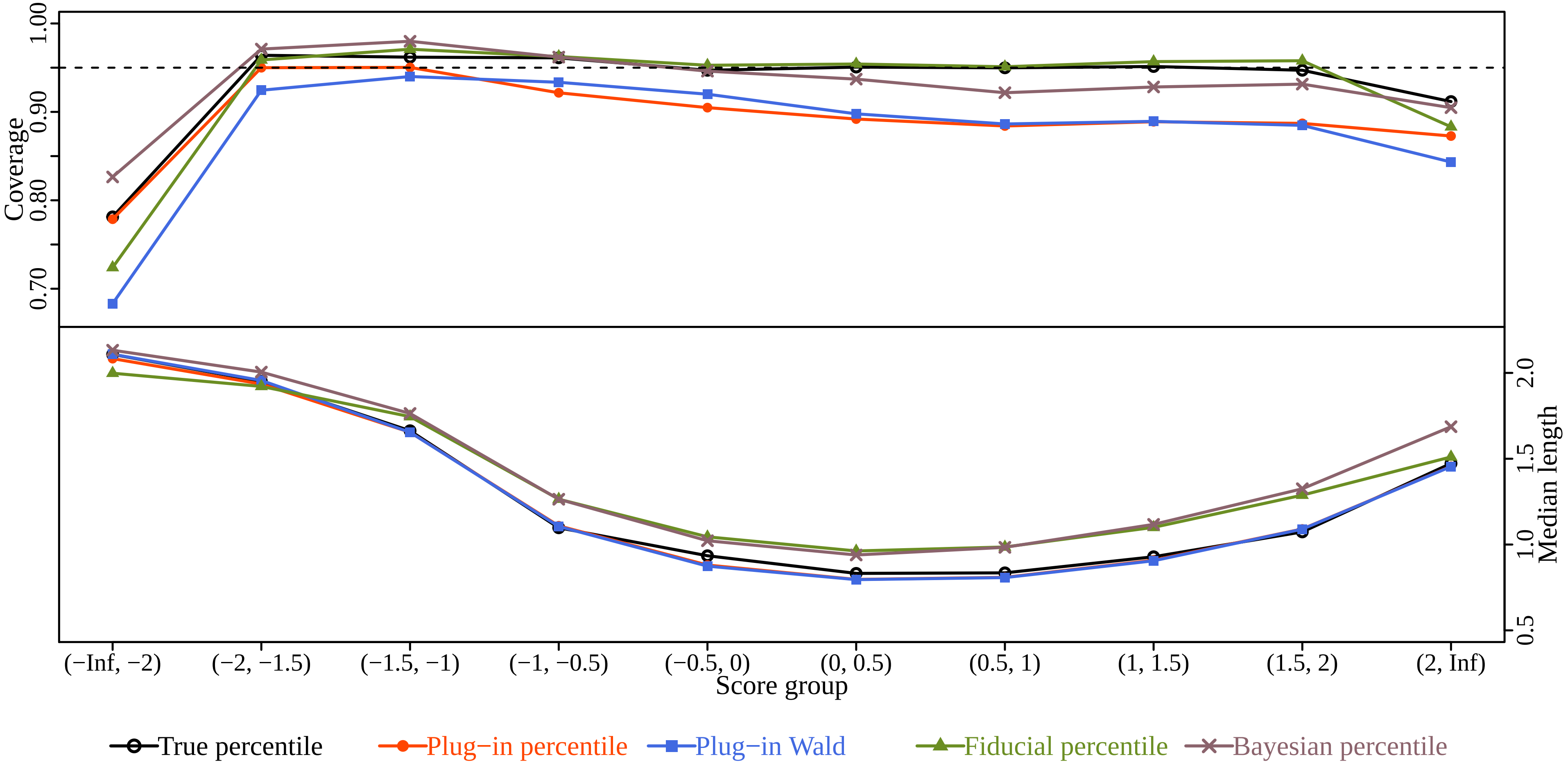

Figure 4.7: Empirical coverage (upper panel) and median length (lower panel) of the five types of prediction intervals (shown in different colors) plotted against the true score groups. Here, the sample size $n = 100$ and the number of items $m = 9$. The horizontal dashed line marks the nominal coverage probability 0.95.

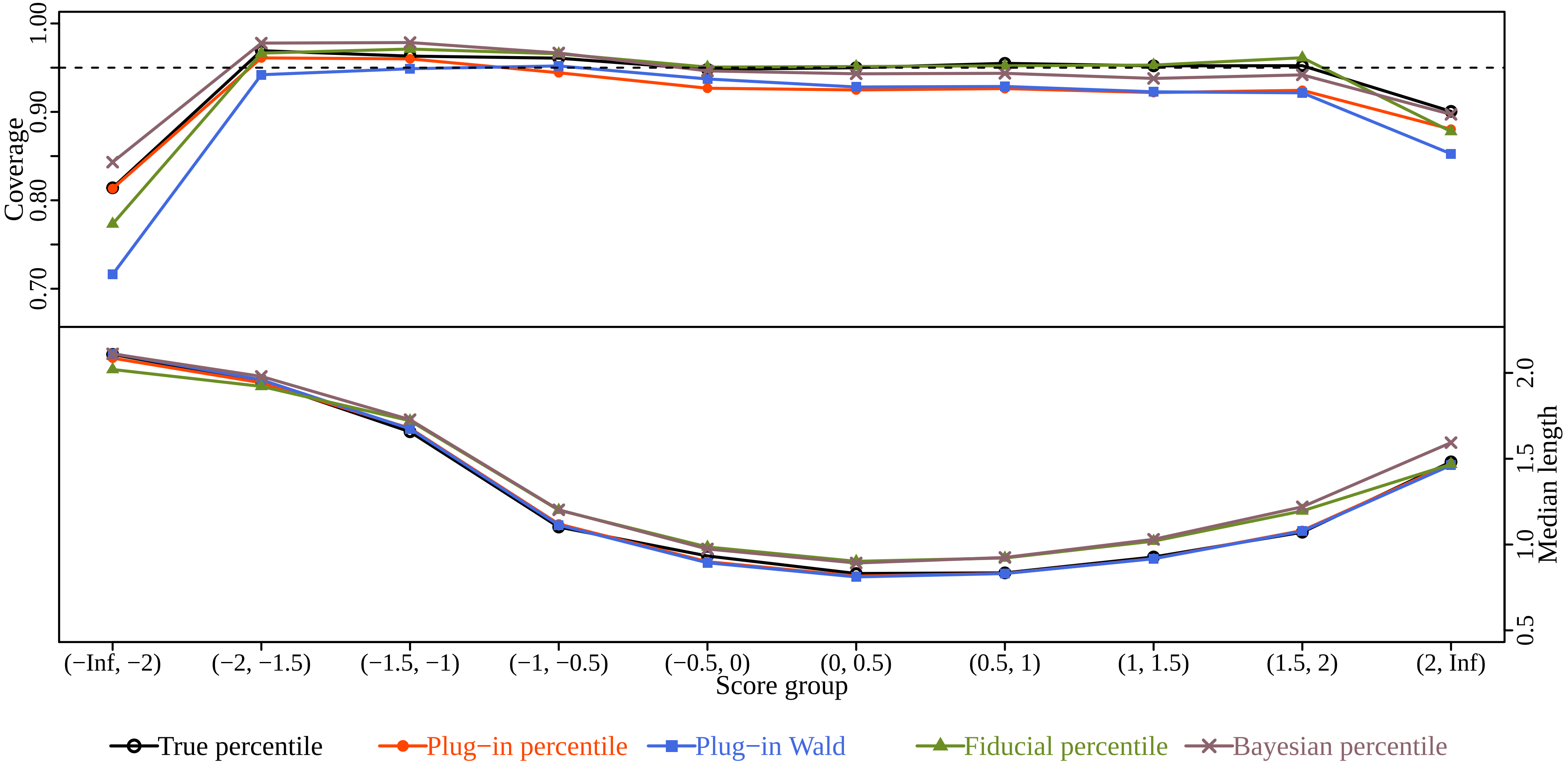

Figure 4.8: Empirical coverage (upper panel) and median length (lower panel) of the five types of prediction intervals (shown in different colors) plotted against the true score groups. Here, the sample size $n = 200$ and the number of items $m = 9$. The horizontal dashed line marks the nominal coverage probability 0.95.

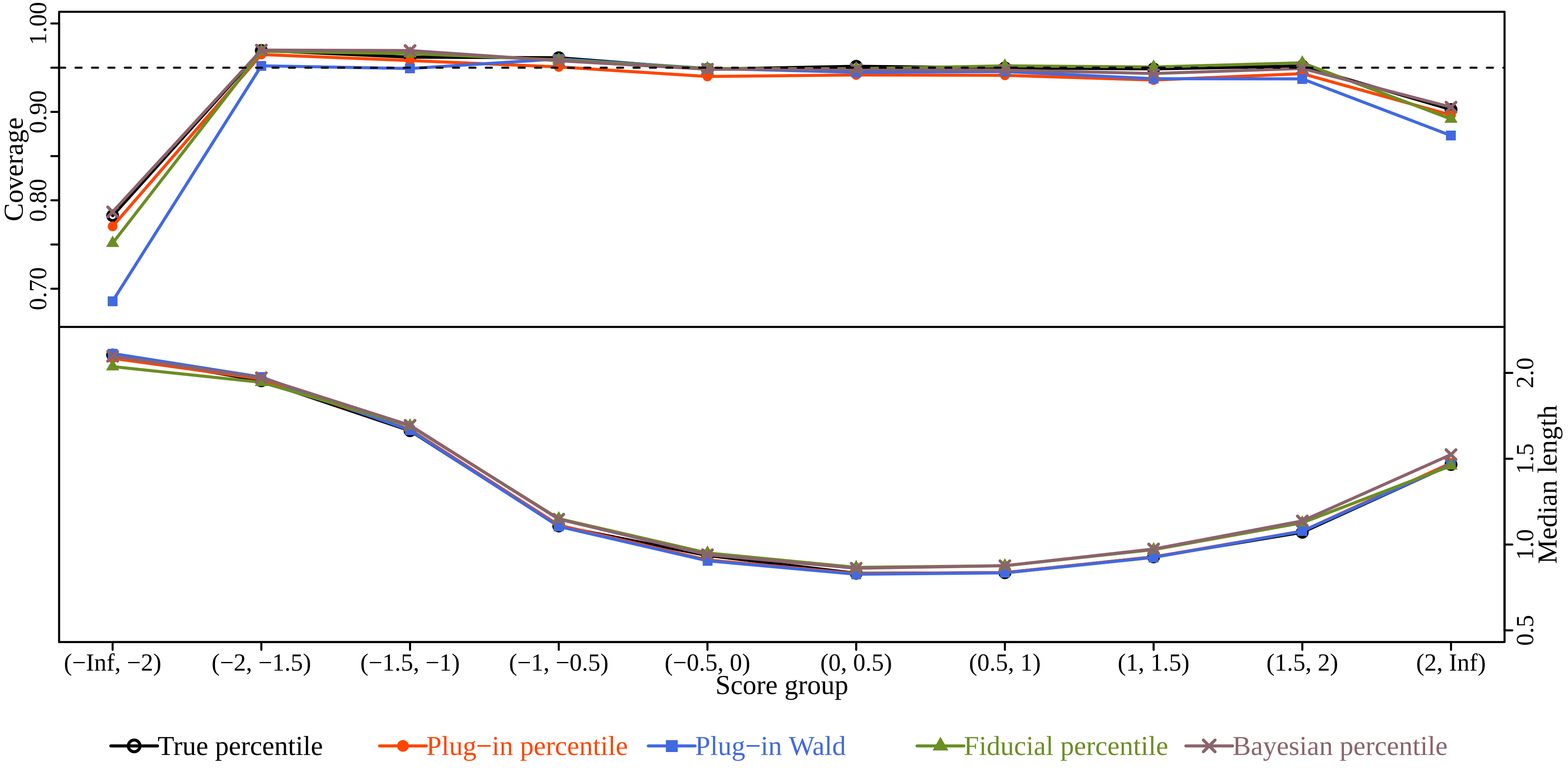

Figure 4.9: Empirical coverage (upper panel) and median length (lower panel) of the five types of prediction intervals (shown in different colors) plotted against the true score groups. Here, the sample size $n = 500$ and the number of items $m = 9$. The horizontal dashed line marks the nominal coverage probability 0.95.

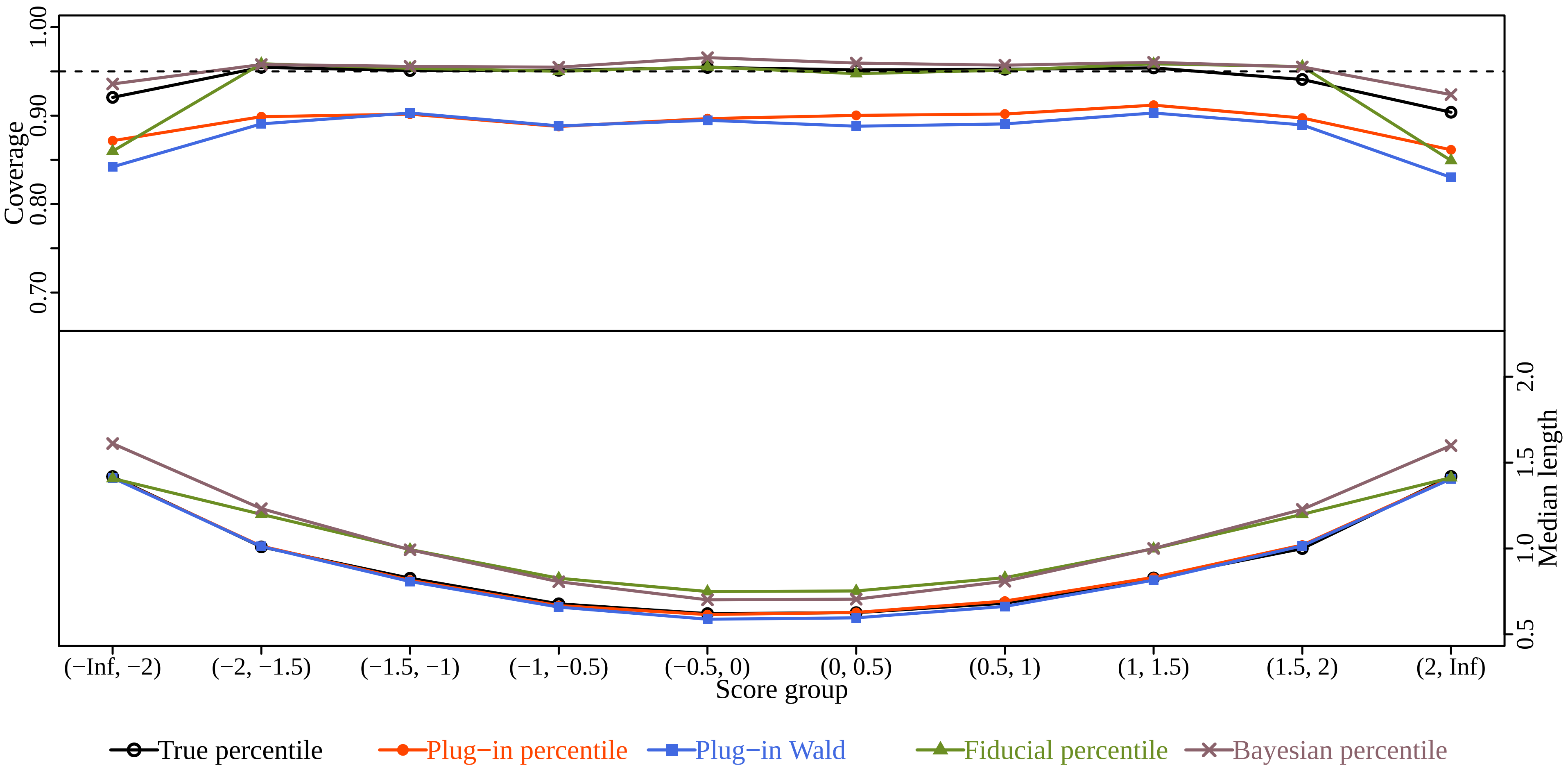

Figure 4.10: Empirical coverage (upper panel) and median length (lower panel) of the five types of prediction intervals (shown in different colors) plotted against the true score groups. Here, the sample size $n = 100$ and the number of items $m = 18$. The horizontal dashed line marks the nominal coverage probability 0.95.

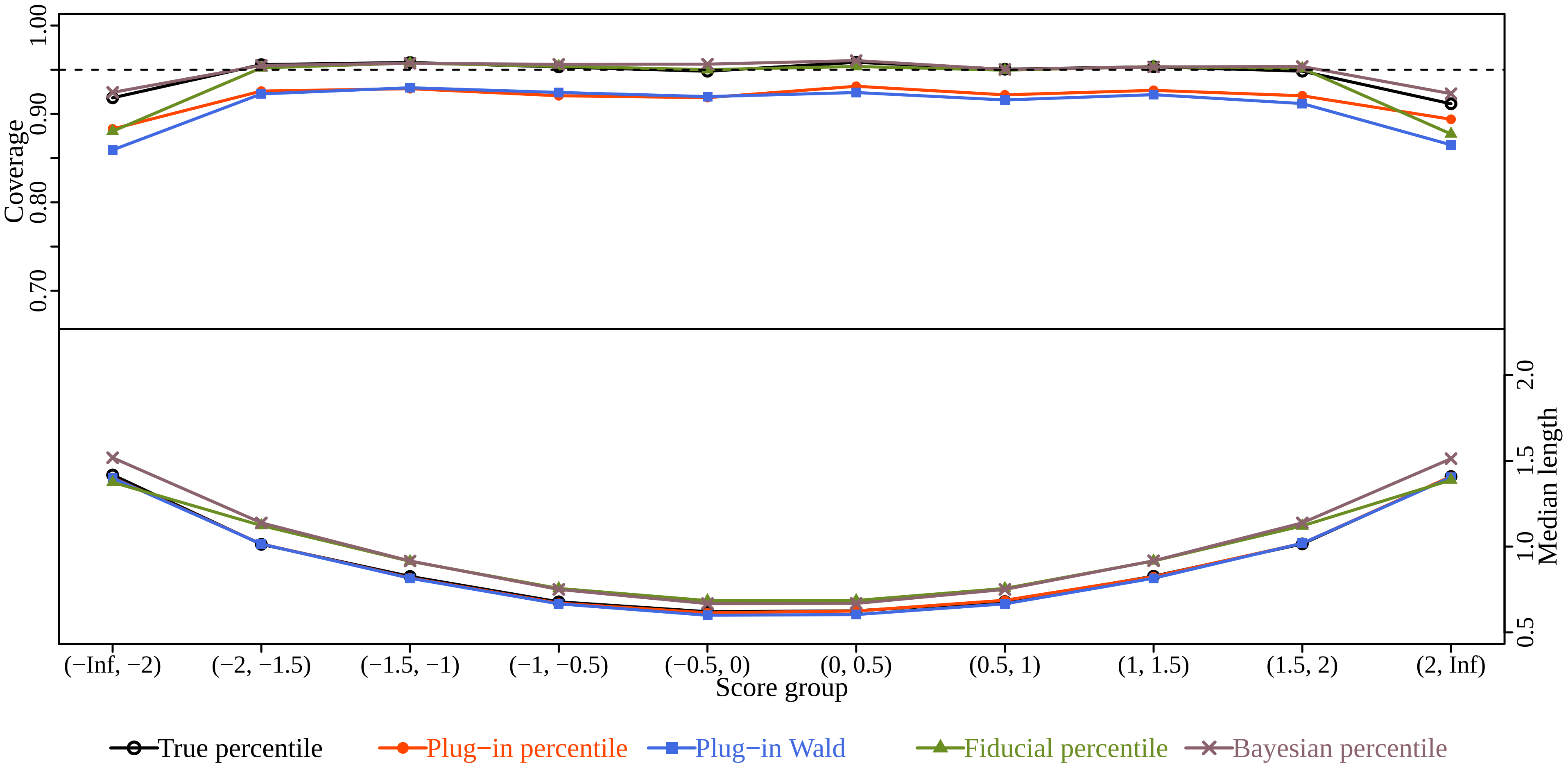

Figure 4.11: Empirical coverage (upper panel) and median length (lower panel) of the five types of prediction intervals (shown in different colors) plotted against the true score groups. Here, the sample size $n = 200$ and the number of items $m = 18$. The horizontal dashed line marks the nominal coverage probability 0.95.

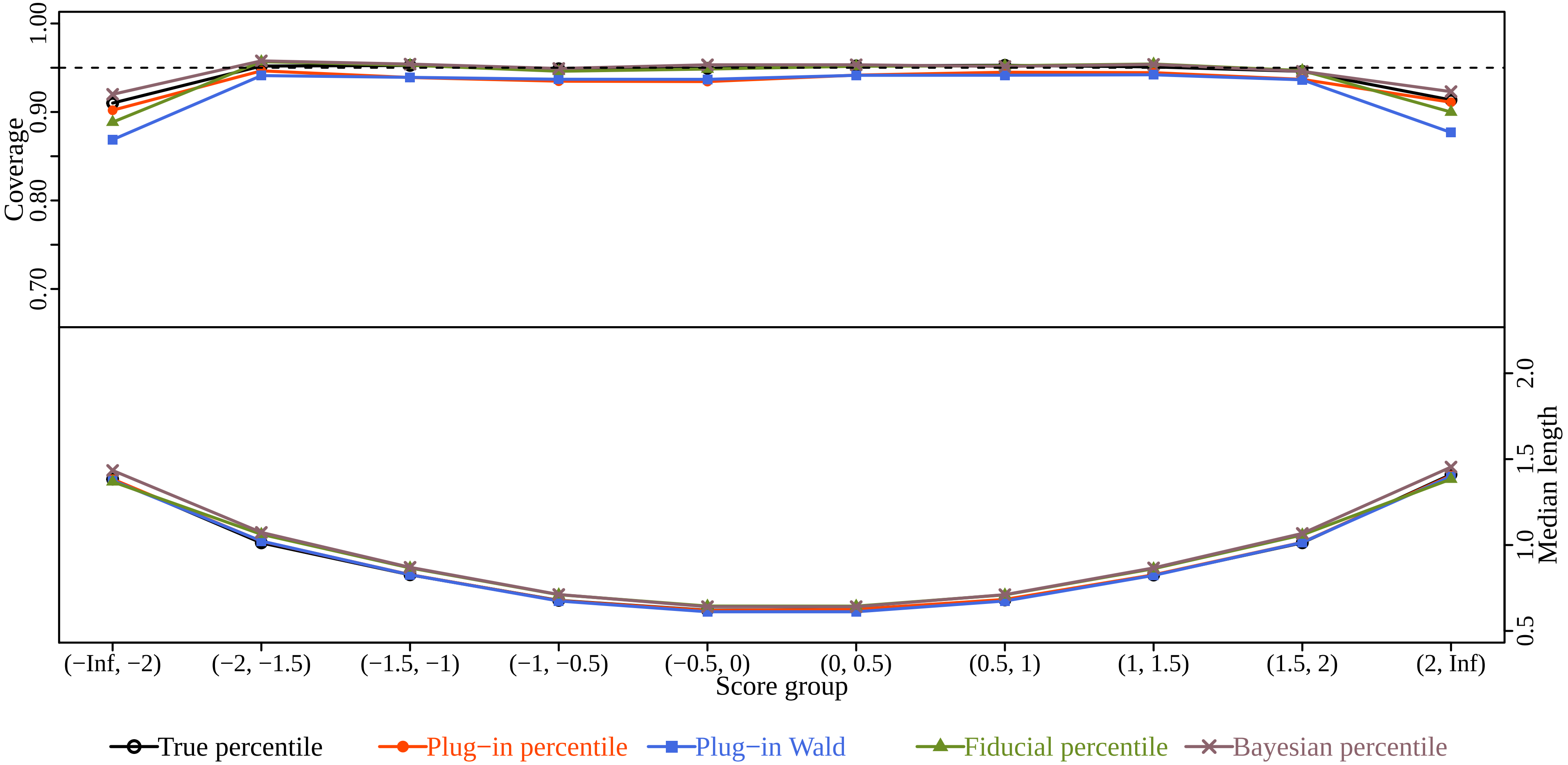

Figure 4.12: Empirical coverage (upper panel) and median length (lower panel) of the five types of prediction intervals (shown in different colors) plotted against the true score groups. Here, the sample size $n = 500$ and the number of items $m = 18$. The horizontal dashed line marks the nominal coverage probability 0.95.

Because the data generating item parameter values are used in the computation, the behavior of the true posterior percentile PIs, i.e., our reference approach, is not contingent upon the sample size of the calibration sample. As expected, their coverages are almost always on target, and their lengths are often the shortest. For extreme score groups, i.e., $(-\infty, -2)$ and $(2, \infty)$, even the true posterior distributions do not support adequate coverage probability, due to the fact that the number of items is limited. In particular, the lowest score groups are poorly recovered in short tests ($m = 9$; Figure 4.7 to 4.9), for the reason that the true thresholds are skewed to the positive side (see Table 4.1) and thus little information about the lower-end of latent variable scores can be garnered from the response patterns.

The plug-in intervals under-cover in small samples (when $n = 100$, the coverage can be lower than 90%), because they fail to take into account the sampling variability of the ML estimates, and thus, intuitively speaking, they are not as wide as what they should be. In extreme score groups, the Wald approach fares worse than the percentile one, which suggests that approximating the posterior distribution by a normal one is not appropriate for those observations. As a consequence, we do not recommend the use of plug-in intervals in small sample applications of the GRM, although that is the most widely adopted practice in substantive research.

In terms of empirical coverage, fiducial and Bayesian percentile PIs are the closest to the true posterior percentile ones, and thus are preferred over the plug-in PIs. Bayesian PIs outperform fiducial ones in extreme score groups; however, when a small sample ($n = 100$) combines with a short test ($m = 9$), Bayesian PIs slightly under-cover for most groups on the positive half of the latent variable scale. Surprisingly, the Bayesian approach is well-behaved in making predictive inference about extreme latent variable scores, even thought the corresponding CIs for extreme intercept/threshold/difficulty parameters (see the previous section) are poor. We also observe that the fiducial and Bayesian PIs are substantially wider compared to the other candidates, which is anticipated as the result of properly accounting for the sampling variability. Recently, Xie, Liu, Chang, and Chen (2014) studied the theoretical properties of predictive distributions, extending the findings of Lawless and Fredette (2005). They established under extra regularity conditions that predictive densities obtained from confidence distributions dominate the plug-in densities in terms of the average Kullback-Leibler discrepancy from the target density. It is conjectured that our findings are subject to an analogous theoretical interpretation.

4.4 Unidimensional models: Goodness of fit testing

In this section, we investigate the Type I error and power performance of the fiducial predictive check (FPC) in identifying score group and bivariate model misfit of the unidimensional GRM. For simplicity, we focus our attention on only one sample size and test length combination, i.e., $n = 200$ and $m = 18$. The previously generated graded response data were re-used to calculate the Type I error results; methods having empirical rejection rates close to the nominal α-level are considered well-calibrated. To evaluate power, two alternative data generating models similar to those studied by Liu and Maydeu-Olivares (2014) were used. The first model is a three-dimensional simple-structure GRM with correlated latent variables (factor inter-correlations = 0.3). Items $1, 4, \ldots, 16$ load on the first factor, items $2, 5, \ldots, 17$ load on the second factor, and items $3, 6, \ldots, 18$ load on the third factor. The factor loadings have the same numerical values as tabulated in Table 4.1. As a result, all the factors have matched loadings. The same threshold parameters were used as well; therefore, the first factor is indicated by items having symmetric thresholds, the second by items having moderately skewed thresholds, and the third by items having skewed thresholds. The second model is a unidimensional GRM characterized by the same item parameters as listed in Table 4.1 but a non-normal latent variable $Z_i \sim \frac{1}{2}\mathcal{N}(-1.5, 1) + \frac{1}{2}\mathcal{N}(1.5, 1)$. In both cases, the unidimensional GRM with a normal latent variable is misspecified, and a good fit checking procedure should be able to reject the fitted model as often as possible.

To reduce the computational burden, we incorporated a further thinning interval of 10 to the generated Markov chain in each replication; the fiducial predictive p-values thereby were calculated based on 500 Monte Carlo samples using Algorithm 2. Both the centering (Section

2.4.1) and partial (Section 2.4.2) approaches were implemented, and the extent to which they are able to ensure asymptotically uniform p-values was empirically evaluated and compared. In the centering method, we evaluate the expectations of the test statistics (under the null model), i.e., Equation 2.31 for score-group fit statistics and Equation 2.34 for pairwise fit statistics, at a mean and variance adjusted diagonally weighted least square estimator that is available in M*plus* (`estimator = WLSMV`). For the partial predictive approach, the efficient sample size (ESS, Equation 2.29), were recorded in each replication; means and standard deviations of ESS values were reported in addition to the empirical rejection rates.

In a test composed of 18 five-category items, the sum score ranges from 0 to $18 \times 4 = 72$. In the current simulation study, the score range was divided into 10 equi-width intervals, and the score-group fit statistics (Equation 2.30), previously defined at individual sum-score levels, were computed for the corresponding score groups instead; those statistics were denoted $T_{(1)}, \ldots, T_{(10)}$. For comparison, the reduced M_2 statistic (Cai and Hansen, 2013), denoted $M_2^\star$, was also computed to probe the score-group fit. $M_2^\star$ statistic is a quadratic form of all residual first and second moments (i.e., mean for each item, and cross-product for each pair). The residual mean for each item j is defined as the discrepancy between the observed and model-implied average scores:

$$\hat{e}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} - \sum_{\mathbf{y}_i} y_{ij} f(\hat{\boldsymbol{\theta}}, \mathbf{y}_i). \tag{4.4}$$

Similarly, the residual cross-product for a pair of items j and k is defined as

$$\hat{e}_{jk} = \frac{1}{n}\sum_{i=1}^{n} y_{ij} y_{ik} - \sum_{\mathbf{y}_i} y_{ij} y_{ik} f(\hat{\boldsymbol{\theta}}, \mathbf{y}_i), \tag{4.5}$$

which is in fact the difference between equations 2.33 and 2.34 evaluated at the ML estimates $\hat{\boldsymbol{\theta}}$. The weight matrix of the quadratic form is a specific generalized inverse of the asymptotic covariance matrix of the residual moments, and thus the resulting statistic follows asymptotically a chi-square distribution (Joe and Maydeu-Olivares, 2010). Cai and Hansen (2013) found that $M_2^\star$ statistic has a better calibrated Type I error performance than its predecessor M_2 (Maydeu-Olivares and Joe, 2006) when the test is long and the items are polytomous.

For bivariate fit assessment, we compared the FPC with two diagnostics recommended by Liu and Maydeu-Olivares (2014): namely, the bivariate residual cross-product z-test, and the bivariate Pearson's X^2-test with a method-of-moment correction. The former test statistic is a standardized version of Equation 4.5 that has an asymptotic standard normal distribution. The latter procedure amounts to computing Pearson's X^2 statistics for bivariate marginal tables (i.e., a 5×5 contingency table for two five-category items). Asymptotically, the statistic follows a mixture of independent chi-square distributions under the null model, which is further approximated by a scaled chi-square distribution having matched first and second moments. In both methods, the Hessian-form estimate of the Fisher information matrix is used. For ease of graphical presentation of the results, bivariate fit statistics were computed only for item pairs in the first half of the test.

To visualize the Type I error behavior of the test statistics of interest, we plot their empirical rejection rates against nominal α-levels, i.e., the empirical cdf of the p-values, ranging from 0 to 0.2. The diagonal line on the plot corresponds to a uniform p-value. Liberal and conservative statistics correspond to the upper- and lower-diagonal regions on the plot, respectively.

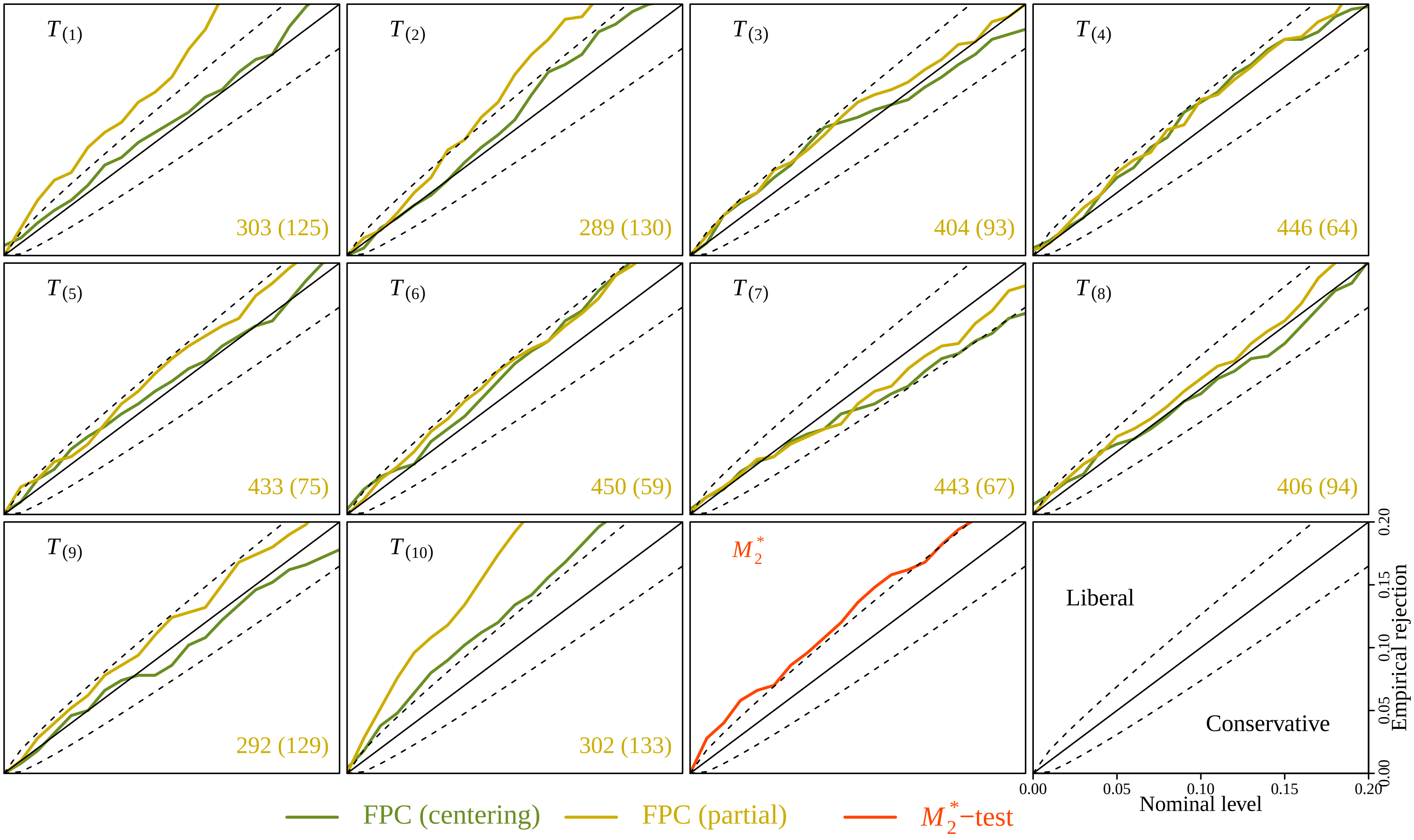

Figure 4.13: Type I error results for score-group goodness of fit tests. In each panel, the x-axis is the nominal α-level, and the y-axis is the empirical rejection rate pooling across 500 replications. The diagonal solid line indicates empirical rejection equal to the nominal level, and the dashed lines give a 95% normal approximation confidence band. Different statistics are displayed in different colors. The numbers shown on the lower-right corner are the means and standard deviations (in parenthesis) of ESS associated with the partial FPC approach.

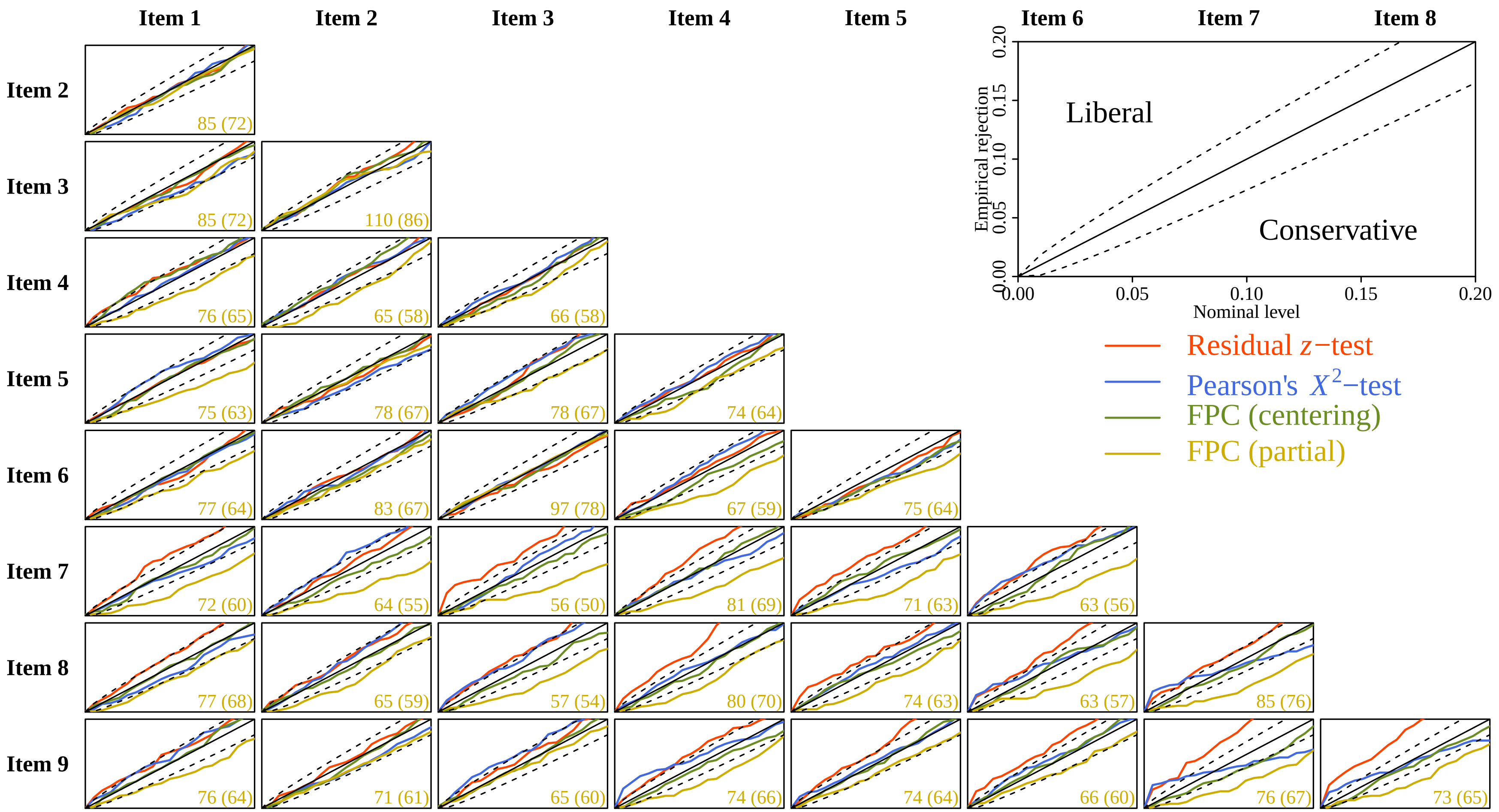

Figure 4.14: Type I error results for bivariate goodness of fit tests. In each panel, the x-axis is the nominal α-level, and the y-axis is the empirical rejection rate pooling across 500 replications. The diagonal solid line indicates empirical rejection equal to the nominal level, and the dashed lines give a 95% normal approximation confidence band. Different statistics are displayed in different colors. The numbers shown on the lower-right corner are the means and standard deviations (in parenthesis) of ESS associated with the partial FPC approach.

Figure 4.13 shows the results for score-group goodness of fit tests. The reference procedure, i.e., the $M_2^\star$ test, is slightly liberal. Cai and Hansen (2013) showed that the limited information M_2 statistic under-performs when first- and second-order marginal contingency tables are sparse. The same rationale applies to the $M_2^\star$ statistic, and that may cause the observed liberality in the current simulation study, in which the data-generating item parameter values are extreme and likely to produce sparse lower-order margins. The centering FPC approach works reasonably well in all but the last score groups, i.e., using statistic $T_{(10)}$, in which it rejects slightly more than the nominal level. The partial FPC procedures yield on-target Type I errors in the middle score groups (from $T_{(3)}$ to $T_{(9)}$; $T_{(2)}$ is acceptable when α is small), but become unacceptably liberal in the extreme groups ($T_{(1)}$ and $T_{(10)}$). The ESS values associated with the sampling weights are on average greater than 300 in most score groups; they tend to vary more in the extreme groups than in the middle ones, which may contribute to the inflated Type I error rates.

Results for bivariate goodness of fit tests are presented in Figure 4.14. The residual product z-test over-rejects for those pairs involving high-communality items (i.e., items 7, 8, and 9). In addition, in very few cases (≤ 10 for each pair) we obtain negative asymptotic variance for the bivariate residual cross-product due to numerical error, which leads to an undefined test statistic; these cases were excluded when computing empirical rejection rates. The p-value obtained from Pearson's X^2-test, although less liberal, has a significant deviation from uniformity: It tends to over-reject for small α-levels (< 0.05), and then suddenly switchs to under-reject as α increases. The unsatisfying behaviors of the two reference methods are likely traceable to the sparseness of the two-way marginal tables resulting from the combination of the small sample size and extreme true parameter values. The centering FPC is less prone to such effects; it yields empirical rejection rates close to the nominal α-level for all 36 pairs of items plotted in Figure 4.14. The partial FPC approach, on the other hand, is conservative for some pairs of items. The ESS values are always below 100, which indicates that those weighted Monte Carlo samples used to calculate the p-values are noticeably degenerate; this may be the leading cause of the conservativeness.

We briefly summarize the Type I error results. We found that traditional goodness of fit testing procedures associated with the ML estimation may not work well in small sample-size conditions. The partial FPC approach works reasonably well in assessing score-group fit when observed proportions in middle score groups are used as a test statistics, but over-rejects when it is applied to proportions in extreme score groups. For bivariate fit assessment, the partial approach suffers from noticeable degeneracy of sampling weights, and is overly conservative. Based on the Type I error results, we recommend the centering FPC method, whose rejection of the correctly specified models is well-controlled by the designated α-level of the test.

Similar graphical tables are reported in the two conditions when the unidimensional GRM is misspecified. Inasmuch as the fitted model is wrong, we anticipate small p-values from fit checking procedures; in other words, they ought to reject substantially more than the nominal level. The more powerful the test is, the more the empirical cdf of the p-value climbs above the line indicating uniformity.

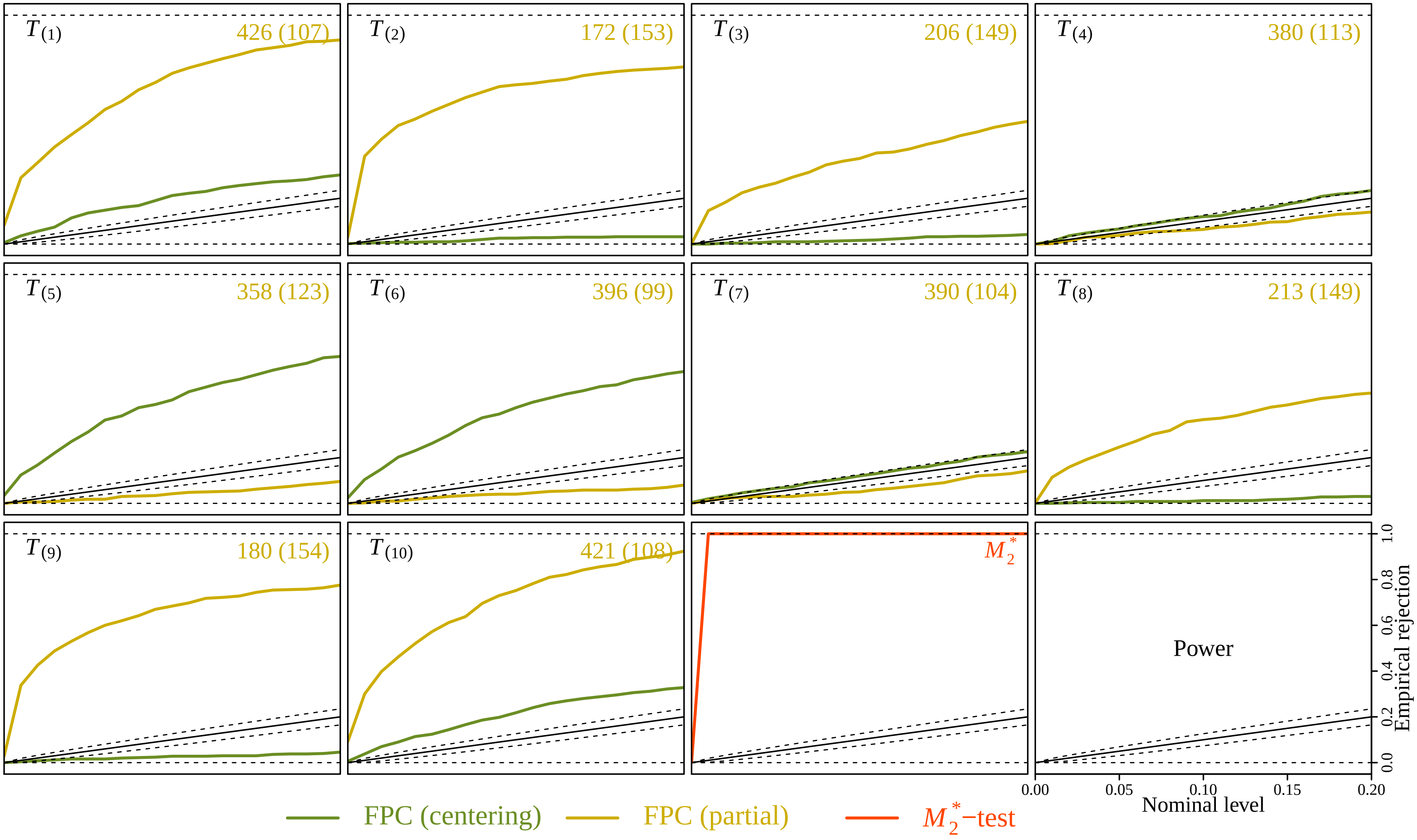

Figure 4.15: Power results for score-group goodness of fit tests: True model = 3-dimensional. In each panel, the x-axis is the nominal α-level, and the y-axis is the empirical rejection rate pooling across 500 replications. The slanted solid line indicates empirical rejection equal to the nominal level, and the slanted dashed lines give a 95% normal approximation confidence band. The horizontal dashed line indicates the maximum power 1. Different statistics are displayed in different colors. The numbers displayed in each panel are the means and standard deviations (in parenthesis) of ESS associated with the partial FPC approach.

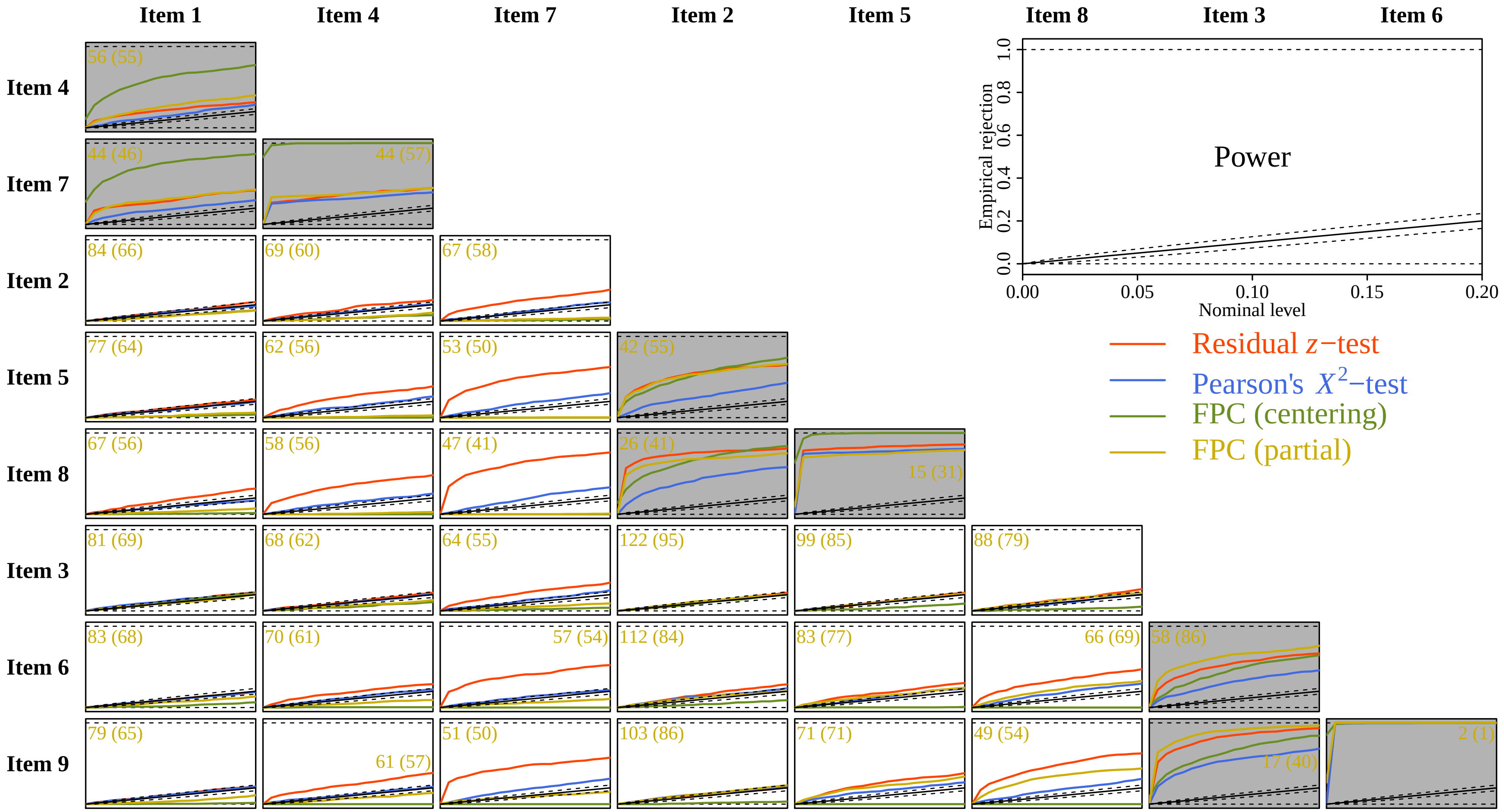

Figure 4.16: Power results for bivariate goodness of fit tests: True model = 3-dimensional. Items are reordered so that those loaded on the same factor are grouped together; panels associated with these within-factor pairs are highlighted with a light-gray background. In each panel, the x-axis is the nominal α-level, and the y-axis is the empirical rejection rate pooling across 500 replications. The slanted solid line indicates empirical rejection equal to the nominal level, and the slanted dashed lines give a 95% normal approximation confidence band. The horizontal dashed line indicates the maximum power 1. Different statistics are displayed in different colors. The numbers displayed in each panel are the means and standard deviations (in parenthesis) of ESS associated with the partial FPC approach.

When the data were generated from a three-factor model, the $M_2^\star$ test has the highest power to reject the fitted unidimensional model (see Figure 4.15). Because $M_2^\star$ incorporates all bivariate residual cross-product moments in its computation, it is sensitive to under- and over-fit to residual "correlations" among item responses, which is, in an intuitive sense, what differentiates a multidimensional factor analytic model from a unidimensional one. For FPC, the centering and partial approaches exhibit disparate patterns: The centering approach has moderate power in the middle groups (i.e., statistics $T_{(5)}$ and $T_{(6)}$) and very little power otherwise; meanwhile, the partial approach has moderate power in more extreme groups (i.e., statistics $T_{(2)}$, $T_{(3)}$, $T_{(8)}$, and $T_{(9)}$) and little power elsewhere. Note that the partial FPC also rejects substantially more than the nominal level for $T_{(1)}$ and $T_{(10)}$; however, the inflated Type I error rates in those cases should be taken into account.

Results for bivariate tests are tabulated in Figure 4.16. All testing procedures under investigation tend to reject more within-factor pairs (highlighted by the gray background) than between-factor ones, and the power increases as the slopes increase. This pattern is similar to that observed in Liu and Maydeu-Olivares (2014). For within-factor pairs, the power of the centering FPC maintains at roughly the same level regardless of the true threshold values involved in the pair, while the power of the other three methods increases as the thresholds become more extreme. As a result, the centering FPC is preferred for symmetric items (items 1, 4, and 7), and the partial FPC and the z-test are preferred for skewed items (items 3, 6, and 9); the X^2-test, however, always has the lowest power for detecting within-factor pairs. For between-factor pairs, the residual z-tests outperform the X^2-test and both FPC approaches. The proposed FPCs are under-powered for detecting between-factor pairs in general, especially the centering approach that is non-responsive to any of those pairs.

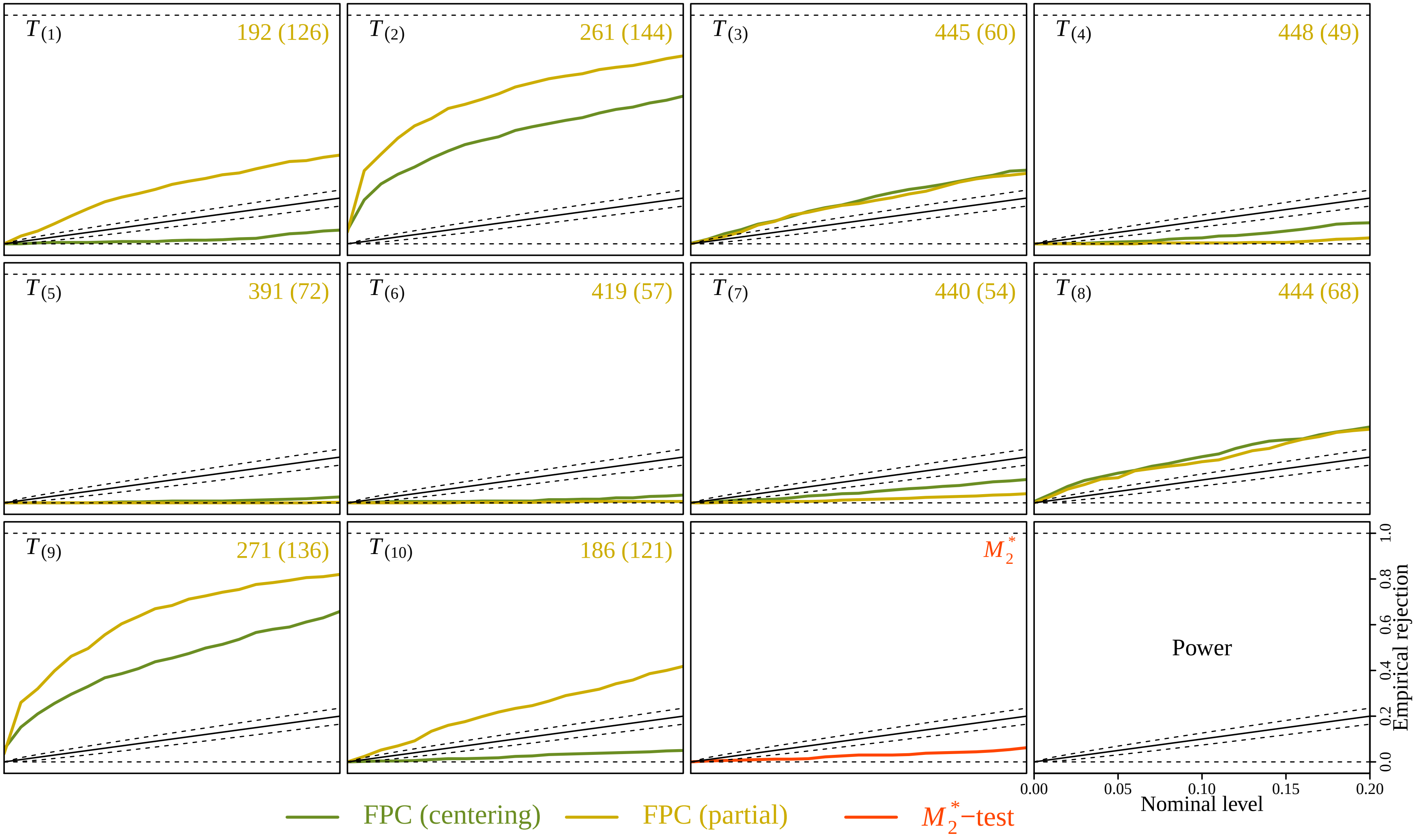

Figure 4.17: Power results for score-group goodness of fit tests: True model = mixture. In each panel, the x-axis is the nominal α-level, and the y-axis is the empirical rejection rate pooling across 500 replications. The slanted solid line indicates empirical rejection equal to the nominal level, and the slanted dashed lines give a 95% normal approximation confidence band. The horizontal dashed line indicates the maximum power 1. Different statistics are displayed in different colors. The numbers displayed in each panel are the means and standard deviations (in parenthesis) of ESS associated with the partial FPC approach.

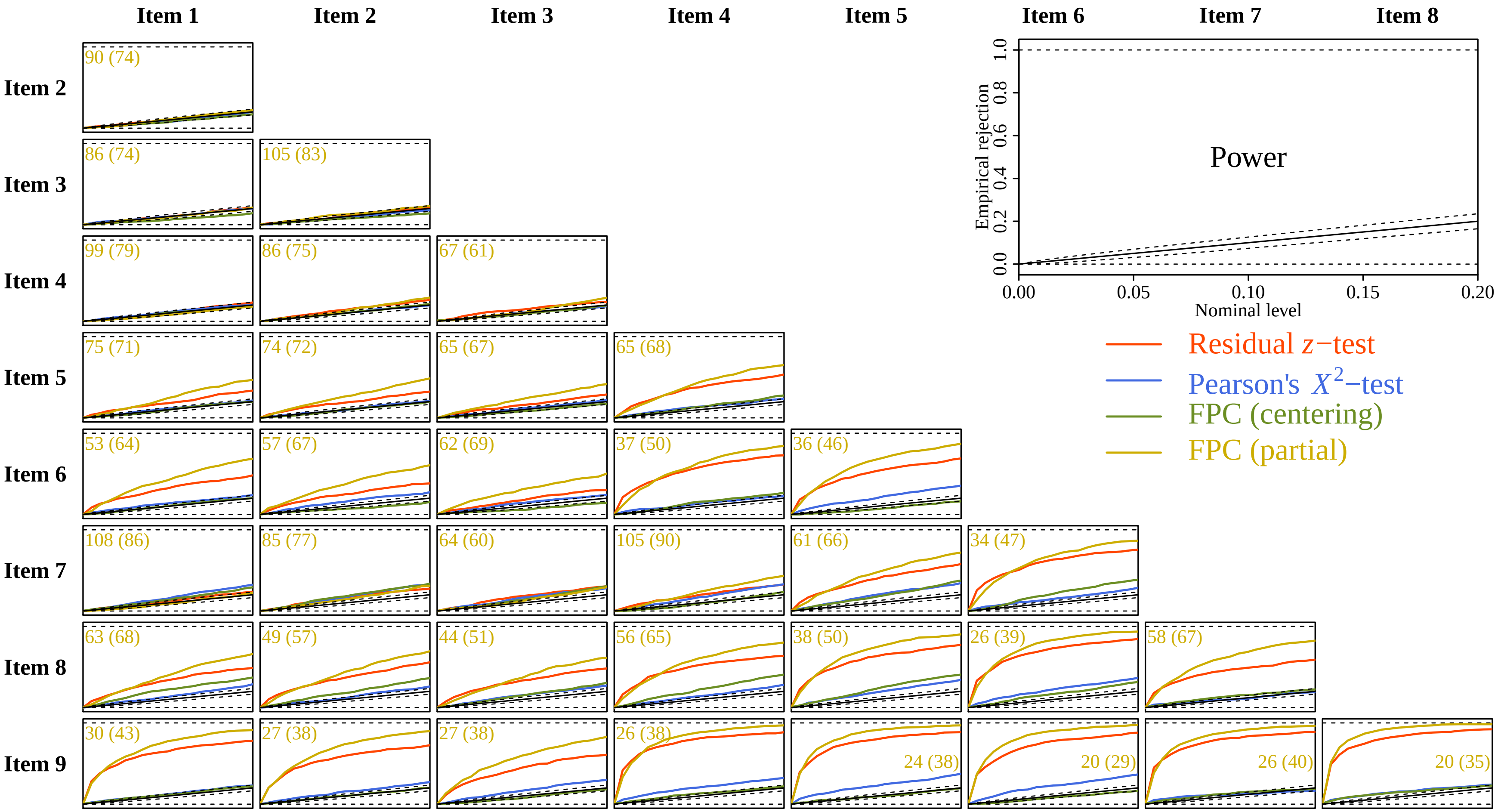

Figure 4.18: Power results for bivariate goodness of fit tests: True model = mixture. In each panel, the x-axis is the nominal α-level, and the y-axis is the empirical rejection rate pooling across 500 replications. The slanted solid line indicates empirical rejection equal to the nominal level, and the slanted dashed lines give a 95% normal approximation confidence band. The horizontal dashed line indicates the maximum power 1. Different statistics are displayed in different colors. The numbers displayed in each panel are the means and standard deviations (in parenthesis) of ESS associated with the partial FPC approach.

Finally, we summarize the simulation results for data generated from a unidimensional GRM model in which the latent variable is a mixture of two independent normal variates; graphical tables are presented as Figure 4.17 and 4.18.

Among all score-group fit diagnostics of interest, only statistics $T_{(2)}$ and $T_{(9)}$ are recognizably responsive to the misspecified latent variable distribution; in particular, the partial FPC is more powerful than the centering FPC. This is attributable to the shape of the latent variable distribution in the data generating model: The distribution has two modes located at the low and high ends of the standard normal scale, respectively, so the observed proportion of extreme score groups are under-estimated when a standard normal latent variable is fitted. We also observe that the $M_2^\star$ test rejects even less than the nominal level. The fact that $M_2^\star$ and the observed score profile combine information in the full item response contingency table differently might be the main source of the differential power, which itself has little to do with the comparison between likelihood-based and fiducial inference. Future investigations focusing on the latter comparison are encouraged: For example, statistics based on a summary of univariate and bivariate margins should be developed for FPC, and their performance should be compared with M_2.

For pairwise tests, the power to reject the fitted model increases as the true slope values associated with the pair increase. For the low-slope items in the current simulation, the generating model is almost indistinguishable from the independence model, which makes irrelevant the latent variable distribution; that explains why misspecified latent variable distribution is seldom identified for low-slope pairs. The partial FPC is the most powerful procedures, followed by the bivariate z-test; in contrast, the centering FPC and the X^2-test do not reject much more than the nominal α-level.

For power performance, different test statistics summarize information in different ways; consequently, for a specific type of data generating model differing from the fitted one, some diagnostics might be more sensitive than others. In the current work, we only consider misfitting the dimensionality or the shape of the latent variable distribution, which are alternative models to the unidimensional GRM that have been widely used in practice. Both types of overall model misspecifications can be detected by test statistics based on the sum score profile; in contrast, the $M_2^\star$ statistic fails to identify the incorrectly specified latent variable distribution. Therefore, the former approach is preferred. At the item pair level, however, no clear winner can be found among the candidates being investigated. The centering FPC rejects most of the fitted models for within-factor pairs under the three-dimensional alternative, but rejects no more than the nominal level for between-factor pairs and under the mixture alternative. The partial FPC, despite its decent power to detect the mixture model, is also not able to identify between-factor pairs. The z-test has some power to detect most pairs under both alternative models; however, it suffers from an inflated Type I error, and occasionally the statistic cannot be computed due to a negative asymptotic variance estimate.

4.5 Bifactor models: Simulation design

In the final part of this chapter, we compare fiducial and likelihood-based interval estimators in recovery of bifactor model parameters. We considered a fully crossed design with two levels of sample size, $n = 200$ and 500, and two levels of test length, $m = 9$ and 18; under each condition, 500 data set were simulated. The data generating model has a general latent variable that loads on all items in the test and three secondary latent variables each of which links to one-third of the items; the four latent variables are orthogonal to each other. Each item has $K_j = 3$ categories.

Similar to the previous simulation study with unidimensional GRMs, we converted item parameters to the standardized scale and manipulated the values. The standardized factor loading vector under a multidimensional GRM model can be expressed as

$$\boldsymbol{\lambda}_j = \frac{\boldsymbol{\beta}_j/1.7}{\sqrt{1+\boldsymbol{\beta}_j{}^\top\boldsymbol{\beta}_j/1.7^2}}, \tag{4.6}$$

extending the previous expressions (Equation 4.1). In the data-generating model, the loading/slope vector for each item j has only two non-zero elements, corresponding to the related primary and secondary latent variables for that item. In our simulation study, the two design factors that determine the true loading parameters are: a) communality $\boldsymbol{\lambda}_j{}^\top\boldsymbol{\lambda}_j$, and b) relative impact of the general factor, measured by the proportion of common variance explained by the common factors. Values 0.1, 0.5, and 0.9 were chosen to represent low, medium, and high communality levels. The weak, moderate, and strong general factors explain 25%, 50%, and 75% of the common variance, respectively. Similar to Equation 4.2, the multidimensional version of the threshold parameter is defined as

$$\tau_{jk} = \frac{\alpha_{jk}/1.7}{\sqrt{1 + \boldsymbol{\beta}_j{}^\top\boldsymbol{\beta}_j/1.7^2}}, \ k = 1, 2. \tag{4.7}$$

The threshold parameters used in data generation were $(-1.25 \ 0.25)^\top$ for odd items, and $(-0.25 \ 1.25)^\top$ for even items.

In addition to the standardized and the original slope-intercept parameterization, we are also interested in a multidimensional generalization of the item difficulty parameter that is interpreted somewhat differently from the unidimensional version of Equation 4.3. In a unidimensional GRM, the difficulty parameter $\delta_{jk} = -\alpha_{jk}/\beta_j$ can be alternatively explained as the solution of z_i to the equation $\alpha_{jk} + \beta_j z_i = 0$, which pins down the center of the corresponding logistic curve, i.e., $1/(1 + e^{-\alpha_{jk}-\beta_j z_i}) = 1/2$. When $\boldsymbol{\beta}_j$ is r-dimensional, however, the corresponding linear equation $\alpha_{jk} + \boldsymbol{\beta}_j{}^\top\mathbf{z}_i = 0$ determines an $(r-1)$-dimensional linear subspace, in which there are infinitely many values that maps onto a 50% probability. In this case, we compute the distance from the origin of the latent variable space to the subspace, and consider it an overall measure of item intensity (see, e.g., Reckase, 2009). The formula for such distance, which is subsequently referred to as the item difficulty parameter, can be straightforwardly derived using geometry:

$$\delta_{jk} = \inf\{\|\mathbf{z}_i\| : \alpha_{jk} + \boldsymbol{\beta}_j{}^\top\mathbf{z}_i = 0\} = -\frac{\alpha_{jk}}{\sqrt{\boldsymbol{\beta}_j{}^\top\boldsymbol{\beta}_j}}. \tag{4.8}$$

Table 4.2 shows the data generating values of the five types of parameters for the first nine items; these values were repeated twice in the 18-item conditions.

Table 4.2: Data-generating parameter values for the bifactor GRM ($m = 9$)

Item	Communality	General factor	Loading				Thresholds	
			λ_{j1}	λ_{j2}	λ_{j3}	λ_{j4}	τ_{j1}	τ_{j2}
1	low	strong	0.27	0.16	0.00	0.00	-1.25	0.25
2	low	moderate	0.22	0.00	0.22	0.00	-0.25	1.25
3	low	weak	0.16	0.00	0.00	0.27	-1.25	0.25
4	medium	strong	0.61	0.35	0.00	0.00	-0.25	1.25
5	medium	moderate	0.50	0.00	0.50	0.00	-1.25	0.25
6	medium	weak	0.35	0.00	0.00	0.61	-0.25	1.25
7	high	strong	0.82	0.47	0.00	0.00	-1.25	0.25
8	high	moderate	0.67	0.00	0.67	0.00	-0.25	1.25
9	high	weak	0.47	0.00	0.00	0.82	-1.25	0.25

Item	Slope β_j				Intercepts		Difficulties	
	β_{j1}	β_{j2}	β_{j3}	β_{j4}	α_{j1}	α_{j2}	δ_{j1}	δ_{j2}
1	0.49	0.28	0.00	0.00	2.24	-0.45	-3.95	0.79
2	0.40	0.00	0.40	0.00	0.45	-2.24	-0.79	3.95
3	0.28	0.00	0.00	0.49	2.24	-0.45	-3.95	0.79
4	1.47	0.85	0.00	0.00	0.60	-3.01	-0.35	1.77
5	1.20	0.00	1.20	0.00	3.01	-0.60	-1.77	0.35
6	0.85	0.00	0.00	1.47	0.60	-3.01	-0.35	1.77
7	4.42	2.55	0.00	0.00	6.72	-1.34	-1.32	0.26
8	3.61	0.00	3.61	0.00	1.34	-6.72	-0.26	1.32
9	2.55	0.00	0.00	4.42	6.72	-1.34	-1.32	0.26

The candidate interval estimators in this section are fiducial percentile CIs, the Hessian-form, and the cross-product-form ML Wald CIs. The Gibbs sampler for fiducial estimation was configured similarly to the unidimensional runs, with an exception that the parameter bound in the initial bounding box (Equation 3.8) was set to $M = 50$. Again, the value of M was selected to ensure desirable coverage while retaining efficiency; a larger value of M, implying less shrinkage for parameters, was specified for the fitted bifactor model due to the concern that it is not as well-identified as the unidimensional model. ML estimates and standard errors were computed using the software M*plus* 7.0 with the same quadrature and convergence settings. Note that although the latent variable is four dimensional in the fitted model, the effective dimensionality for each item is only two; the software implemented by default an automatic dimension reduction strategy (see Gibbons and Hedecker, 1992; Cai, 2010c) so that the numerical integration was conducted on a two-dimensional space.

4.6 Bifactor models: Parameter recovery

We computed empirical coverage probabilities and median lengths of the three type of interval estimators for all seven types of parameters, and tabulated them in Figure 4.19 to 4.22.

There were identification issues when fitting the bifactor GRM with ML estimation: For short tests ($m = 9$), there were 20 replications for which the Hessian matrix of the log-likelihood is non-positive-definite when $n = 200$, and 19 such replications when $n = 500$; for long tests ($m = 18$), there were 59 such replications when $n = 200$, and 13 such replications when $n = 500$. In those cases, Wald-type intervals with the Hessian-form standard errors cannot be computed. Not only does the poor identification lead to numerically ill-conditioned Hessian matrices, it also reduces the coverage of the resulting interval estimators. For high-communality items (items 7, 8, and 9), both types of Wald CIs have low coverage for intercept and slope parameters; in the small-sample conditions, the coverage can be even lower than 0.8. Their performance slightly improves when the sample size increases to 500, but the empirical coverage probabilities are still significantly lower than the nominal 95% level. The Hessian-form intervals have low coverage for large loading parameters and small difficulty parameters, which is similar to what has been observed in the unidimensional simulations, and is conjectured to be ascribed to the failure of the quadratic approximation to the log-likelihood function when the true values approach the boundary. Across all the conditions and parameterizations under investigation, the cross-product-form standard errors are uniformly larger than the Hessian-form ones, and thus in most cases fare ineffective.

In comparison, fiducial CIs are able to maintain on-target coverages for most parameters of interest. They can be slightly liberal for large factor loading values (e.g., both primary and secondary loadings for item 7, 8 and 9); but even in those cases, the coverages are almost

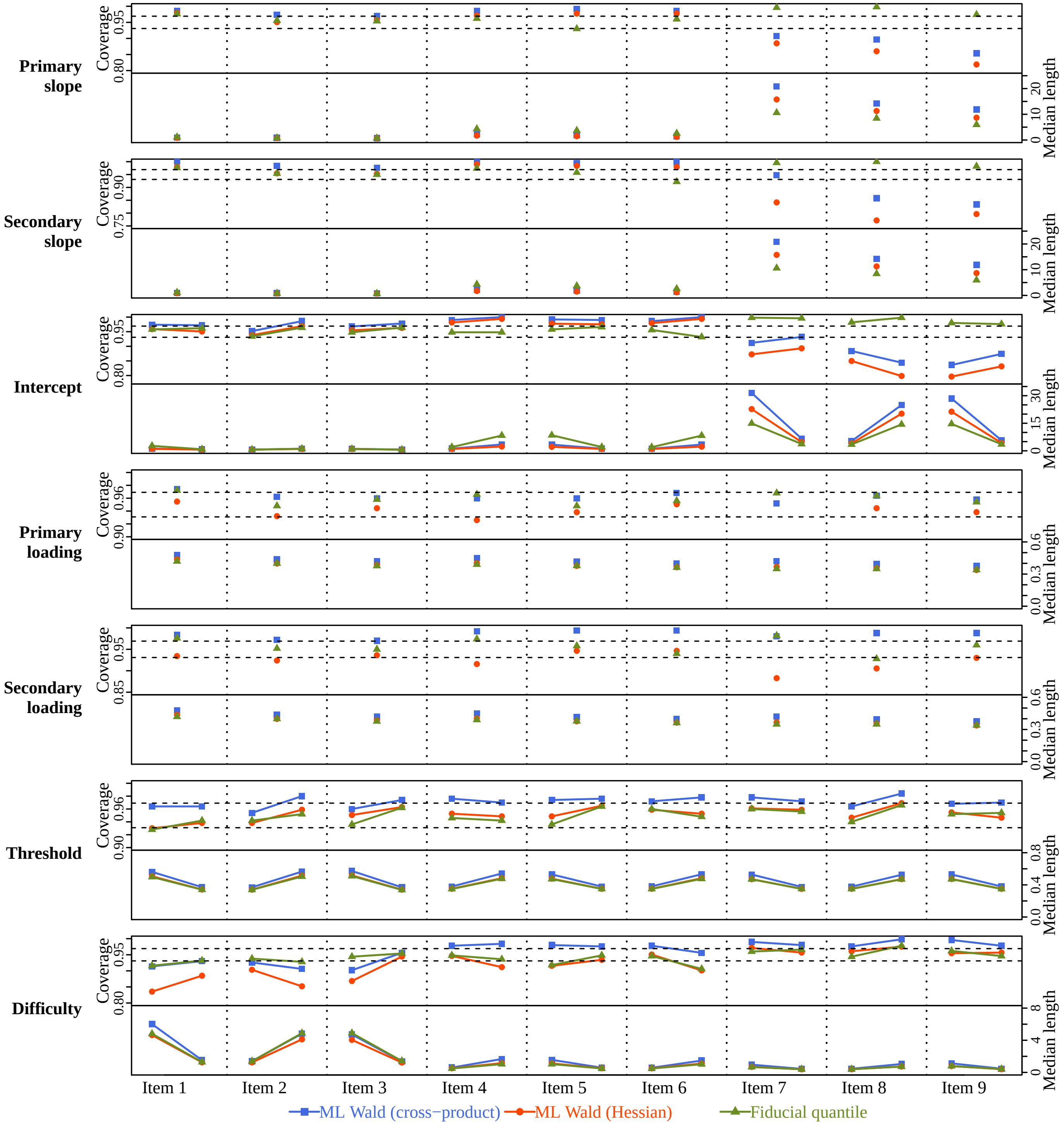

Figure 4.19: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 200$ and the number of items $m = 9$. Each row corresponds to one type of parameters, in which coverage are plotted in the upper panel and median length on the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

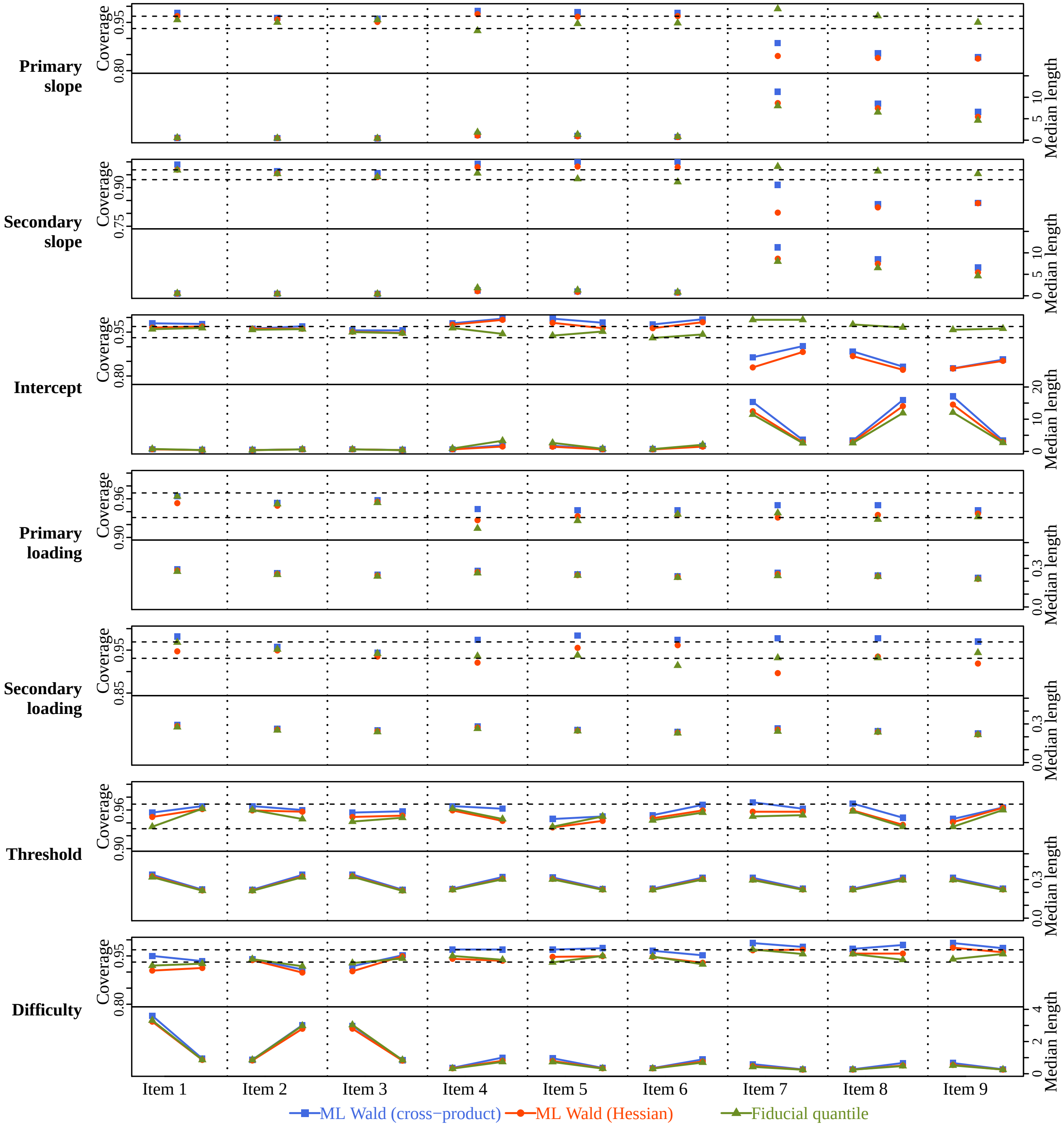

Figure 4.20: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 500$ and the number of items $m = 9$. Each row corresponds to one type of parameters, in which coverage are plotted on the upper panel and median length on the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

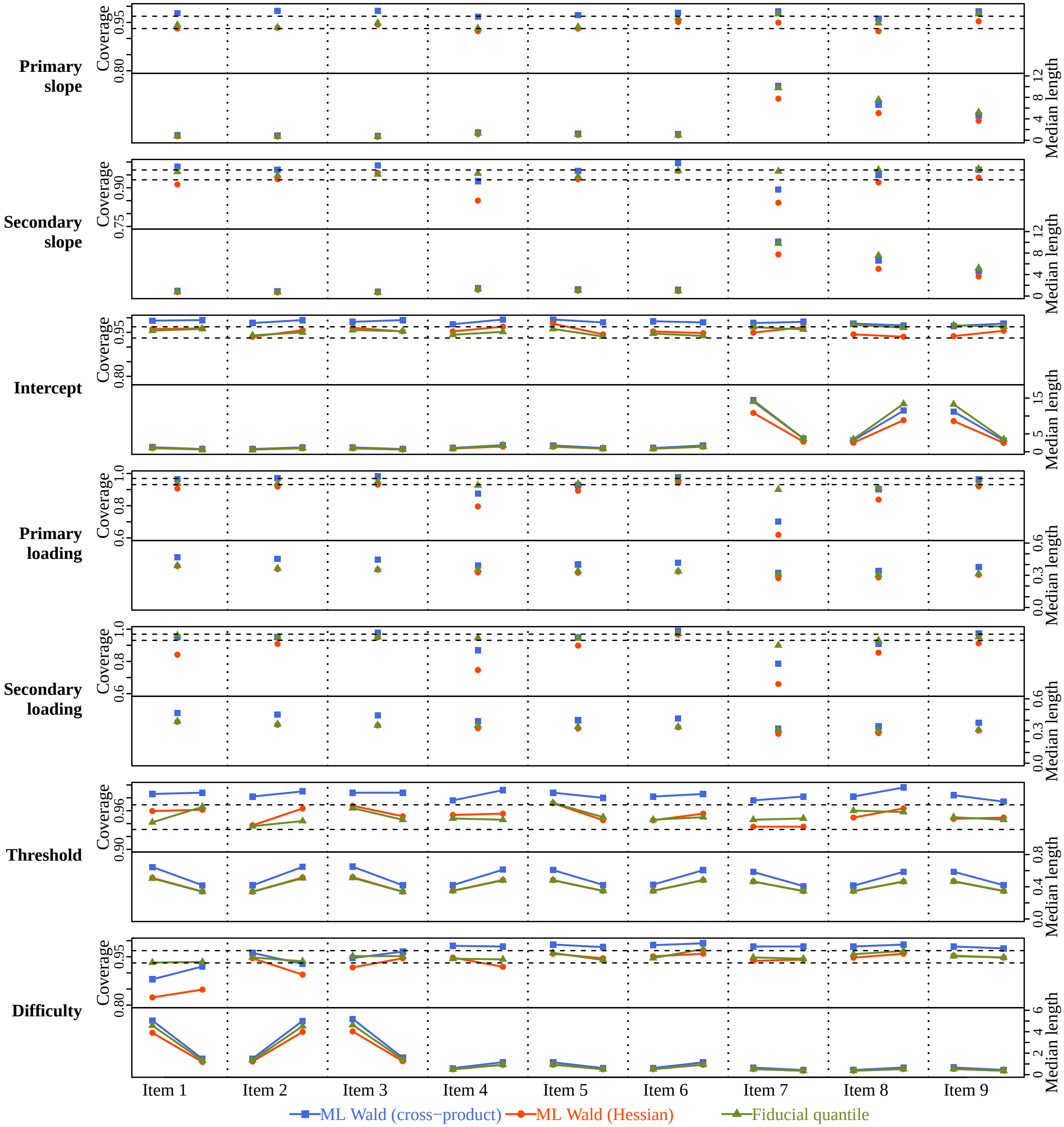

Figure 4.21: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 200$ and the number of items $m = 18$. Each row corresponds to one type of parameters, in which coverage are plotted on the upper panel and median length on the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

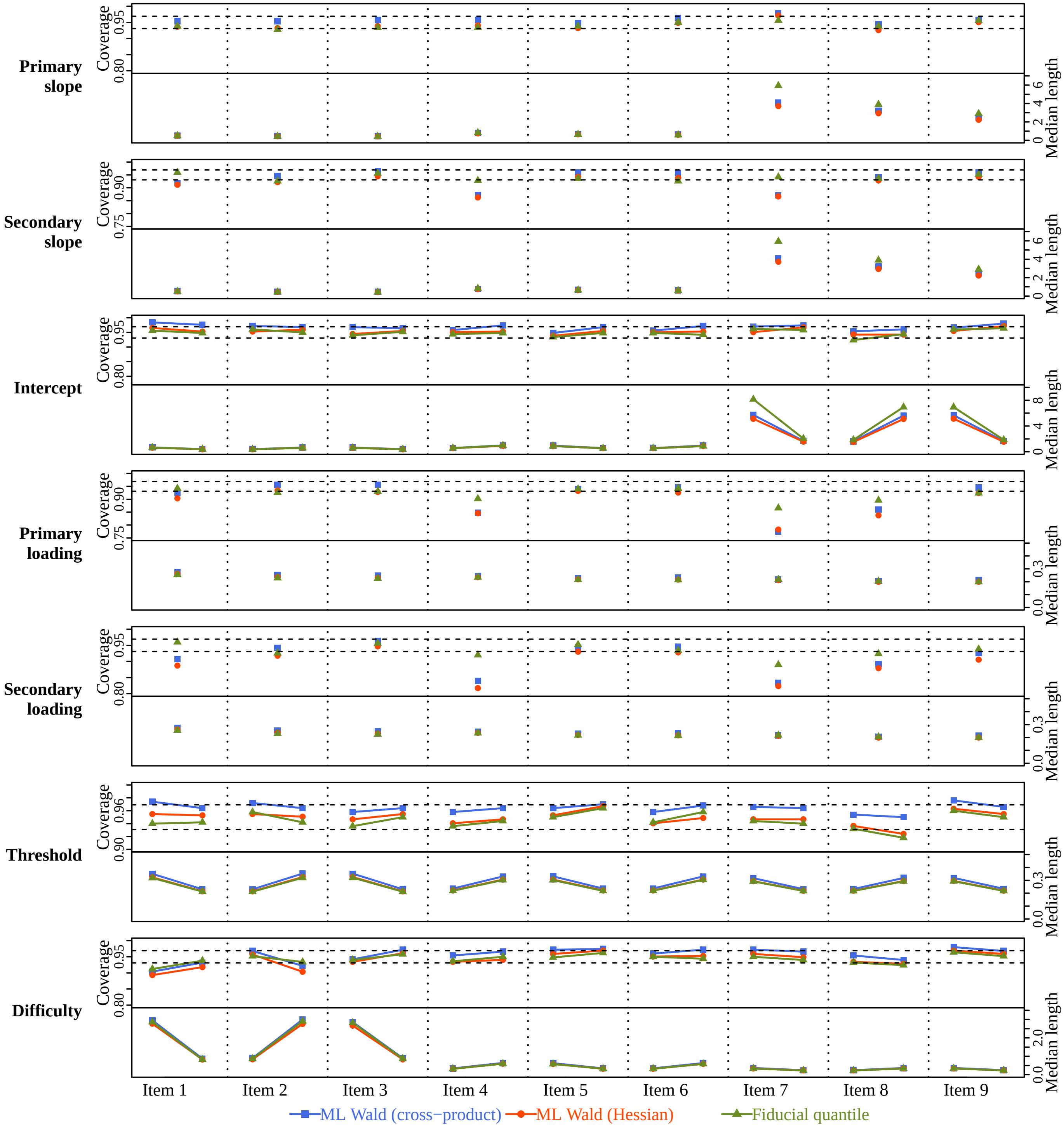

Figure 4.22: Empirical coverage and median length of the four types of interval estimators (shown in different colors). Here, the sample size $n = 500$ and the number of items $m = 18$. Each row corresponds to one type of parameters, in which coverage are plotted on the upper panel and median length on the lower panel, and parameters belonging to different items are separated by vertical dotted lines. The two horizontal dashed lines on the coverage panel gives a 95% normal-approximation confidence band for the nominal level 0.95.

always higher than those of the Wald CIs. However, the good coverage is accompanied with a noticeable cost of efficiency: For instance, wider intervals are resulted for the slopes and intercepts of items 4, 5, and 6 when the tests are short, and those of items 7, 8, and 9 when the tests are long.

4.7 Conclusion

We conclude that GFI is able to offer reliable statistical inference in extreme situations, i.e., the combination of small sample sizes and extreme true parameter values, where conventional likelihood-based and Bayesian approaches may fail. This suggests that GFI is particularly useful in substantive studies when a large calibration sample is not practical (e.g., sampling from a clinical population) and/or some of the items of interest have very low endorsement rates (e.g., suicidal attempt items in a depression scale). Methodologists have long been hesitant to calibrate items with extremely low endorsement rates, because they are likely to cause troubles in the numerical search of the ML solution. Even if the ML estimates and standard errors can be obtained, inferences thereof are less trustworthy based on our simulation results, especially when the calibration sample is not large. Now, with the aid of the proposed Gibbs sampler, well-calibrated fiducial CIs for extreme item parameters are obtainable in very small samples, and thus those low-endorsement items can be more appropriately handled.

Item parameters control the sparseness of the univariate margin. Together with the Type I error results of FPC, we conjecture that GFI is in general less affected by the sparseness of the item response contingency table, which has been identified as the major source of the discrepancy between the empirical performance and asymptotic theory of conventional ML-based inference procedures. The higher-order asymptotic properties of the fiducial distribution, which has yet been fully established, may serve as the theoretical base for such an observation.

Although the proposed implementation of GFI has been applied to unidimensional and bifactor GRMs with a certain degree of success, there are limitations to be addressed by future research.

Better mechanisms to determine an optimal value of the parameter bound M involved in Equation 3.8 is a potential topic for future research. It is a crucial tuning parameter required by the sampling algorithm, and we found in our pilot study that it does exert an influence on the results when the sample size is relatively small compared to the magnitude of the data-generating parameter values. A practical recommendation based on our experience is to run the sampler with multiple choices of M values from high to low, and select the smallest one before the interval estimates start to vary. In some scenarios, it might also be reasonable to allow different M values for different items.

We should also explore other choices of test statistics to be used in junction with the flexible FPC framework. It has been observed in the power simulations that the current choices of statistics may have low power to detect certain types of model misspecification at both the global and pairwise scope. For example, an score-group statistic based on all bivariate margins, analogous to the M_2 and $M_2^\star$ statistics, may be more powerful for revealing misspecified dimensionality of the latent variable. For item pairs, the importance weights calculated in the partial FPC tend to concentrate on very few Monte Carlo draws, and thus may trigger the unstable performance. It could be induced by the normal approximation of the statistic's likelihood (see Appendix E), or by the specific choice of test statistic, i.e., the cross-product moment (Equation 2.33). Thorough investigations at both theoretical and empirical levels should be conducted in order to resolve this problem.

CHAPTER 5: EMPIRICAL EXAMPLE

In this chapter, ordinal response data from the Patient Reported Outcomes Measurement Information System (PROMIS; Irwin, Stucky, Langer, Thissen, DeWitt, Lai, Varni, Yeatts, and DeWalt, 2010) study of chronic illness are analyzed using the proposed implementation of generalized fiducial inference (GFI). Several interesting inferential examples are presented, including goodness of fit assessment and interval/set estimation for complex transformations of parameters, all of which are derived straightforwardly from a Monte Carlo sample of the fiducial distribution generated by the proposed Gibbs sampler. We illustrate that GFI, similar to Bayesian inference, is more flexible and straightforward than traditional likelihood-based methods in terms of accounting for the variability of parameter estimation—an important part of many inferential procedures which has long been ignored in practice. GFI does not require prior specification, which circumvents the subjective judgment involved in Bayesian analysis.

The data set comprises 455 complete responses to 22 short-form items that are designed to measure three aspects of emotional distress: anger (6 items), anxiety (8 items), and depression (8 items). Although the proposed method is able to handle missing data in a natural fashion, we applied listwise deletion in order to mimic the scenario of small-sample calibration. All items have five response categories. The common response scale ranges form 0 to 4: 0 = never, 1 = almost never, 2 = sometimes, 3 = often, and 4 = almost always. The items stems are tabulated in Table 5.1.

A unidimensional GRM was fitted as the base model. It has been repeatedly observed that anger, anxiety and depression are highly correlated; as a result, the one-dimensional latent variable is likely to reflect a general dimension of emotional distress. However, items measuring the same symptomatology tend to co-vary more than can be explained by the overall emotional distress latent variable. The global and local fit of the unidimensional model was checked using the fiducial predictive check (FPC). Next, we used a three-dimensional exploratory item factor analysis (EIFA) to investigate the underlying factor structure of the scale. Similar to Liu and Hannig (2014), we constructed confidence intervals (CIs) for rotated factor loadings and factor intercorrelations, which are implicitly defined transformations of item slopes. Finally, we fit a bifactor model with a primary dimension on which all items load and three secondary dimensions for the items of the three subscales. Using the bifactor model, we discuss inferential examples including drawing bivariate confidence contours for a pair of item parameters and confidence bands (CBs) for functional transformations of item parameters.

All the models were fitted by the same Fortran program that has been used in the simulation study. Parameter estimates from the default limited information estimator in M*plus* (i.e., `estimator = WLSMV`) were obtained in advance and set as the initial parameter values $\boldsymbol{\theta}^{(0)}$; the corresponding factor score estimates (i.e., `save = fscores`) were used as the starting values for the normal variates $\mathbf{z}^{(0)}$. Item parameters were restricted to the bounding box (Equation 3.8) with $M = 50$; the starting Algorithm 7 was then executed to produce $\mathbf{a}^{(0)}$, which completes the initialization stage of the sampler. For each model of interest, a total of 60000 Markov chain Monte Carlo cycles were obtained; after burning in the first 10000 cycles to reduce the impact of starting values, we recorded 1/10 of the rest 50000 Monte carlo samples by thinning. Therefore, all the statistical procedures discussed in the sequel, unless specifically indicated, are based on 5000 draws.

5.1 A unidimensional model

First, a unidimensional GRM was fitted to all 22 emotional distress items, and the model fit was checked in sum-score profile and bivariate marginal tables by both centering and partial FPCs. For fit to sum-score levels, we split the entire score range into 10 groups and computed the observed proportion in each group as test statistics; the results are shown in

Table 5.1: PROMIS emotional distress short-form items

Label	Item stem
Ang1	I felt fed up.
Ang2	I felt mad.
Ang3	I felt upset.
Ang4	I was so angry I felt like throwing something.
Ang5	I was so angry I felt like yelling at somebody.
Ang6	When I got mad, I stayed mad.
Anx1	I worried about what could happen to me.
Anx2	I was afraid that I would make mistakes.
Anx3	I felt nervous.
Anx4	I felt like something awful might happen.
Anx5	I felt scared.
Anx6	I worried when I went to bed at night.
Anx7	I thought about scary things.
Anx8	I felt worried.
Dep1	I felt alone.
Dep2	I felt like I couldnt do anything right.
Dep3	I felt everything in my life went wrong.
Dep4	I felt sad.
Dep5	I thought that my life was bad.
Dep6	I could not stop feeling sad.
Dep7	I felt lonely.
Dep8	I felt unhappy.

Figure 5.1. We also examined model fit to each pair of items with the bivariate cross-product statistic (Equation 2.33); a graphical summary of results can be found in Figure 5.2.

It is anticipated that there exist residual associations among the items that measure the same domain; however, both the centering and partial FPCs based on the observed sum-score profile do not suggest model misfit (see Figure 5.1). The smallest p-value, i.e., 0.02, is observed for the lowest score group when the centering FPC is used. Although the numerical value is less than 0.05, it is likely to be a false positive case out of the multiple hypothesis tests displayed in Figure 5.1. It has been discovered in the simulation study that the sum-score profile approach is not particularly sensitive to the multidimensional alternative model, even when the true factors are distinct (correlation = 0.3). For the emotional distress scale, we expect much higher inter-correlations among the three domains (which is confirmed later by the EIFA); therefore, the non-significant results might be attributed to the further dampened power of the testing procedures.

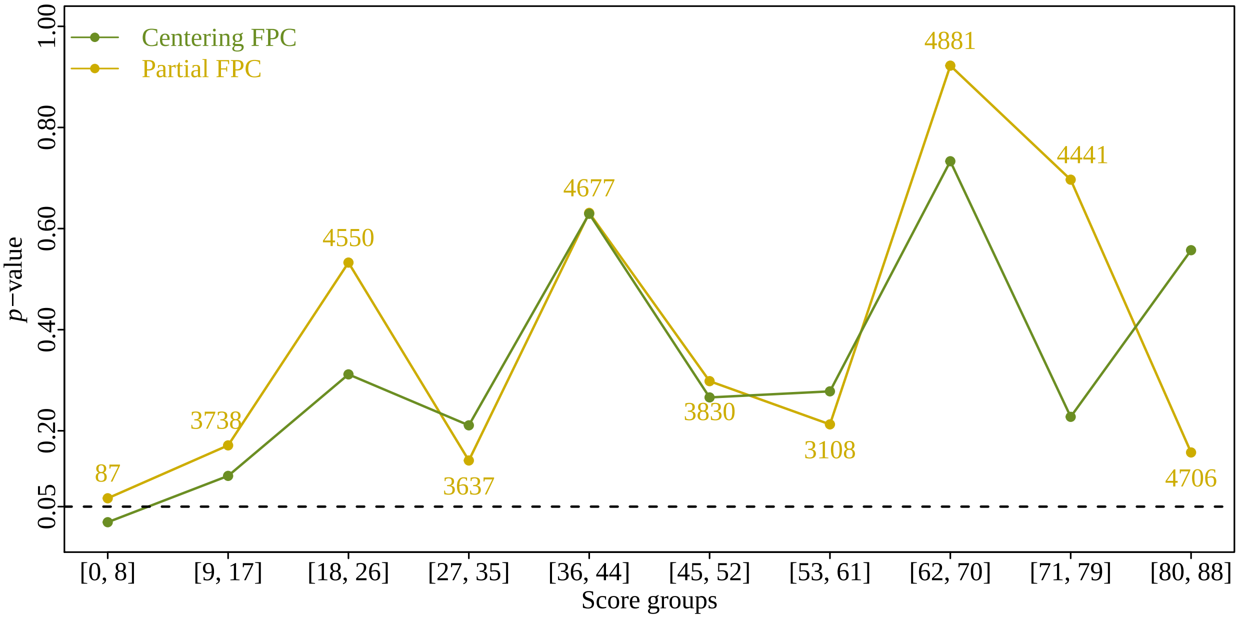

Figure 5.1: Fiducial predictive p-values for the sum-score group fit of a unidimensional GRM. The predictive p-values are plotted against the score groups. The centering and partial FPCs are shown in different colors; the numbers in the plot are the efficient sample sizes (ESS') associated with partial predictive p-values. The horizontal dashed line marks the nominal α-level 0.05.

In contrast, the bivariate tests suggest a clear pattern of symptom-specific clustering of items (see Figure 5.2): The bivariate covariance for items measuring the same domain tend to be significantly under-estimated by the fitted unidimensional model, while the cross-domain associations are likely to be over-estimated. The direction of misfit, which is shown by color codes in Figure 5.2, is determined differently for the two methods of FPC. For the centering approach, a positive residual after subtracting the asymptotic mean of the test statistic indicates under-estimation, and similarly a negative residual indicates over-estimation. For the partial approach, the direction is determined by which tail of the fiducial predictive distribution the observed test statistic falls into. We conclude that Figure 5.2 indicates the existence of three factors for anger, anxiety, and depression items, respectively; however, we also notice that the depression items (i.e., the third diagonal block) show less severe misfit, and thus the corresponding factor is relatively weak. In addition to the obvious diagonal blocks of within-domain pairs, items Anx2 ("I was afraid that I would make mistakes") and Dep2 ("I felt like I couldn't do anything right") also exhibit a significantly positive residual dependency, which is identified by both the centering and partial FPC. This particular pair of items was also found by Liu and Thissen (2014) using their score test of local dependence. They explained that since both items are related to "making mistakes", the correlation is higher than can be explained by a general factor of emotional distress.

5.2 A three-dimensional exploratory model

Our bivariate fit diagnostics suggest that the scale has three underlying dimensions; next, we used a three-dimensional EIFA to examine such a structure. In EIFA, the standardized loading-threshold parameterization is often reported, for the reason that it is on a scale that facilitates the computation of variance/covariance of test items explained by the common factors. In addition, analytic rotation methods are often applied (see Browne, 2001, for a review) to the estimated factor loadings to obtain more interpretable patterns of item-factor dependency. Our goal in this section is to obtain point estimates and CIs for rotated factor loadings and factor inter-correlations.

The EIFA amounts to the minimally constrained model for a given number of factors. Here, the fitted three-factor EIFA model was parameterized as Equation 1.1 with $\boldsymbol{\beta}_1 = (\beta_{11}\ 0\ 0)^\top$, $\boldsymbol{\beta}_2 = (\beta_{21}\ \beta_{22}\ 0)^\top$, and $\boldsymbol{\beta}_j = (\beta_{j1}\ \beta_{j2}\ \beta_{j3})^\top$ for $j = 3, \cdots, 22$. Then the (multidimensional) unrotated factor loadings $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{22}$ can be calculated by Equation 4.6. Let $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1\ \cdots\ \boldsymbol{\lambda}_{22})^\top$ be the unrotated factor loading matrix. A well-known property of the exploratory model is rotational indeterminancy. Let $\mathbf{Q}$ be a 3×3 oblique rotation matrix that is invertible and the inverse is normalized, i.e., $\mathbf{Q}^{-1}\mathbf{Q}^{-\top}$ has unity on the diagonal. Define the rotated factor loadings $\tilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\mathbf{Q}$, and the factor inter-correlations $\boldsymbol{\Phi} = \mathbf{Q}^{-1}\mathbf{Q}^{-\top}$. Then, rotational indeterminancy refers to the fact that

$$\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \boldsymbol{\Lambda}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{Q}^{-\top}\mathbf{Q}^\top\boldsymbol{\Lambda}^\top = \tilde{\boldsymbol{\Lambda}}\boldsymbol{\Phi}\tilde{\boldsymbol{\Lambda}}^\top. \tag{5.1}$$

Equation 5.1 implies that for all eligible rotation matrices $\mathbf{Q}$ the rotated and unrotated solutions fit the data equally well; however, a properly selected rotation matrix may greatly simplify the pattern of the loading matrix and thus ease the interpretation. In particular, we select the $\mathbf{Q}$ matrix that minimizes the Crawford-Ferguson Quartimax criterion of simple structure (Crawford and Ferguson, 1970). The constrained optimization problem can be solved by a gradient projection algorithm, which has been implemented in the R package `GPArotation` (Bernaards and Jennrich, 2005). Consequently, the resulting rotated factor loadings $\tilde{\boldsymbol{\Lambda}}$ and correlation matrix $\boldsymbol{\Phi}$ are implicit non-linear transformations of the unrotated loadings $\boldsymbol{\Lambda}$. For maximum likelihood (ML) estimation, the Delta-method standard errors for the rotated ML solutions can be obtained via implicit differentiation; see Jennrich (1973) for details. With GFI, however, the fiducial distribution of the rotated solutions can be easily approximated by applying the transformation given in Equation 4.6 and then the rotation algorithm to each Monte Carlo sample from the marginal fiducial distribution of item slopes.

In the rotation algorithm, the sign and order of the factors are not identified, because altering them does not affect the value of the criterion function being minimized. In order to harmonize the orientation of the resulting factors across the 5000 Monte Carlo samples, we apply a matching procedure similar to that described by Asparouhov and Muthén (2012) in the context of Bayesian EIFA. Specifically, the sum of the rotated factor loadings related to each factor is constrained to be greater than zero in order to identify the sign. As for ordering, we select the particular permutation of the three factors such that the correspondingly permuted factor loading matrix has the least sum of squared differences from the one obtained from the fiducial median of the unrotated loadings.

The fiducial densities, point estimates, and confidence intervals for rotated factor loadings and factor inter-correlations are tabulated in Figures 5.3 and 5.4.

As expected, the quartimax rotation of three factors corresponds with with the three subsets of items measuring anger, anxiety, and depression, respectively. Occasionally, we observe items with cross-loadings. For example, items “I felt upset” (Ang6) and “I was afraid that I would make mistakes" (Anx2) also load on the third factor dominated by depression items. Some of the weaker cross-loadings might not be significantly different from 0 after the false coverage-statement rate (FCR; Benjamini and Yekutieli, 2005) is controlled. We recommend adjusting for FCR in practice (see Liu and Hannig, 2014, for a detailed description of the procedure); however, it is omitted here for simplicity. In addition, items "I felt alone" (Dep1) and "I felt lonely" (Dep7) have very strong loadings (fiducial median ≈ 0.9) on the third factor, and thus their fiducial densities are highly skewed. Our simulation findings suggest that fiducial percentile CIs are more reliable than the Wald-type CIs when the parameters are close to the boundary; thus, we are confident in the reported intervals.

The factor inter-correlations are moderately high, with the CIs covering the range 0.5–0.7; this implies that anger, anxiety, and depression are likely to co-occur. Motivated by the high correlation estimates among the three factors, we continue to distinguish the proportion of individual differences in general emotional distress shared by all three symptomatology domains from those domain-specific granularities. This leads naturally to our next analysis: i.e., fitting a bifactor model with a general factor on which all items load, and three domain-specific factors for the anger, anxiety, and depression items, respectively.

5.3 A bifactor model

The point estimates (fiducial median) and the confidence limits (percentile CI) of all slope and intercept parameters in the bifactor GRM are in 5.2. Interestingly, the pair of items "I felt alone" (Dep1) and "I felt lonely" (Dep7) are the only significant indicators of the depression-specific dimension; the two item stems are almost identical ("alone" versus "lonely"). This phenomena has been referred to as "theta-theft" in the existing literature[1]: i.e., a latent variable is defined by a small subset of locally dependent items which are similar to each other in construct-irrelevant aspects (Thissen and Steinberg, 2010, p.131). For the current analysis, it implies that the particular secondary dimension only captures the wording similarity of the two locally dependent items above and beyond the general emotional

---

[1] In the IRT literature, the Greek letter $\boldsymbol{\theta}$ has been commonly used to denote latent variables, which is different from our notation $\mathbf{Z}$ in the current work.

distress, and that, with those two items in the analysis, the construct of depression itself is almost indistinguishable from the macro-level construct of emotional distress.

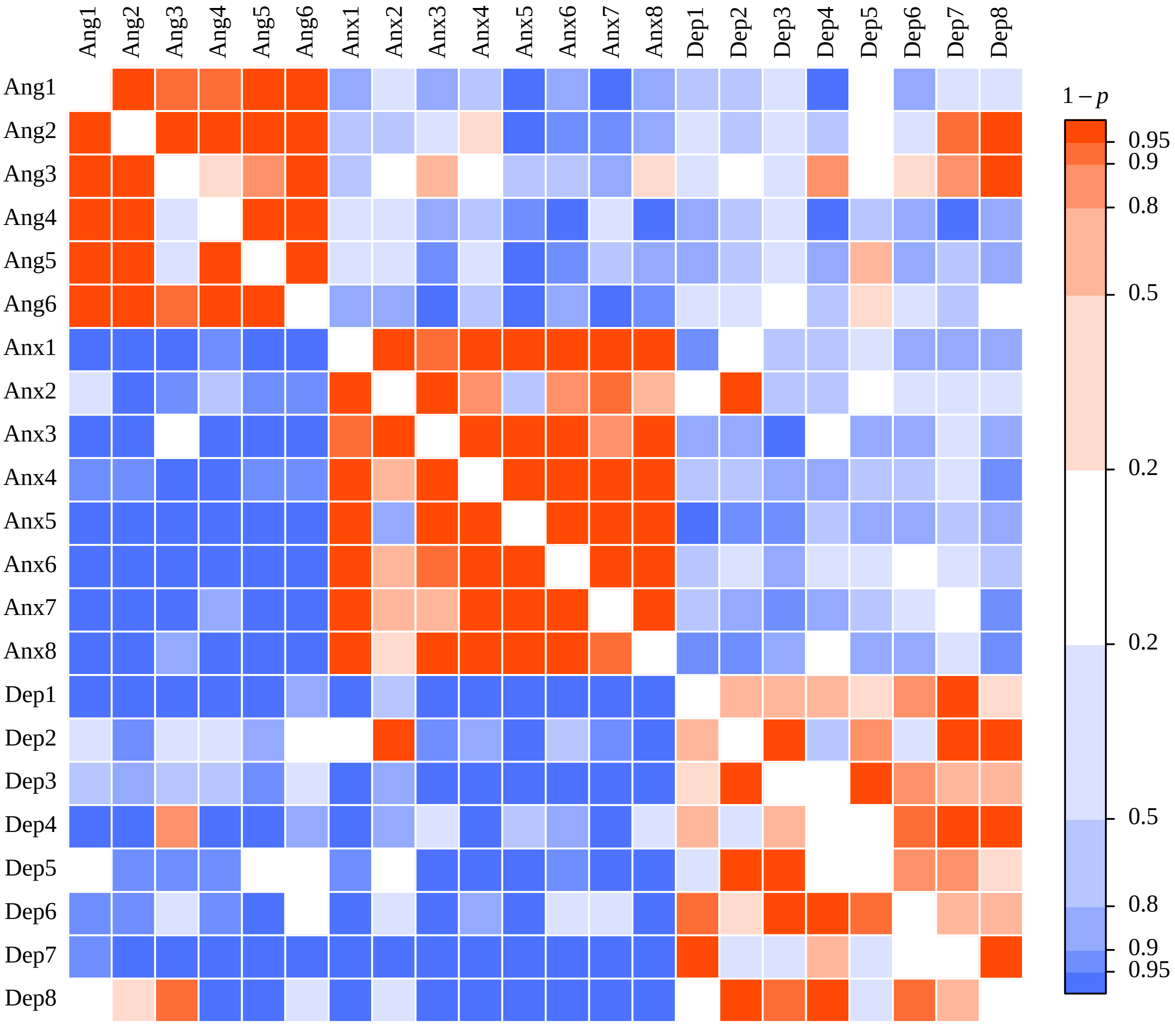

Figure 5.2: Fiducial predictive p-values for the bivariate goodness of fit of a unidimensional GRM. The lower triangle displays one minus the p-values resulted from the centering FPC, and the upper triangle shows those resulted from the partial FPC. The warm half of the color scale indicates under-estimation of bivariate associations, and the cold half indicates over-estimation.

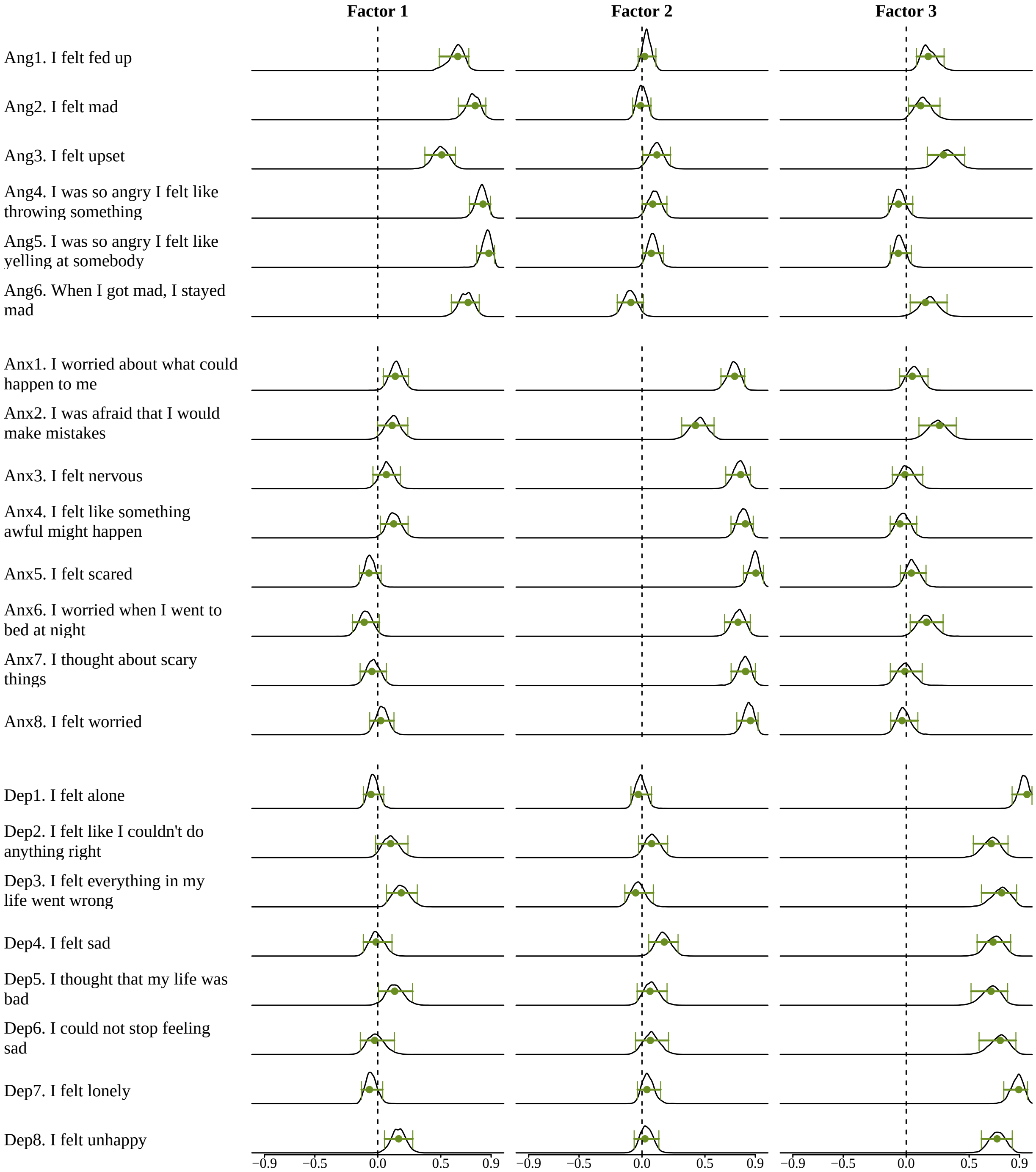

Figure 5.3: Fiducial median and 95% fiducial percentile CIs for rotated factor loadings in the three-factor EIFA. The tabular layout has a row for each item; item stems are listed in the leftmost column. The three columns of graphics correspond to the three factors, dominated by anger, anxiety, and depression items, respectively. Within each cell, the estimated fiducial density is shown in the background; the fiducial median (dot) and the percentile CI (interval) are superimposed. The 0 points on the factor loading scale are shown by vertical dashed lines.

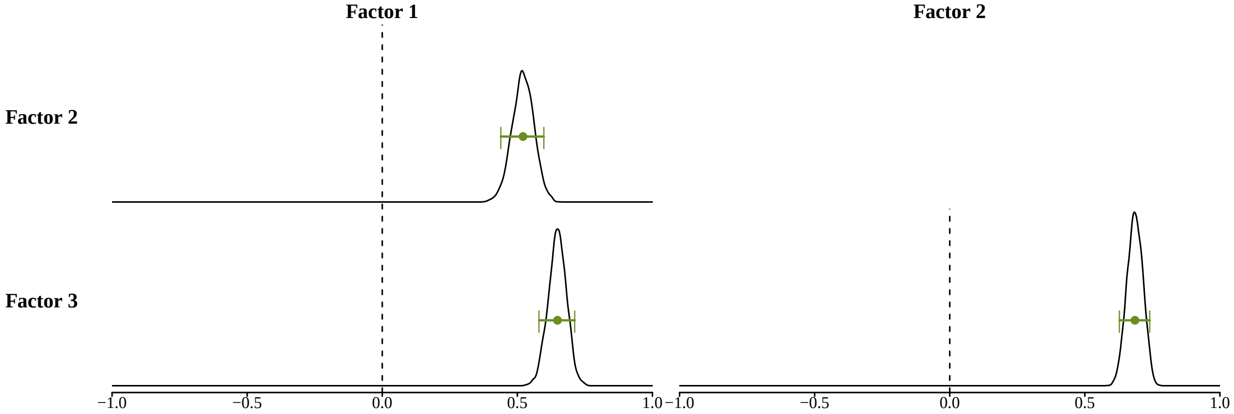

Figure 5.4: Fiducial median and 95% fiducial percentile CIs for factor inter-correlations in the three-factor EIFA. The tabular layout resembles a correlation matrix. Within each cell, the estimated fiducial density is shown in the background; the fiducial median (dot) and the percentile CI (interval) are superimposed. The 0 points on the correlation scale is highlighted by vertical dashed lines.

Table 5.2: Fiducial point and interval estimates for item slopes and intercepts under the bifactor GRM

Item	Primary slope			Secondary slope			Intercept 1			Intercept 2			Intercept 3			Intercept 4		
	L	M	U	L	M	U	L	M	U	L	M	U	L	M	U	L	M	U
Ang1	1.40	**1.70**	2.05	0.91	**1.20**	1.55	-0.02	**0.28**	0.56	-1.40	**-1.07**	-0.77	-3.82	**-3.31**	-2.85	-6.12	**-5.29**	-4.52
Ang2	1.74	**2.10**	2.54	1.27	**1.68**	2.14	1.35	**1.74**	2.18	-0.79	**-0.42**	-0.08	-4.75	**-4.02**	-3.43	-7.88	**-6.66**	-5.65
Ang3	1.92	**2.25**	2.64	0.68	**0.99**	1.31	0.81	**1.15**	1.47	-0.87	**-0.53**	-0.22	-4.58	**-3.94**	-3.41	-6.43	**-5.56**	-4.79
Ang4	1.52	**1.92**	2.40	1.44	**1.90**	2.43	-0.79	**-0.42**	-0.06	-2.30	**-1.83**	-1.41	-5.00	**-4.26**	-3.57	-6.84	**-5.78**	-4.84
Ang5	1.79	**2.27**	2.84	1.75	**2.24**	2.88	0.24	**0.67**	1.07	-1.65	**-1.14**	-0.74	-4.76	**-3.89**	-3.23	-7.60	**-6.30**	-5.28
Ang6	1.29	**1.59**	1.98	0.99	**1.30**	1.65	-0.64	**-0.33**	-0.04	-2.15	**-1.76**	-1.42	-4.45	**-3.86**	-3.32	-6.21	**-5.34**	-4.56
Anx1	1.90	**2.29**	2.77	1.33	**1.68**	2.08	-0.31	**0.05**	0.40	-2.02	**-1.59**	-1.21	-4.56	**-3.91**	-3.37	-6.51	**-5.60**	-4.82
Anx2	1.37	**1.65**	1.95	0.49	**0.72**	0.98	0.24	**0.48**	0.73	-1.05	**-0.79**	-0.52	-3.55	**-3.12**	-2.70	-5.50	**-4.77**	-4.14
Anx3	1.46	**1.78**	2.15	1.19	**1.53**	1.90	0.22	**0.53**	0.83	-1.45	**-1.08**	-0.77	-4.88	**-4.20**	-3.62	-7.27	**-6.19**	-5.27
Anx4	1.71	**2.12**	2.59	1.49	**1.88**	2.36	-0.75	**-0.38**	-0.05	-2.80	**-2.29**	-1.87	-5.94	**-5.07**	-4.33	-8.05	**-6.81**	-5.75
Anx5	1.95	**2.45**	3.04	1.80	**2.30**	2.93	-1.12	**-0.67**	-0.27	-3.09	**-2.52**	-2.00	-6.72	**-5.60**	-4.70	-9.79	**-8.08**	-6.72
Anx6	1.53	**1.90**	2.31	1.19	**1.59**	1.98	-1.40	**-1.02**	-0.68	-2.98	**-2.49**	-2.05	-5.33	**-4.57**	-3.91	-6.59	**-5.64**	-4.82
Anx7	1.19	**1.51**	1.87	1.22	**1.57**	1.97	-0.73	**-0.42**	-0.12	-2.13	**-1.75**	-1.39	-4.63	**-4.00**	-3.44	-6.04	**-5.19**	-4.45
Anx8	1.60	**1.97**	2.41	1.49	**1.88**	2.31	0.05	**0.38**	0.73	-2.05	**-1.65**	-1.26	-5.06	**-4.34**	-3.73	-7.43	**-6.29**	-5.37
Dep1	3.28	**4.33**	6.69	1.22	**2.12**	4.02	-2.74	**-1.54**	-0.92	-6.45	**-4.02**	-3.07	-12.83	**-8.28**	-6.46	-15.79	**-10.30**	-7.97
Dep2	2.04	**2.47**	2.99	-0.49	**-0.04**	0.42	-1.34	**-0.98**	-0.67	-2.98	**-2.48**	-2.05	-5.27	**-4.50**	-3.86	-7.04	**-6.00**	-5.11
Dep3	2.81	**3.56**	4.79	-1.51	**-0.53**	0.13	-2.31	**-1.67**	-1.15	-4.57	**-3.46**	-2.73	-7.66	**-5.81**	-4.70	-10.12	**-7.70**	-6.22
Dep4	1.99	**2.39**	2.81	-0.16	**0.24**	0.64	-0.01	**0.28**	0.58	-1.81	**-1.46**	-1.14	-5.03	**-4.33**	-3.74	-7.37	**-6.32**	-5.39
Dep5	2.33	**2.85**	3.52	-0.89	**-0.32**	0.22	-1.81	**-1.36**	-0.97	-3.48	**-2.86**	-2.35	-6.02	**-5.05**	-4.28	-9.09	**-7.48**	-6.24
Dep6	1.71	**2.08**	2.52	-0.23	**0.16**	0.55	-1.80	**-1.45**	-1.12	-3.16	**-2.69**	-2.27	-5.25	**-4.53**	-3.90	-7.22	**-6.11**	-5.16
Dep7	3.09	**4.10**	8.27	1.34	**2.29**	5.59	-2.87	**-1.42**	-0.82	-7.04	**-3.58**	-2.64	-15.00	**-7.76**	-5.96	-18.69	**-9.68**	-7.44
Dep8	2.45	**2.93**	3.52	-0.20	**0.18**	0.58	0.11	**0.45**	0.81	-1.89	**-1.48**	-1.11	-5.78	**-4.93**	-4.21	-8.16	**-6.98**	-5.95

L: Fiducial 2.5 percentile
M: Fiducial median (highlighted in bold)
U: Fiducial 97.5 percentile

The highest fiducial density regions (at nominal coverage levels 75%, 90%, and 95%) for each item's primary and secondary factor loading pairs are displayed in Figures 5.5 to 5.7. To obtain the contours on each two-dimensional parameter space, we used the R package `ks` (Duong, 2014) to estimate a two-dimensional density by kernel smoothing from the 5000 fiducial samples. We selected the bandwidth by the plug-in method (Wand and Jones, 1994) using function `Hpi()` in the `ks` package. The implementation relies on the optimizer `nlm()`, which was found to be very inefficient when the number of data points is large. As a result, a further thinned sample of 500 Monte Carlo draws was extracted for bandwidth selection, while the entire sample was still used for the subsequent density estimation.

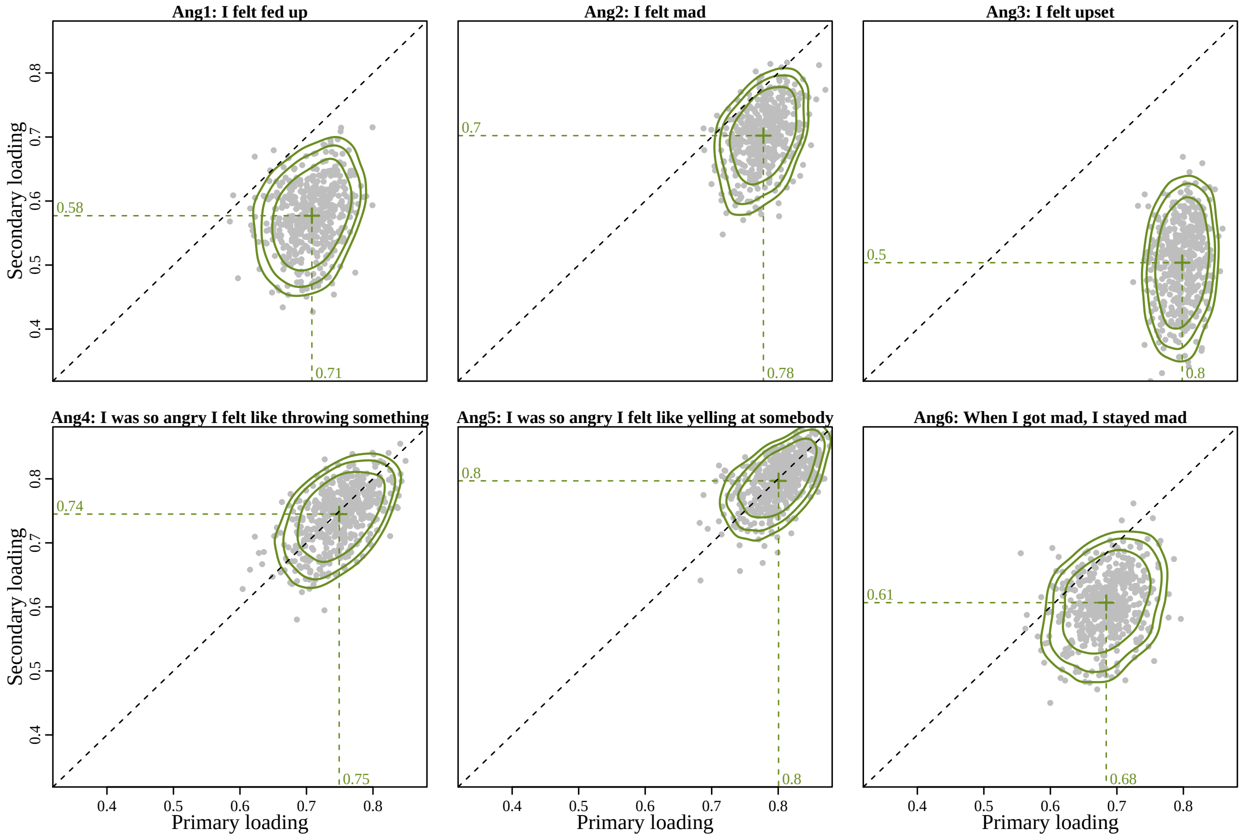

Figure 5.5: Two-dimensional confidence regions for the primary and secondary loadings of anger items. The points are 500 draws selected via a further thinning interval of 10. The three contours shown on each panel are the 75%, 90%, and 95% highest fiducial density regions. The fiducial medians are highlighted by green crosses, and the numerical values are also shown in green text. The diagonal dashed line indicates that an item contributes evenly to the primary and secondary factors.

We are able to visualize the relative contributions of primary and secondary factors to each item on those bivariate plots: The primary one dominates if the point cloud is below the diagonal line. This is the case for the anger item “I felt upset” (Ang7), the anxiety item “I was afraid that I would make mistakes” (Anx8), and all but the two locally dependent depression items. For the “alone/lonely” pair that governs the depression dimension, the point clouds are non-elliptical. In those cases, a normal approximation, which produces ellipses, is likely to yield contours of a completely different shape, affected by the downward tail trailing along the vertical direction.

In Figures 5.8 to 5.12, we demonstrate drawing CBs for functional transformations of item parameters, in order to account for the sampling variability induced by parameter estimation.

We first consider the marginal expected score curve, defined for each item $j = 1, \dots, 22$ on each latent dimension $d = 1, \dots, 4$ as

$$s(z_{id}, \boldsymbol{\theta}_j) = \sum_{k=1}^{K_j - 1} k \int_{\mathbb{R}^3} f_j(\boldsymbol{\theta}_j, k|\mathbf{z}_i) d\Phi(\mathbf{z}_{i,-d}), \tag{5.2}$$

in which $\mathbf{z}_{i,-d}$ is a three-dimensional vector corresponding to all but the dth dimensions. For the sake of succinctness, the expected score function has been preferred over the item response function in visualizing item characteristics of ordinal items. In computation, the three-dimensional integral in Equation 5.2 needs to be approximated by quadrature; the quadrature grid being used here amounts to an outer product of three identical lists of 21 equally-spaced points from -3 to 3. The effective dimension of integration is in fact two, because each item only has two non-zero slopes. On each secondary dimension in the fitted bifactor model, the marginal expected score curve for those items not loading on this particular dimension is flat (its level depends on the primary slope and the intercepts), and thus are not shown in Figures 5.9 to 5.11.

For fixed z_{id}, Equation 5.2 is just a single transformed parameter, and then its 95% fiducial CI can be computed as usual. Pooling across all $z_{id} \in \mathbb{R}$, we obtain the pointwise CB, the green dashed lines in Figures 5.8 to 5.11. A 95% pointwise CB only ensures that at each level of z_{id} the true marginal expected score is covered 95% of the time over repeated sampling; sometimes, however, an extended statement for all possible values of z_{id} is also in demand. Here, a procedure adapted from Thissen and Wainer (1990) is used to construct simultaneous CBs, the yellow dashed lines in Figures 5.8 to 5.11. Note that equation 5.2 depends on six free item parameters, denoted $\boldsymbol{\theta}_j$: four intercepts and two slopes. In order to determine a simultaneous CB for each curve, we first calculate a six-dimensional kernel density estimate (again, using the R package `ks`) from the 5000 fiducial draws of $\boldsymbol{\theta}_j$, each of which is a point defined on the six-dimensional parameter space. Next, we selected the draws enclosed in the 95% highest density region, and set the lower (upper) confidence limit at each z_{id} level to the minimum (maximum) value of Equation 5.2 among the selected $\boldsymbol{\theta}_j$ values. Because the true six-dimensional parameter vector falls in the 95% highest fiducial density region (approximately) 95% of the time over repeated sampling, the entire true marginal expected score curve is then covered by the simultaneous CB in at least those 95% replications.

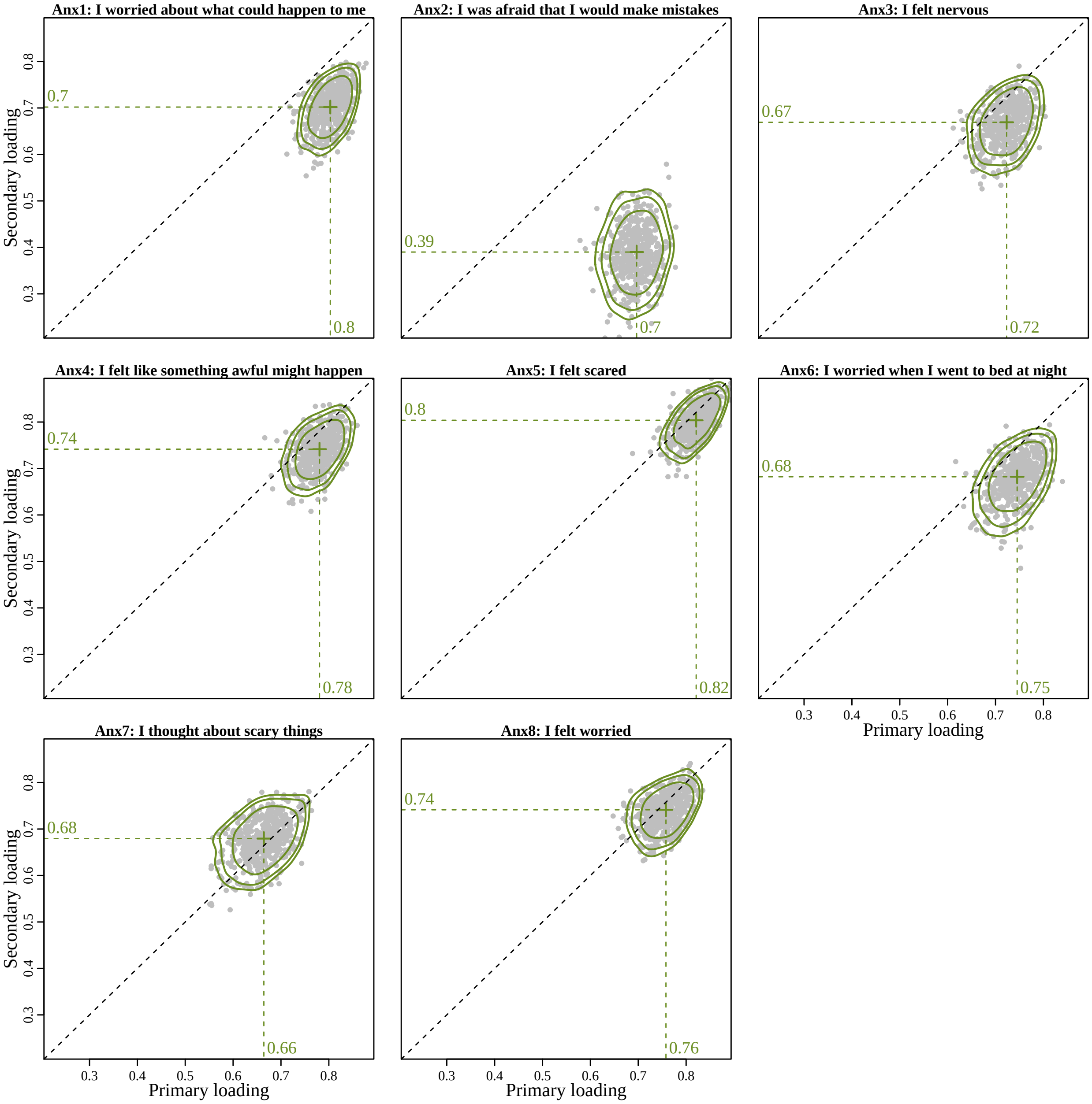

Figure 5.6: Two-dimensional confidence regions for the primary and secondary loadings of anxiety items. The points are 500 draws selected via a further thinning interval of 10. The three contours shown on each panel are the 75%, 90%, and 95% highest fiducial density regions. The fiducial medians are highlighted by green crosses, and the numerical values are also shown in green text. The diagonal dashed line indicates that an item contributes evenly to the primary and secondary factors.

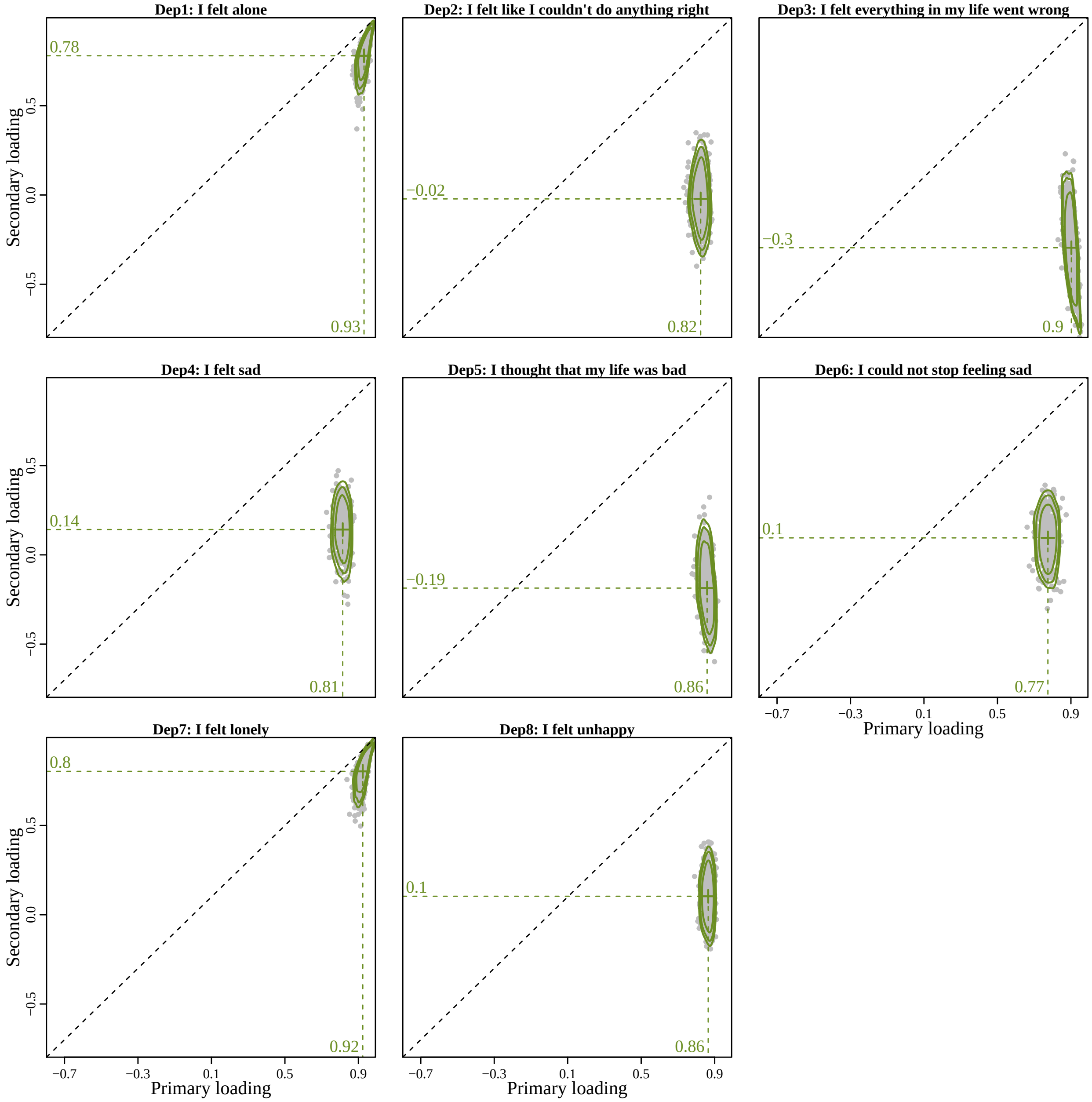

Figure 5.7: Two-dimensional confidence regions for the primary and secondary loadings of depression items. The points are 500 draws selected via a further thinning interval of 10. The three contours shown on each panel are the 75%, 90%, and 95% highest fiducial density regions. The fiducial medians are highlighted by green crosses, and the numerical values are also shown in green text. The diagonal dashed line indicates that an item contributes evenly to the primary and secondary factors.

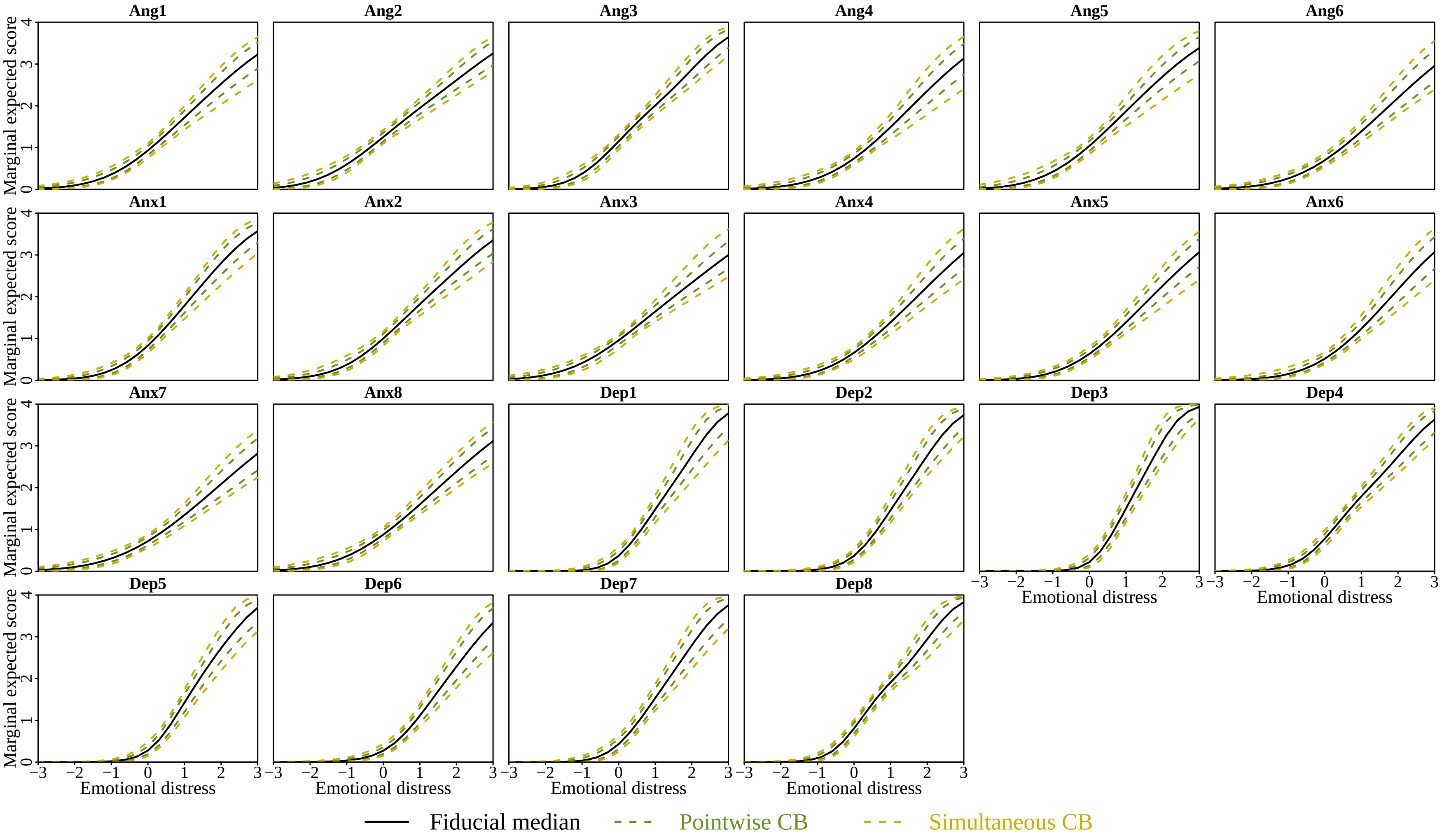

Figure 5.8: Fiducial median and 95% confidence bands (CBs) of the marginal expected score curve for general emotional distress. Pointwise and simultaneous CBs are shown in different colors.

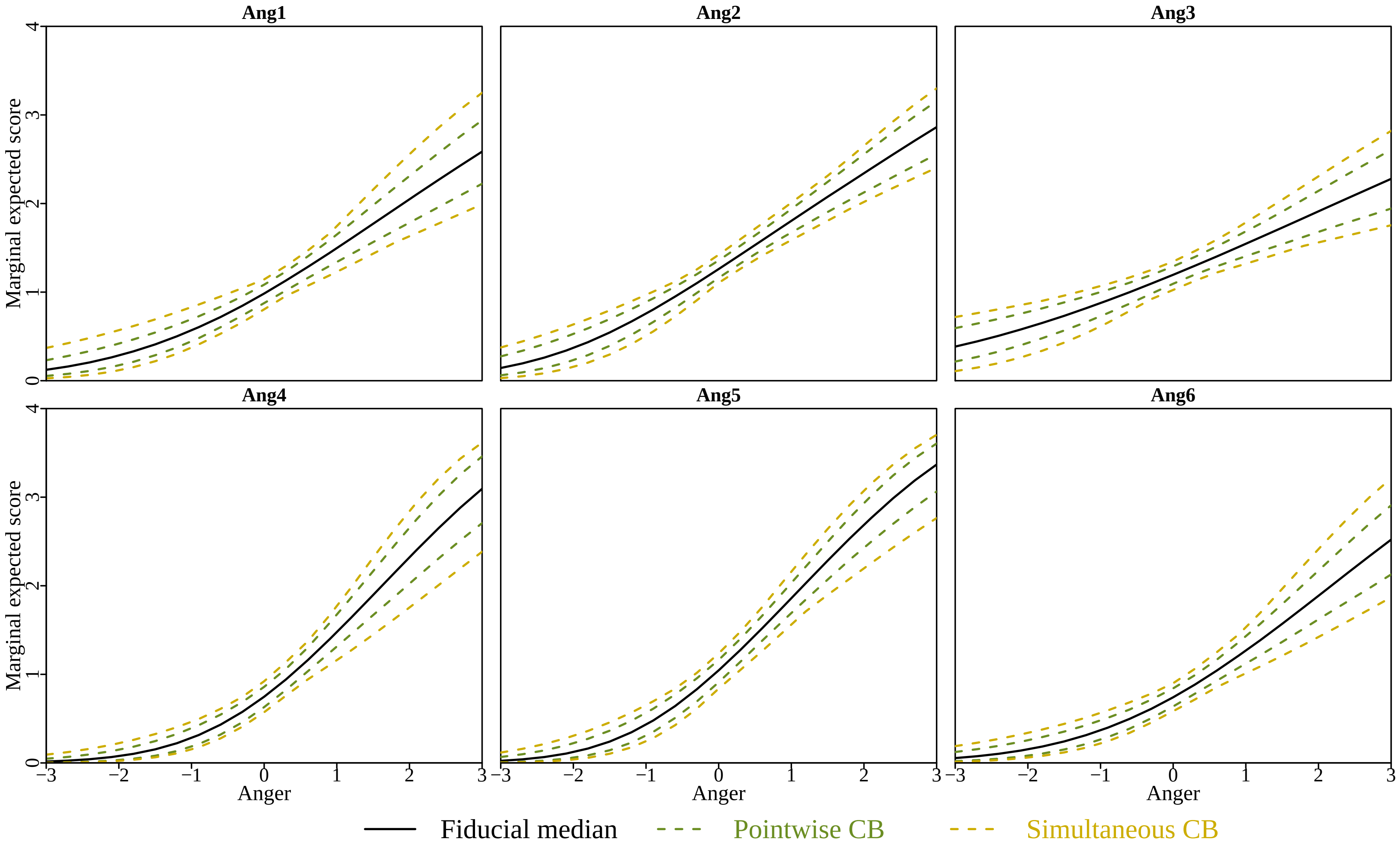

Figure 5.9: Fiducial medians and 95% confidence bands (CBs) of the marginal expected score curve for anger. Only items in the anger subscale are shown here. Pointwise and simultaneous CBs are shown in different colors.

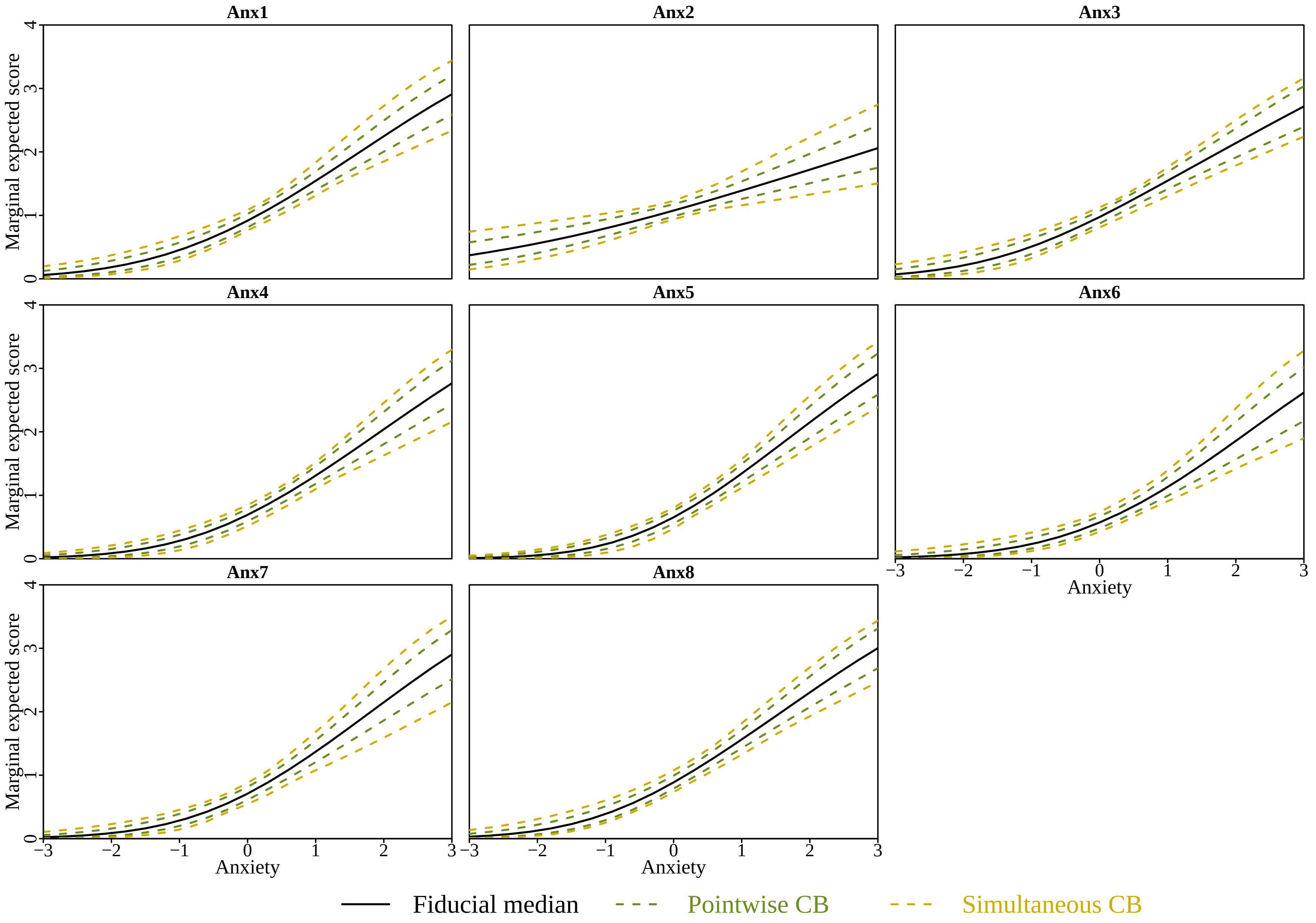

Figure 5.10: Fiducial medians and 95% confidence bands (CBs) of the marginal expected score curve for anxiety. Only items in the anxiety subscale are shown here. Pointwise and simultaneous CBs are shown in different colors.

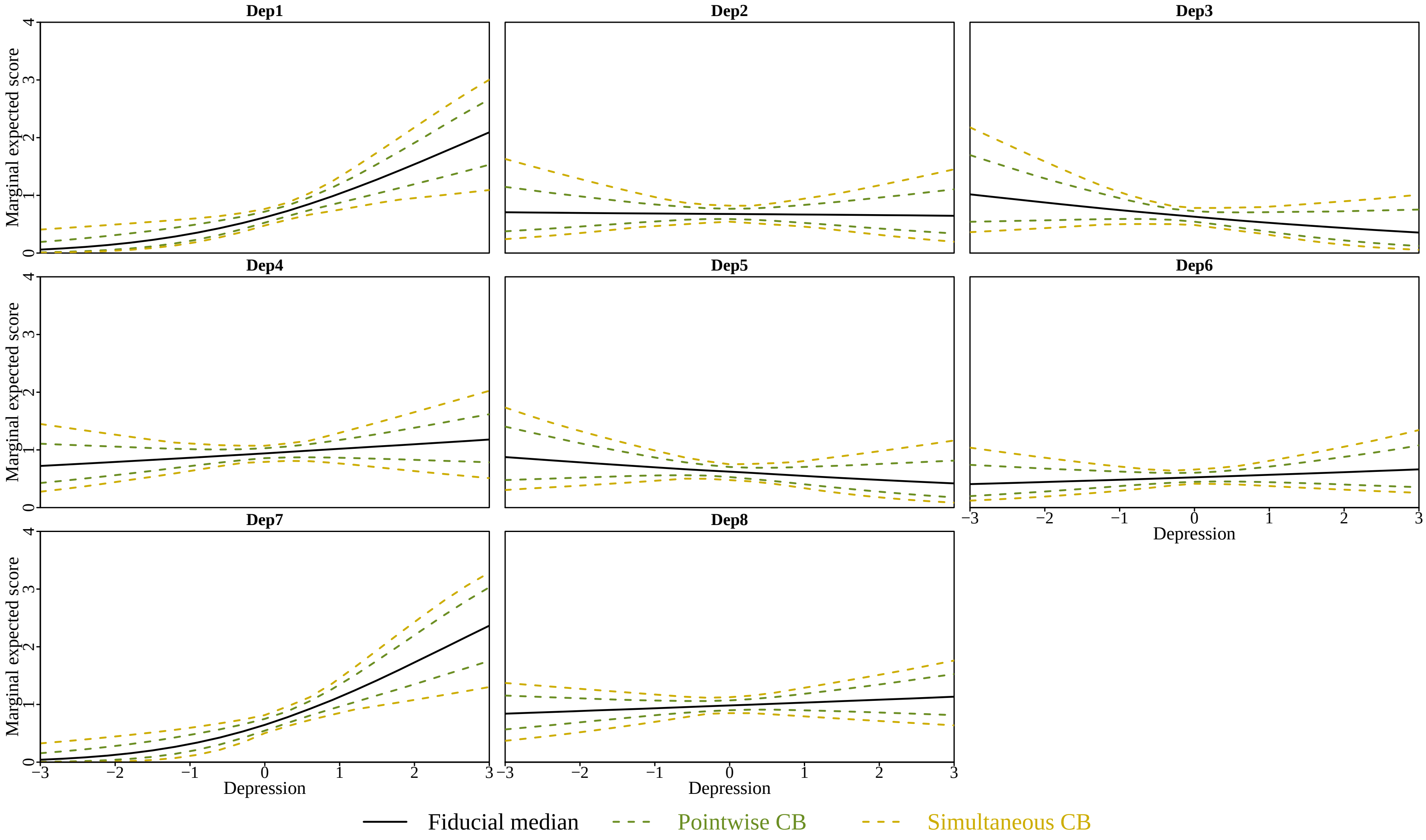

Figure 5.11: Fiducial medians and 95% confidence bands (CBs) of the marginal expected score curve for depression. Only items in the depression subscale are shown here. Pointwise and simultaneous CBs are shown in different colors.

Finally, we study the impact of sampling error in reliability analysis. Various methods have been proposed to quantify the reliability of a scale in the context of item response theory (IRT). A popular method is to compute the Fisher information matrix with respect to the latent variables $\mathbf{z}_i$ while treating the item parameters $\boldsymbol{\theta}$ as known; for easy reference, we denote it by $\boldsymbol{\mathcal{J}}(\mathbf{z}_i, \boldsymbol{\theta})$. Also let $f(\boldsymbol{\theta}, \mathbf{y}_i|\mathbf{z}_i) = \prod_{j=1}^{22} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i)$ be the conditional likelihood of an individual response pattern $\mathbf{y}_i$. It can be verified by direct calculation that

$$\begin{aligned}\boldsymbol{\mathcal{J}}(\mathbf{z}_i, \boldsymbol{\theta}) =& - E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \mathbf{z}_i \partial {\mathbf{z}_i}^\top} \log f(\boldsymbol{\theta}, \mathbf{y}_i|\mathbf{z}_i)\right]\\ =& \sum_{j=1}^{22}\sum_{k=0}^{4} \frac{1}{f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i)}\left\{\frac{e^{\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i)}}{[1+e^{\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{z}_i)}]^2} - \frac{e^{\tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}_i)}}{[1+e^{\tau_{j,k+1}(\boldsymbol{\theta}_j, \mathbf{z}_i)}]^2}\right\}^2 \boldsymbol{\beta}_j {\boldsymbol{\beta}_j}^\top.\end{aligned} \tag{5.3}$$

The inverse of $\boldsymbol{\mathcal{J}}(\mathbf{z}_i, \boldsymbol{\theta})$ gives the asymptotic covariance matrix for the ML estimates of the latent variables. Most often in practice, however, item parameters need to be calibrated, and thus the plug-in versions of information/asymptotic covariance matrix evaluated at the point estimates are subject to sampling variability.

For the four-dimensional model being fitted here, the 4 × 4 asymptotic covariance matrix $\boldsymbol{\mathcal{J}}(\mathbf{z}_i, \boldsymbol{\theta})$ is defined for each four-dimensional vector $\mathbf{z}_i \in \mathbb{R}^4$. For each dimension d, we define the marginal standard error function as following:

$$\upsilon_d(z_{id}, \boldsymbol{\theta}) = \sqrt{\int \sigma_d^2(\mathbf{z}_i, \boldsymbol{\theta}) d\Phi(\mathbf{z}_{i,-d})}, \tag{5.4}$$

in which $\sigma_d^2(\mathbf{z}_i, \boldsymbol{\theta})$ is the dth diagonal element of $\boldsymbol{\mathcal{J}}(\mathbf{z}_i, \boldsymbol{\theta})^{-1}$, and the integral is taken with respect to the remaining three dimensions. $\upsilon_d(z_{id}, \boldsymbol{\theta})$ gauges the average precision of the scale at each level z_{id} of a particular dimension d. Again, numerical integration is needed in the actual computation of Equation 5.4; because the previous dimension reduction trick is no longer applicable, we reduce the number of quadrature points to 11 on each dimension. The fiducial median and 95% pointwise CB for the marginal standard error functions are shown in Figure 5.12. In principle, simultaneous CBs can be obtained in a fashion similar to what we have done for expected score curves; however, the total number of item parameters involved in the calculation of equations 5.3 and 5.4 are $22 \times 6 = 132$, and we are not aware of any existing software that is able to handle such a high-dimensional kernel density estimation problem.

We observe in Figure 5.12 that the depression-specific latent variable is poorly measured at all levels; this is anticipated because the dimension is only effectively indicated by the "longly/alone" pair. The general dimension is often more precisely assessed compared to the domain-specific dimensions, for the reason that all 22 items provide information about individual differences in emotional distress symptoms. Because all threshold parameters are skewed to the high end of the latent variable scale, the measurement error at negative z_{id} levels are high.

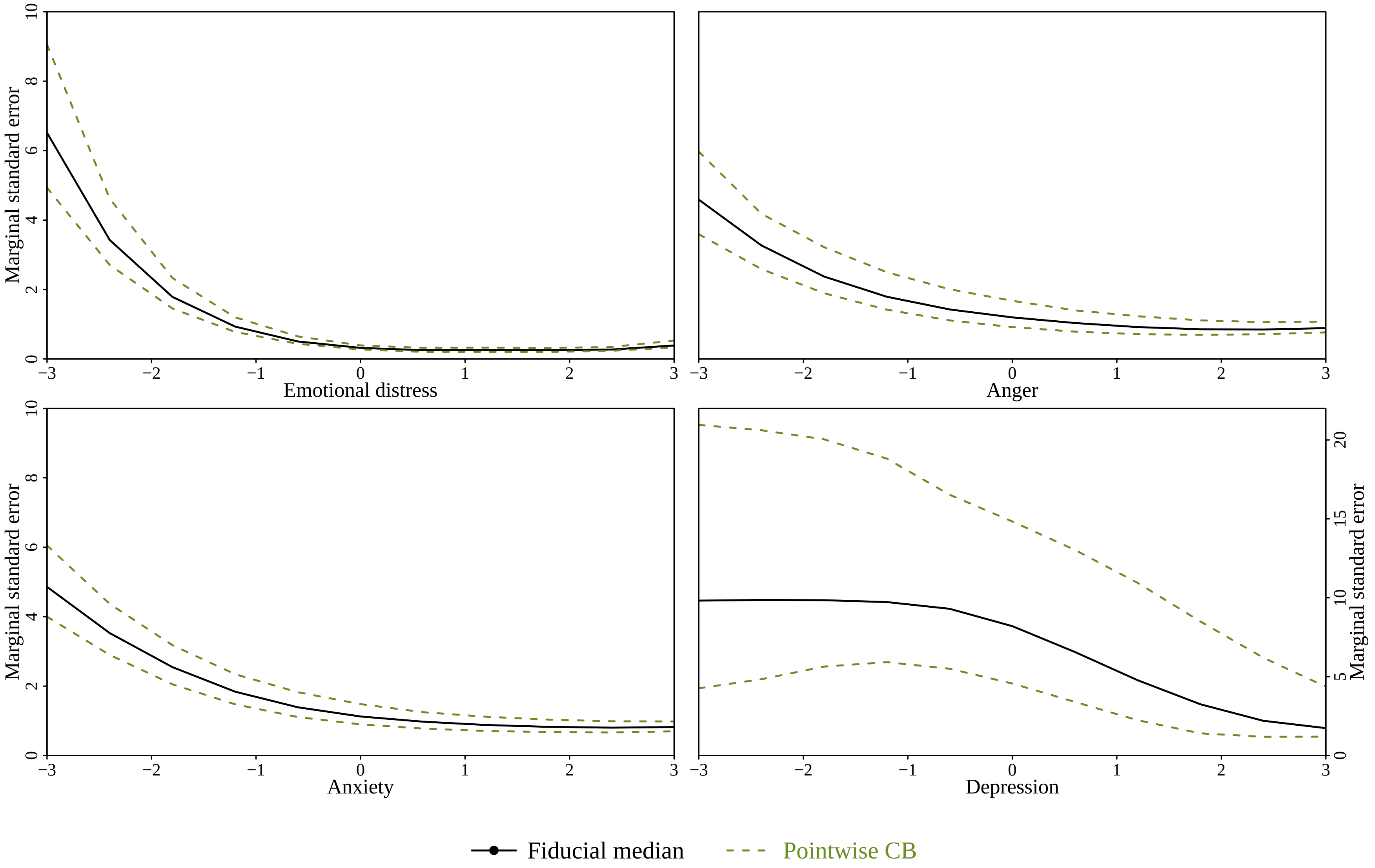

Figure 5.12: The fiducial median and 95% confidence bands (CBs) for marginal standard error curves for the four dimensions. Pointwise CBs are shown in colored dashed lines. Note that the lower-right panel for depression has a different y-axis from the rest.

5.4 Summary

This analysis of the PROMIS emotional distress data shows that GFI offers a powerful and omnibus toolkit to address various inferential problems based on the GRM. Making use of the Monte Carlo sample generated by the proposed Gibbs sampling algorithm, we are able to check the compatibility of the model to the data, construct CIs for implicitly defined functions of model parameters, and draw CBs for functional transformations. Some of the analyses are novel: for example, calculating simultaneous CBs for item expected score curves (each of which depends on six item parameters) and CBs for marginal standard errors functions of the test (which depends on all 132 item parameters). The same level of flexibility can be achieved by adopting the full Bayesian framework, but at the cost of prior specification. For the PROMIS emotional distress domains, no *a priori* information about model parameters is available, making it difficult to specify a reasonable subjective prior. Moreover, the calibration sample is not large enough. From our experience in the reported simulation study (Chapter 4), the small-sample performance of non-informative Bayesian methods is inferior to GFI.

There are several limitations in the presented analyses.

First, we did not correct for the false discovery rate (FDR; Benjamini and Hochberg, 1995; Thissen, Steinberg, and Kuang, 2002) when conducting multiple hypothesis tests, nor the false coverage-statement rate (FCR; Benjamini and Yekutieli, 2005) when reporting multiple CIs/CBs. Our goal has been to illustrate the use of GFI instead of studying the scale itself, and thus we do not feel pressed to do so. Furthermore, all the p-values and confidence bounds are computed empirically based on only 5000 draws, which is too few to control the Monte Carlo error after applying the adjustment. For example, if we consider 100 tests, then the smallest p-value will be compared to $0.05/100 = 0.0005$; the estimated 0.0005 quantile of a 5000 Monte Carlo sample is almost the minimum. In applied research, we do suggest correcting for FDR/FCR in order to justify the validity of the results. If necessary, a longer Markov chain should be obtained to reduce the Monte Carlo error in approximating extreme quantiles of the fiducial distribution.

Second, the R package `ks` for non-parametric estimation of multidimensional density functions can only handle up to six-dimensional data, which barely suffices in the simultaneous CB calculation for marginal expected score curves in a bifactor model. Apart from the dimensionality limit, it calls R's optimizer `nlm()` when determining the smoothing bandwidth, which lessens its efficiency[2]. It is necessary to write a separate program in higher level computer languages such as Fortran or C, in order to make the proposed method applicable in larger-scale problems.

Finally, our goodness of fit tests based on the sum-score profile fail to detect the misfit of a unidimensional model, whereas the bivariate tests clearly indicate an underlying three-dimensional structure. Li and Cai (2012) remarked based on their simulation results that misspecified latent dimensionality does not damage the fit to sum-score profile so much as it does to bivariate margins. They revealed that an overall fit statistic summarizing the residuals in bivariate subtable cells has a much higher power than another statistic summarizing the residuals at sum-score levels, whereas the comparative power is reversed when the generating model remains unidimensional but has a misspecified distribution. Both the simulations reported in Chapter 4 and the current data example seem to confirm their findings. We acknowledge that perhaps there is no single diagnostic that is able to identify all types of discrepancies between the true and fitted models, because different diagnostics combine information in different manners. Practically, we recommend choosing test statistics in line with *a priori* information of plausible misfitting mechanisms. As a topic for future research, fit diagnostics that are sensitive to alternative data generating models should be developed and incorporated into FPC.

---

[2] There is an option to switch to `optim()`, but there is no speed gain from our experience.

CHAPTER 6: DISCUSSION AND CONCLUSION

We have derived generalized fiducial inference (GFI) for a family of multidimensional graded response models (GRM). It has been rigorously established that GFI for the GRM yields asymptotically correct inference in the frequentist sense, equivalent to likelihood-based and Bayesian methods that have been extensively studied in the literature (Chapter 2). Furthermore, we have shown by Monte Carlo simulations (Chapter 4) that GFI using the proposed Gibbs sampler (Chapter 3) is reliable for parameter recovery, scoring, and goodness of fit testing, even in situations when the sample size is too small and/or the data generating parameters are too extreme for likelihood-based and Bayesian counterparts to behave well. The usefulness and flexibility of the proposed method have been illustrated with an empirical example (Chapter 5). We conclude that GFI is a preferred inferential framework for calibrating ordinal items in small samples, and that sampling variability, which is a more salient issue in small-sample data analyses, can be gauged easily with a Monte Carlo approximation of the fiducial distribution.

There are several remaining challenges to be addressed by future research.

First, theoretical interpretations of the superiority of GFI over normal approximation in small samples should be sought. It has been conjectured that GFI has more favorable higher-order asymptotic properties. Some preliminary investigation of a higher-order expansion of the fiducial distribution function can be found in Pal Majumder and Hannig (2015), in which an ingenious "shrinkage argument" (Ghosh and Bickel, 1990) was exploited to derive conditions under which the fiducial probability coincides with the coverage probability up to a certain order. Their results might not be directly applicable to the current GRM application; however, a similar theoretical justification may establish more solid grounds for the use of the current work in small samples.

Second, effort should be devoted to improve the efficiency of the sampling algorithm. We believe that a more efficient algorithm would facilitate the dissemination of the proposed method, which has been demonstrated to outperform existing ones under certain circumstances. The combinatorial and computational mathematics of the polygon cutting problem should be studied in order to alleviate the computational burden presented in the current implementation of the updating stage (Section 3.3 of Chapter 3). Alternative Monte Carlo methods such as sequential Monte Carlo (SMC; Doucet, De Freitas, and Gordon, 2001) has been successfully used for GFI in the context of other models (e.g., linear-normal mixed effect models; Cisewski and Hannig, 2012); as a non-iterative method, SMC may be more efficient than the Gibbs sampler.

Finally, the application of GFI to other popular item response models should be explored. In the current work, we only focus on a family of GRMs in which the covariance structure of the latent variables is known, e.g., exploratory models. In practice, however, simple structure models (also known as independent cluster models) and general two-tier models (i.e., replacing the general factor in a bifactor model by multiple factors with unconstrained covariance structure) might be preferred over exploratory models for estimation efficiency and ease of interpretation. In addition, it is also of interest to derive GFI for un-ordered polytomous item response models such as Bock's nominal response model (Bock, 1972), and models with categorical latent variables such as cognitive diagnostic models (Rupp, Templin, and Henson, 2010) and latent class models (Lazarsfeld and Henry, 1968). We have observed in our simulation study that GFI is more reliable than ML in the presence of model identification difficulties, which renders it a promising alternative for many psychometric models in which ML estimator has been known to be ill-behaved.

APPENDIX A: BASIC PROPERTIES

A.1 Calculating the fiducial density

Proof of Lemma 1. We start with the following fundamental probability calculation:

$$
\begin{aligned}
&P\left\{\mathbf{V} \leq \boldsymbol{\theta}, Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\right\} \\
=&\sum_{\mathcal{P}} P\left\{\mathbf{V} \leq \boldsymbol{\theta}, \bigcap_{(I',\mathbf{k}_{I'})\in\mathcal{P}} E_{I',\mathbf{k}_{I'}} \bigcap_{(I',\mathbf{k}_{I'})\notin\mathcal{P}} E^c_{I',\mathbf{k}_{I'}}\right\} \\
=&\sum_{\mathcal{P}} \sum_{(I,\mathbf{k}_I)\in\mathcal{P}} P\left\{\mathbf{V}_{I,\mathbf{k}_I} \leq \boldsymbol{\theta}, \bigcap_{(I',\mathbf{k}_{I'})\in\mathcal{P}} E_{I',\mathbf{k}_{I'}} \bigcap_{(I',\mathbf{k}_{I'})\notin\mathcal{P}} E^c_{I',\mathbf{k}_{I'}}, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I}\right\} \\
=&\sum_{I}\sum_{\mathbf{k}_I} P\left\{\mathbf{V}_{I,\mathbf{k}_I} \leq \boldsymbol{\theta}, E_{I,\mathbf{k}_I}, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I}\right\} \\
=&\sum_{I}\sum_{\mathbf{k}_I} P\left\{\mathbf{V}_{I,\mathbf{k}_I} \leq \boldsymbol{\theta}, E_{I,\mathbf{k}_I}\right\} P\left\{\mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} | E_{I,\mathbf{k}_I}\right\}. \qquad \text{(A.1)}
\end{aligned}
$$

The last line of Equation A.1 holds because selection rules are assumed to be independent of the logistic and normal variates, and subsequently independent of $\mathbf{V}_{I,\mathbf{k}_I}$. Equation 2.9 shows that $P\left\{\mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} | E_{I,\mathbf{k}_I}\right\} = w_{I,\mathbf{k}_I}(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$; therefore, the remaining task is to derive the expression of $P\left\{\mathbf{V}_{I,\mathbf{k}_I} \leq \boldsymbol{\theta}, E_{I,\mathbf{k}_I}\right\}$ and then differentiate with respect to $\boldsymbol{\theta}$.

Consider a single item first. When $\mathbf{V}_{I_j,\mathbf{k}_{I_j}} = \boldsymbol{\theta}'_j$, $E_{I_j,\mathbf{k}_{I_j}}$ means that for all $i \in I_j$, $\tau_{jk_{ij}}(\boldsymbol{\theta}'_j, \mathbf{z}_i) = A^\star_{ij}$, and that $\boldsymbol{\theta}'_j$ should not conflict with the half-spaces of the other observations: i.e., for all $i \in I^c_j$, $\tau_{j,y_{ij}+1}(\boldsymbol{\theta}'_j, \mathbf{z}_i) \leq A^\star_{ij} < \tau_{jy_{ij}}(\boldsymbol{\theta}'_j, \mathbf{z}_i)$. Thus, conditional on $\mathbf{Z}^\star = \mathbf{z}$, we have

$$
\begin{aligned}
&P\left\{\mathbf{V}_{I_j,\mathbf{k}_{I_j}} \leq \boldsymbol{\theta}_j, E_{I_j,\mathbf{k}_{I_j}} \mid \mathbf{Z}^\star = \mathbf{z}\right\} \\
=&\int_{\boldsymbol{\theta}'_j \leq \boldsymbol{\theta}_j} d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}'_j, \mathbf{z}_{I_j}) \prod_{i\in I_j} \frac{e^{\tau_{jk_{ij}}(\boldsymbol{\theta}'_j,\mathbf{z}_i)}}{[1+e^{\tau_{jk_{ij}}(\boldsymbol{\theta}'_j,\mathbf{z}_i)}]^2} \prod_{i\in I^c_j} f_j(\boldsymbol{\theta}'_j, y_{ij}|\mathbf{z}_i) d\boldsymbol{\theta}'_j, \qquad \text{(A.2)}
\end{aligned}
$$

in which the determinant and the first product are due to the change of variables from $(A^\star_{ij})_{i\in I_j}$ to $\mathbf{V}_{I_j,\mathbf{k}_{I_j}}$ (the standard logistic density $\psi(x) = e^x/(1+e^x)^2$), and the second product corresponds to the logistic probabilities of those inequalities that the other observations should satisfy.

Due to the conditional independence assumption,

$$\begin{aligned}
&P\left\{\mathbf{V}_{I,\mathbf{k}_I} \leq \boldsymbol{\theta}, E_{I,\mathbf{k}_I}\right\} \\
=&\int_{\mathbb{R}^{nr}} \prod_{j=1}^{m} P\left\{\mathbf{V}_{I_j,\mathbf{k}_{I_j}} \leq \boldsymbol{\theta}_j, E_{I_j,\mathbf{k}_{I_j}} \mid \mathbf{Z}^\star = \mathbf{z}\right\} d\Phi(\mathbf{z}) \\
=&\int_{\mathbb{R}^{nr}} \int_{\boldsymbol{\theta}' \leq \boldsymbol{\theta}} \prod_{j=1}^{m} \left\{ d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}'_j, \mathbf{z}_{I_j}) \frac{e^{\tau_{jk_{ij}}(\boldsymbol{\theta}'_j,\mathbf{z}_i)}}{[1+e^{\tau_{jk_{ij}}(\boldsymbol{\theta}'_j,\mathbf{z}_i)}]^2} \prod_{i\in I_j^c} f_j(\boldsymbol{\theta}'_j, y_{ij}|\mathbf{z}_i) \right\} d\boldsymbol{\theta}' d\Phi(\mathbf{z}).
\end{aligned} \tag{A.3}$$

Equation 2.8 is established by substituting Equation A.3 back into Equation A.1, switching the order of integration, and differentiating with respective to $\boldsymbol{\theta}$. □

A.2 The invariance property

Proof of Proposition 1. Let $a_n(\boldsymbol{\theta}, \mathbf{y})$ be the right-hand side of equation 2.8, and similarly $\tilde{a}_n(\boldsymbol{\xi}, \mathbf{y})$ be the unnormalized version of $\tilde{g}_n(\boldsymbol{\xi}|\mathbf{y})$. It suffices to show

$$\int_B a_n(\boldsymbol{\theta}, \mathbf{y}) d\boldsymbol{\theta} = \int_{\mathbf{q}^{-1}(B)} \tilde{a}_n(\boldsymbol{\xi}, \mathbf{y}) d\boldsymbol{\xi}. \tag{A.4}$$

for every measurable set B on the parameter space. Let $\boldsymbol{\tau}_{I,\mathbf{k}_I}(\boldsymbol{\theta}, \mathbf{z}_I) = (\tau_{jk_{ij}}(\boldsymbol{\theta}_j, \mathbf{z}_i))_{i\in I_j, j=1,\ldots,m}$. Then the Jacobian determinant $|\det(\partial\boldsymbol{\tau}_{I,\mathbf{k}_I}(\boldsymbol{\theta}, \mathbf{z}_I)/\partial\boldsymbol{\theta})| = \prod_{j=1}^m d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}_j, \mathbf{z}_{I_j})$. Also write

$$R_{I,k_I}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}) = \prod_{j=1}^{m} \left\{ \prod_{i\in I_j} \frac{e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}}{[1+e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}]^2} \prod_{i\in I_j^c} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) \right\}. \tag{A.5}$$

By our differentiability assumption and the multivariate chain rule,

$$\left|\det\left(\frac{\partial\boldsymbol{\tau}_{I,\mathbf{k}_I}(\mathbf{q}(\boldsymbol{\xi}), \mathbf{z}_I)}{\partial\boldsymbol{\xi}}\right)\right| = \left|\det\left(\frac{\partial\boldsymbol{\tau}_{I,\mathbf{k}_I}(\mathbf{q}(\boldsymbol{\xi}), \mathbf{z}_I)}{\partial\mathbf{q}(\boldsymbol{\xi})}\right)\right| \cdot \left|\det\left(\frac{\partial\mathbf{q}(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}}\right)\right|. \tag{A.6}$$

Hence, we have

$$
\begin{aligned}
&\int_B a_n(\boldsymbol{\theta}, \mathbf{y}) d\boldsymbol{\theta} \\
=&\sum_I \sum_{\mathbf{k}_I} w_{I,\mathbf{k}_I}(\mathbf{y}) \int_{\mathbb{R}^{nr}} \int_B \left| \det \left(\frac{\partial \boldsymbol{\tau}_{I,\mathbf{k}_I}(\boldsymbol{\theta}, \mathbf{z}_I)}{\partial \boldsymbol{\theta}} \right) \right| R_{I,\mathbf{k}_I}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}) d\boldsymbol{\theta} d\Phi(\mathbf{z}) \\
=&\sum_I \sum_{\mathbf{k}_I} w_{I,\mathbf{k}_I}(\mathbf{y}) \int_{\mathbb{R}^{nr}} \int_{\mathbf{q}^{-1}(B)} \left| \det \left(\frac{\partial \boldsymbol{\tau}_{I,\mathbf{k}_I}(\mathbf{q}(\boldsymbol{\xi}), \mathbf{z}_I)}{\partial \mathbf{q}(\boldsymbol{\xi})} \right) \right| R_{I,\mathbf{k}_I}(\mathbf{q}(\boldsymbol{\xi}), \mathbf{y}, \mathbf{z}) \left| \det \left(\frac{\partial \mathbf{q}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \right) \right| d\boldsymbol{\xi} d\Phi(\mathbf{z}) \\
=&\sum_I \sum_{\mathbf{k}_I} w_{I,\mathbf{k}_I}(\mathbf{y}) \int_{\mathbb{R}^{nr}} \int_{\mathbf{q}^{-1}(B)} \left| \det \left(\frac{\partial \boldsymbol{\tau}_{I,\mathbf{k}_I}(\mathbf{q}(\boldsymbol{\xi}), \mathbf{z}_I)}{\partial \boldsymbol{\xi}} \right) \right| R_{I,\mathbf{k}_I}(\mathbf{q}(\boldsymbol{\xi}), \mathbf{y}, \mathbf{z}) d\boldsymbol{\xi} d\Phi(\mathbf{z}) \\
=&\int_{\mathbf{q}^{-1}(B)} \tilde{a}_n(\boldsymbol{\xi}, \mathbf{y}) d\boldsymbol{\xi}. \qquad \text{(A.7)}
\end{aligned}
$$

□

APPENDIX B: A BERNSTEIN-VON MISES THEOREM

Proof of Theorem 1. We start from re-expressing the fiducial density in Equation 2.10. Recall that we require $w_{I,\mathbf{k}_I}(\mathbf{y}) = w_{I',\mathbf{k}_{I'}}(\mathbf{y}) = w_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{y})$ whenever $\mathbf{y}_I = \mathbf{y}_{I'}$ and $\mathbf{k}_I = \mathbf{k}_{I'}$; therefore, the key observation is that the (outer) sum over index sets I in Equation 2.10 can be reduced to a finite sum over sub-sample response patterns $\mathbf{y}_I$. Recall that $I = \bigcup_{j=1}^m I_j$, which has at least $\max_j q_j$ and at most $\sum_{j=1}^m q_j$ elements. Let $G_n = \binom{n}{\sum_{j=1}^m q_j}$ be the total number of size-$\sum_{j=1}^m q_j$ sub-samples. Also let $p_n(\mathbf{y}_I) = G_n^{-1}\sum_I \mathbb{I}\{\mathbf{Y}_I = \mathbf{y}_I\}$. By the standard theory of U-statistics, $p_n(\mathbf{y}_I) \overset{P_{\boldsymbol{\theta}_0}}{\to} \pi_0(\mathbf{y}_I)$, in which $\pi_0(\mathbf{y}_I)$ is determined by the data-generating parameter values $\boldsymbol{\theta}_0$, and $\pi_0(\mathbf{y}_I) = 0$ if $|I| < \sum_{j=1}^m q_j$. Then the fiducial density (Equation 2.8) can be written as

$$g_n(\boldsymbol{\theta}|\mathbf{y}) \propto b_n(\boldsymbol{\theta}, \mathbf{y}) f_n(\boldsymbol{\theta}, \mathbf{y}), \tag{B.1}$$

In Equation B.1, $f_n(\boldsymbol{\theta}, \mathbf{y})$ is the sample likelihood, and

$$b_n(\boldsymbol{\theta}, \mathbf{y}) = G_n \sum_{\mathbf{y}_I} p_n(\mathbf{y}_I) \left[\sum_{\mathbf{k}_I} w_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{y}) b_{\mathbf{y}_I,\mathbf{k}_I}(\boldsymbol{\theta})\right], \tag{B.2}$$

in which

$$\begin{aligned} b_{\mathbf{y}_I,\mathbf{k}_I}(\boldsymbol{\theta}) = &\int \prod_{j=1}^m d_{I_j,\mathbf{k}_{I_j}}(\boldsymbol{\theta}_j, \mathbf{z}_{I_j}) \\ &\cdot \prod_{i\in I} \left\{ \prod_{j\in J(i)} \frac{e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}}{[1+e^{\tau_{jk_{ij}}(\boldsymbol{\theta}_j,\mathbf{z}_i)}]^2} \prod_{j\notin J(i)} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) \right\} d\Phi(\mathbf{z}_I) \\ &\Big/ \int \prod_{i\in I}\prod_{j=1}^m f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}_I). \end{aligned} \tag{B.3}$$

Equation B.2 is a repetition of Equation 2.13. Also let

$$a_n(\boldsymbol{\theta}, \mathbf{y}) = f_n(\boldsymbol{\theta}, \mathbf{y}) b_n(\boldsymbol{\theta}, \mathbf{y}) \tag{B.4}$$

be the right-hand side (RHS) of Equation B.1.

Next, we consider the local parameter $\mathbf{h} = \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. Some short-hand notation is introduced for conciseness: Let $b_{n,\mathbf{h}} = b_n(\boldsymbol{\theta}_0 + \mathbf{h}/\sqrt{n}, \mathbf{y})/G_n$, $a_{n,\mathbf{h}} = a_n(\boldsymbol{\theta}_0 + \mathbf{h}/\sqrt{n}, \mathbf{y})$, and $f_{n,\mathbf{h}} = f_n(\boldsymbol{\theta}_0 + \mathbf{h}/\sqrt{n}, \mathbf{y})/\sqrt{n}$; also let $b_{n,0} = \sum_{\mathbf{y}_I} \pi_0(\mathbf{y}_I)[\sum_{\mathbf{k}_I} w_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{y}) b_{\mathbf{y}_I,\mathbf{k}_I}(\boldsymbol{\theta}_0)]$, $a_{n,0} = a_n(\boldsymbol{\theta}_0, \mathbf{y})$, and $f_{n,0} = f_n(\boldsymbol{\theta}_0, \mathbf{y})$. Using this new notation, the fiducial density of the local parameter can be written as

$$\bar{g}_n(\mathbf{h}|\mathbf{y}) \propto a_{n,\mathbf{h}} = b_{n,\mathbf{h}} f_{n,\mathbf{h}}. \tag{B.5}$$

For each $\mathbf{y}_I$ and $\mathbf{k}_I$, $b_{\mathbf{y}_I,\mathbf{k}_I}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ (it is in fact differentiable). In addition, we know that $p_n(\mathbf{y}_I) \to \pi_0(\mathbf{y}_I)$ in $P_{\boldsymbol{\theta}_0}$-probability, and that $w_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{y})$ is bounded. Consequently, $|b_{n,\mathbf{h}} - b_{n,0}|$ converges to 0 in $P_{\boldsymbol{\theta}_0}$-probability.

We also consider the Taylor series expansion of $\log f_{n,\mathbf{h}}$ at the true parameter $\boldsymbol{\theta}_0$:

$$\log \frac{f_{n,\mathbf{h}}}{f_{n,0}} = \mathbf{h}^\top \mathbf{S}_n + \frac{1}{2n} \sum_{i=1}^{n} \mathbf{h}^\top \mathbf{H}(\boldsymbol{\theta}_0, \mathbf{y}_i)\mathbf{h} + R_{n,\mathbf{h}}. \tag{B.6}$$

Here, some comments are made for each term of Equation B.6. a) The sequence $\{\mathbf{S}_n\}$ is tight by the convergence result given by Equation 2.16; hence, for each $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset \mathbb{R}^q$ such that $P(K_\varepsilon) > 1 - \varepsilon$ and $\mathbf{S}_n \in K_\varepsilon$ for all n. If we restrict the consideration to K_ε, then the first term of Equation B.6 is bounded for each $\mathbf{h}$. b) By the (Uniform) Law of Large Numbers, the second term converges to $\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}$ in probability (the convergence is uniform for $\mathbf{h}$ in compact sets). c) The remainder term has the following form:

$$R_{n,\mathbf{h}} = \sum_{i=1}^{n} \sum_{|\mathbf{t}|=3} \frac{f^{(\mathbf{t})}(\bar{\boldsymbol{\theta}}, \mathbf{Y}_i)}{\mathbf{t}!} \left(\frac{\mathbf{h}}{\sqrt{n}}\right)^{\mathbf{t}}. \tag{B.7}$$

In Equation B.7, $\mathbf{t} = (t_1, \ldots, t_q)$ is a q-tuple of nonnegative integers serving as a multi-index: $|\mathbf{t}| = \sum_{s=1}^{q} t_s$, $\mathbf{h}^{\mathbf{t}} = h_1^{t_1} \cdots h_q^{t_q}$, $\mathbf{t}! = \frac{q!}{t_1! \cdots t_q!}$, and $f^{(\mathbf{t})} = \frac{\partial^{|\mathbf{t}|} f}{\partial^{t_1}\theta_1 \cdots \partial^{t_q}\theta_q}$, where $h_1, \ldots, h_q$ and $\theta_1, \ldots, \theta_q$ are the coordinates of $\mathbf{h}$ and $\boldsymbol{\theta}$, respectively. $\bar{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + \mathbf{h}/\sqrt{n}$.

Now we proceed to the proof of Theorem 1, i.e., Equation 2.17. By an argument similar to Ghosh and Ramamoorthi (2003), it suffices to show for each $\varepsilon > 0$ that

$$\int_{H_n} \left| \frac{a_{n,\mathbf{h}}}{f_{n,0}} - b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} \right| d\mathbf{h} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0 \tag{B.8}$$

To see this, let $D_n = \int_{H_n} a_{n,\mathbf{h}} d\mathbf{h} / f_{n,0}$. The left-hand side (LHS) of Equation 2.17 can be bounded by

$$D_n^{-1} \int_{H_n} \left| \frac{a_{n,\mathbf{h}}}{f_{n,0}} - b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} \right| d\mathbf{h} + \int_{H_n} \left| D_n^{-1} b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} - \phi_{\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{S}_n, \boldsymbol{\mathcal{I}}_0^{-1}}(\mathbf{h}) \right| d\mathbf{h}. \tag{B.9}$$

Notice that Equation B.8 implies $|D_n - b_{n,0} \int_{H_n} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h}| \overset{P_{\boldsymbol{\theta}_0}}{\to} 0$. We also know that

$$D e^{\frac{1}{2}\mathbf{S}_n{}^\top \boldsymbol{\mathcal{I}}_0^{-1} \mathbf{S}_n} \le \int_{H_n} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} \le D' e^{\frac{1}{2}\mathbf{S}_n{}^\top \boldsymbol{\mathcal{I}}_0^{-1} \mathbf{S}_n}, \tag{B.10}$$

for some suitable constants D and D', because the local parameter space H_n satisfies $\Theta - \boldsymbol{\theta}_0 \subset H_n \subset \mathbb{R}^q$. It follows that D_n^{-1} is $O_p(1)$, and that the first integral in Equation B.9 converges to zero in probability. Further let $T_{1,n}$ be the integral in Equation B.10, and $T_{2,n} = |D_n^{-1} b_{n,0} - T_{1,n}^{-1}|$; then, the second integral of Equation B.9 can be written as $T_{1,n} T_{2,n}$. The sequence $\{T_{1,n}\}$ is tight by Equation B.10, so for each $\eta > 0$ there exists an L_η such that $P(T_{1,n} \le L_\eta) > 1 - \eta$ for all n. Moreover, $T_{2,n} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0$ by Equation B.8. Fix $\varepsilon, \eta > 0$, we have

$$P(T_{1,n} T_{2,n} > \varepsilon) \le P(T_{1,n} T_{2,n} > \varepsilon, T_{1,n} \le L_\eta) + P(T_{1,n} > L_\eta) \le P(T_{2,n} > \varepsilon / L_\eta) + \eta, \tag{B.11}$$

which can be made less than 2η for large enough n. Therefore, $T_{1,n} T_{2,n} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0$. Because both integrals in Equation B.9 converge to 0 in probability, Equation 2.17 is established.

For the remaining part of the proof, we partition the domain of integration of Equation 2.17 into four regions (for n large enough), and establish the desired convergence on each part. The four regions are:

$$A_{1,n} = \{\mathbf{h} : \|\mathbf{h}\| < B \log n\} \cap H_n, \text{ for some large number } B > 0;$$

$$A_{2,n} = \{\mathbf{h} : B \log n \leq \|\mathbf{h}\| < \delta\sqrt{n}\} \cap H_n, \text{ for some small number } \delta > 0;$$

$$A_{3,n} = \{\mathbf{h} : \delta\sqrt{n} \leq \|\mathbf{h}\| \leq B'\sqrt{n}\} \cap H_n, \text{ for another large number } B' > 0;$$

$$A_{4,n} = \{\mathbf{h} : \|\mathbf{h}\| > B'\sqrt{n}\} \cap H_n.$$

In terms of the constants, we first choose δ and B to ensure the convergence on $A_{2,n}$. The convergence on $A_{1,n}$ holds for any $B > 0$, so it also holds for the particular B that we select. Then we consider region $A_{4,n}$ and select B'. Finally we show that the integral convergences for $\mathbf{h}/\sqrt{n}$ in any compact sets excluding $\mathbf{0}$, from which the convergence on $A_{3,n}$ follows.

Region $A_{2,n}$ Because the likelihood function is three times continuously differentiable with respect to $\boldsymbol{\theta}$, and also because there are finitely many (i.e., $\prod_{j=1}^{m} K_j$) individual patterns of $\mathbf{y}_i$, the remainder term (Equation B.7) of the Taylor expansion (Equation B.6) has the following bound for each $\delta > 0$ and $\|\mathbf{h}\| \leq \delta\sqrt{n}$:

$$|R_{n,\mathbf{h}}| \leq M(\delta)\frac{\|\mathbf{h}\|^3}{n^{3/2}} \leq M(\delta)\delta^3, \tag{B.12}$$

as a result of the multinomial theorem and the Cauchy-Schwarz inequality, in which $M(\delta)$ is a constant multiple of $|\max_{|\mathbf{t}|=3,\mathbf{y}_i} \sup_{\|\boldsymbol{\theta}-\theta_0\|\leq\delta} f^{(\mathbf{t})}(\boldsymbol{\theta}, \mathbf{y}_i)|$. Since $M(\delta) \downarrow$ as $\delta \downarrow 0$, Equation B.12 allows us to choose δ small enough such that $|R_{n,\mathbf{h}}| < \frac{1}{4}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}$ for all $\mathbf{h} \in A_{2,n}$. Then for such δ,

$$\begin{aligned}
&\int_{A_{2,n}} \left| \frac{a_{n,\mathbf{h}}}{f_{n,0}} - b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} \right| d\mathbf{h} \\
\leq &\int_{A_{2,n}} \frac{a_{n,\mathbf{h}}}{f_{n,0}} d\mathbf{h} + \int_{A_{2,n}} b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} \\
\leq &\sup_{\mathbf{h}\in A_{2,n}} b_{n,\mathbf{h}} \int_{A_{2,n}} \frac{f_{n,\mathbf{h}}}{f_{n,0}} d\mathbf{h} + b_{n,0} \int_{A_{2,n}} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} \\
\leq &\left(\sup_{\mathbf{h}\in A_{2,n}} b_{n,\mathbf{h}} + b_{n,0} \right) \int_{A_{2,n}} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{4}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} + o_p(1). \qquad \text{(B.13)}
\end{aligned}$$

In the last line of Equation B.13, the parenthesized term is bounded due to the continuity of function $b_{\mathbf{y}_I,\mathbf{k}_I}(\boldsymbol{\theta})$, the boundedness of $w_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{y})$, the boundedness of set $A_{2,n}$, and our selection of δ. The $o_p(1)$ term comes from the uniform convergence of the second term in the Taylor expansion (Equation B.6). Also notice that

$$\int_{A_{2,n}} e^{-\frac{1}{4}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} \leq C e^{-C' B \log n} (\delta\sqrt{n} - B \log n)^q \leq C'' n^{q/2 - C'B}, \qquad \text{(B.14)}$$

where C, C', and C'' are constants. By selecting B large enough, Equation B.14 implies $\int_{A_{2,n}} e^{-\frac{1}{4}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} \to 0$. Finally, an argument using tightness similar to equations B.10 and B.11 shows that the RHS of Equation B.13 converges to 0 in probability.

Region $A_{1,n}$ The convergence on $A_{1,n}$ can be established similarly. Fix an arbitrary $B > 0$. For the particular δ we have selected,

$$\sup_{\mathbf{h}\in A_{1,n}} |R_{n,\mathbf{h}}| \leq M(\delta) \sup_{\mathbf{h}\in A_{1,n}} \frac{\|\mathbf{h}\|^3}{n^{3/2}} \leq M(\delta) B^3 \frac{\log^3 n}{n^{3/2}} = o(1), \qquad \text{(B.15)}$$

in which $M(\delta)$ is the same as in Equation B.12. Then

$$\int_{A_{1,n}} \left| \frac{a_{n,\mathbf{h}}}{f_{n,0}} - b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} \right| d\mathbf{h} \le \int_{A_{1,n}} b_{n,\mathbf{h}} \left| e^{R_{n,\mathbf{h}}} - 1 \right| e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} + \int_{A_{1,n}} |b_{n,\mathbf{h}} - b_{n,0}| e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} + o_p(1). \tag{B.16}$$

In Equation B.16, the $o_p(1)$ term is again due to the uniform convergence of the second term in Equation B.6. Equation B.15 implies that $\sup_{\mathbf{h} \in A_{1,n}} |e^{R_{n,\mathbf{h}}} - 1| \to 0$; together with $|b_{n,\mathbf{h}} - b_{n,0}| \overset{P_{\boldsymbol{\theta}_0}}{\to} 0$ and the boundedness of $A_{1,n}$, both integrals on the RHS of Equation B.16 converges to 0 in probability (the tightness argument needs to be used again).

Region $A_{4,n}$ Assume for a moment that there exists a large number B' such that

$$\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > B'} \min_{\mathbf{y}_i} f(\boldsymbol{\theta}, \mathbf{y}_i) < f(\boldsymbol{\theta}_0, \mathbf{y}_i^\circ)^{2/f(\boldsymbol{\theta}_0, \mathbf{y}_i^\circ)}, \tag{B.17}$$

in which $\mathbf{y}_i^\circ$ is the least plausible individual response pattern under $\boldsymbol{\theta}_0$. Also write $p_n(\mathbf{y}_i^\circ)$ be the observed proportion of $\mathbf{y}_i$; $p_n(\mathbf{y}_i^\circ) \overset{P_{\boldsymbol{\theta}_0}}{\to} f(\boldsymbol{\theta}_0, \mathbf{y}_i^\circ)$. Then, on region $A_{4,n}$ defined by such a B',

$$P\left\{ \frac{\min_{y_i} f(\boldsymbol{\theta}, \mathbf{y}_i)^{p_n(\mathbf{y}_i^\circ)}}{f(\boldsymbol{\theta}_0, \mathbf{y}_i^\circ)} < 1 \right\} \ge P\left\{ p_n(\mathbf{y}_i^\circ) > \frac{f(\boldsymbol{\theta}_0, \mathbf{y}_i^\circ)}{2} \right\} \to 1. \tag{B.18}$$

Therefore, we have

$$\frac{f_{n,\mathbf{h}}}{f_{n,0}} \le \left[\frac{\min_{y_i} f(\boldsymbol{\theta}, \mathbf{y}_i)^{p_n(\mathbf{y}_i^\circ)}}{f(\boldsymbol{\theta}_0, \mathbf{y}_i^\circ)} \right]^n \le \rho^n + o_p(1) \tag{B.19}$$

for some $0 < \rho < 1$. Also note that this likelihood ratio bound is not affected if finitely many observations are removed from $f_{n,\mathbf{h}}$, which is the case after dividing by the denominator of each summand of $b_{n,\mathbf{h}}$. As a result,

$$\int_{A_{4,n}} \left| \frac{a_{n,\mathbf{h}}}{f_{n,0}} - b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} \right| d\mathbf{h} \le \int_{A_{4,n}} \frac{b_{n,\mathbf{h}} f_{n,\mathbf{h}}}{f_{n,0}} d\mathbf{h} + b_{n,0} \int_{A_{4,n}} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} \le K\rho^n + b_{n,0} \int_{A_{4,n}} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} + o_p(1), \tag{B.20}$$

in which K is a constant. Equation B.20 results from the fact that: a) The numerator of Equation B.3 is integrable with respect to the Lebesgue measure on the parameter space, which contributes to the constant K; b) $p_n(\mathbf{y}_I) \overset{P_{\boldsymbol{\theta}_0}}{\to} \pi_0(\mathbf{y}_I)$, so the latter also contributes to K while the difference of the two is merged into the $o_p(1)$ term. The second term on the RHS of Equation B.20 converges to zero by a similar tightness argument and the tail estimate of a multivariate normal distribution. These altogether shows that the LHS of Equation B.20 converges to zero in probability.

Now we prove the result stated by Equation B.17; we denote the RHS of Equation B.17 by η.

First, consider the parameter subspace of α_{jk} and $\boldsymbol{\beta}_j$ for each j and k. Let $L_{jk} = \|(\alpha_{jk}\ \boldsymbol{\beta}_j{}^\top)^\top\|$, and $\mathbf{d}_{jk} = (\alpha_{jk}\ \boldsymbol{\beta}_j{}^\top)^\top / L_{jk} \in \mathbb{R}^{r+1}$ be a unit directional vector, in which the coordinates corresponding to fixed slopes are set to 0. Also introduce the partition $\mathbf{d}_{jk} = (x_{jk}\ \mathbf{e}_j{}^\top)^\top$ separating the direction of the intercept parameter, i.e., the first coordinate x_{jk}, from those of the slopes. Then, we write

$$\tau_{jk}(\boldsymbol{\theta}_j, \mathbf{Z}_i^\star) = \alpha_{jk} + \boldsymbol{\beta}_j{}^\top \mathbf{Z}_i^\star = L_{jk}(x_{jk} + \mathbf{e}_j{}^\top \mathbf{Z}_i^\star), \tag{B.21}$$

in which $x_{jk} + \mathbf{e}_j{}^\top \mathbf{Z}_i^\star \sim \mathcal{N}(x_{jk}, 1 - x_{jk}^2)$. For fixed $\mathbf{d}_{jk}$, define $H_{jk}^{\varepsilon}(y) = \{\mathbf{z}_i \in \mathbb{R}^r : (-1)^{\mathbb{I}\{y \geq k\}}(x_{jk} + \mathbf{e}_j{}^\top \mathbf{z}_i) \geq \varepsilon\}$ for $\varepsilon \geq 0$.

Now pool across multiple items. A direct consequence of Lemma 2, which is presented soon, is that $\mathbb{R}^r \subset \bigcup_{j=1}^{r+1} H_{jk}^0(y_{ij})$ for properly selected $(y_{ij})_{j=1}^{r+1}$ (recall that we assume $m > r$, so there are sufficient items). Then, for any $\varepsilon > 0$, the following bound can be established for the likelihood of an individual response pattern in which the first $r + 1$ items have the selected pattern $(y_{ij})_{j=1}^{r+1}$:

$$
\begin{aligned}
f(\boldsymbol{\theta}, \mathbf{y}_i) &= \int_{\mathbb{R}^r} \prod_{j=1}^{m} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}_i) \\
&\leq \sum_{j=1}^{r+1} \int_{H_{jk}^0(y_{ij})} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}_i) \\
&\leq \sum_{j=1}^{r+1} \int_{H_{jk}^\varepsilon(y_{ij})} f_j(\boldsymbol{\theta}_j, y_{ij}|\mathbf{z}_i) d\Phi(\mathbf{z}_i) + \sum_{j=1}^{r+1} \Phi\{H_{jk}^0(y_{ij}) \backslash H_{jk}^\varepsilon(y_{ij})\} \\
&\leq \sum_{j=1}^{r+1} \frac{1}{1+e^{\varepsilon L_{jk}}} + \sum_{j=1}^{r+1} \Phi\{H_{jk}^0(y_{ij}) \backslash H_{jk}^\varepsilon(y_{ij})\} \qquad \text{(B.22)}
\end{aligned}
$$

In the last line of Equation B.22, each summand of the second term can be made smaller than $\frac{\eta}{2(r+1)}$ by choosing a proper ε; this result can be strengthened to hold uniformly for all directions $\mathbf{d}_{jk}$ on $\mathbb{R}^{r+1}$, as a consequence of Lemma 3. In addition, since there are only finitely many intercept parameters, we can choose a large enough B' (i.e., $\boldsymbol{\theta}$ is sufficiently distant from $\boldsymbol{\theta}_0$) such that $\frac{1}{1+e^{\varepsilon L_{jk}}} < \frac{\eta}{2(r+1)}$ for all j and k. Consequently, for each $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > B'$, we are able to find an individual response pattern $\mathbf{y}_i$ such that the corresponding value of Equation B.22 can be bounded by the desired number η, which establishes the result stated by Equation B.17. The two lemmas required in the foregoing proof are presented next.

Lemma 2. *Consider a sequence of affine hyperplanes* $\{\mathbf{z} \in \mathbb{R}^r : \mathbf{a}_i^\top \mathbf{z} = b_i\}_{i=1}^k$. *Let half-space* H_i *be either* $\mathbf{a}_i^\top \mathbf{z} \geq b_i$ *or* $\mathbf{a}_i^\top \mathbf{z} \leq b_i$. *There exists some choice of* $\{H_i\}_{i=1}^k$ *such that* $\mathbb{R}^r \subset \bigcup_{i=1}^k H_i$, *if and only if* $\mathbf{a}_i$*'s are linearly dependent.*

Proof. ($\Leftarrow$) Suppose $\mathbf{a}_i$'s are linearly dependent. There exists an $\mathbf{a}_i$ that can be written as a non-trivial linear combination of the others. Without loss of generality, let $\mathbf{a}_1$ be such a vector:

$$\mathbf{a}_1 = \sum_{i=2}^{k} c_i \mathbf{a}_i, \qquad \text{(B.23)}$$

in which at least one c_i is non-zero. If $\sum_{i=2}^k c_i b_i \geq b_1$, then for $i = 2, \ldots, k$ set $H_i = \{\mathbf{z} :$

${\mathbf{a}_i}^\top \mathbf{z} \geq b_i\}$ when $c_i \leq 0$ and $H_i = \{\mathbf{z} : {\mathbf{a}_i}^\top \mathbf{z} \leq b_i\}$ when $c_i > 0$. It follows that

$$\bigcap_{i=2}^{k} H_i^c \subset \{\mathbf{z} : \sum_{i=2}^{k} c_i {\mathbf{a}_i}^\top \mathbf{z} > \sum_{i=2}^{k} c_i b_i\} \subset \{\mathbf{z} : {\mathbf{a}_1}^\top \mathbf{z} \geq b_1\}. \tag{B.24}$$

By letting H_1 be the RHS of Equation B.24, we have $\mathbb{R}^r \subset \bigcap_{i=1}^k H_i$. A similar argument can be used to establish the statement when $\sum_{i=2}^k c_i b_i < b_1$.

($\Rightarrow$) Suppose the $\mathbf{a}_i$'s are linearly independent, which implies that the set of equations $\{{\mathbf{a}_i}^\top \mathbf{z} = \mathbf{b}_i\}_{i=1}^k$ has at least one solution, denoted $\mathbf{z}'$. Consider the k-dimensional subspace spanned by the coordinate system $\{\mathbf{a}_i\}_{i=1}^n$ with an origin at $\mathbf{z}'$. For each i, the half-space H_i corresponds to either the positive or negative side of vector $\mathbf{a}_i$, depending on the direction of the inequality. No matter how we choose the H_i's, there will be one out of 2^k "orthants" corresponding to $\bigcap_{i=1}^k H_i^c$ left uncovered, which proves the "only if" part. □

Lemma 3. *Let* $Z_x \sim \mathcal{N}(x, 1 - x^2)$ *be a one-parameter family of normal random variables with* $x \in [-1, 1]$*. Given any* $\eta \in (0, 1/2)$*, there exists an* $\varepsilon > 0$ *such that* $\sup_{x\in[-1,1]} P(|Z_x| \leq \varepsilon) < \eta$*.*

Proof. By symmetry, $\sup_{x\in[0,1]} P(|Z_x| \leq \varepsilon) = \sup_{x\in[-1,1]} P(|Z_x| \leq \varepsilon)$, so we only need to consider non-negative x's in the proof. Note that for all $\varepsilon \in [0, 1)$ and $x > \varepsilon$,

$$P(Z_x \leq \varepsilon) = \Phi\left(\frac{\varepsilon - x}{\sqrt{1 - x^2}}\right) \downarrow 0, \tag{B.25}$$

as $x \uparrow 1$, due to the monotonicity of the functions involved. Now fix an $\eta \in (0, 1/2)$. Equation B.25 implies there exists an $x' \in (1/2, 1)$ such that $P(Z_{x'} \leq 1/2) < \eta$. Then for all $x \in (x', 1]$ and $\varepsilon \in (0, 1/2]$, we have

$$P(|Z_x| \leq \varepsilon) \leq P(Z_x \leq \varepsilon) \leq P(Z_{x'} \leq \varepsilon) < \eta. \tag{B.26}$$

For $x \in [0, x']$, the variance of Z_x is bounded from below by $1 - x'^2$. We select ε' such that

$P(|Z_{x'} - x'| \leq \varepsilon') < \eta$. Then by Anderson's inequality,

$$P(|Z_x| \leq \varepsilon') \leq P(|Z_x - x| \leq \varepsilon') \leq P(|Z_{x'} - x'| \leq \varepsilon') < \eta. \tag{B.27}$$

The statement follows by setting $\varepsilon = \min\{1/2, \varepsilon'\}$. $\square$

Region $A_{3,n}$ Let K_{-0} be any compact subset of Θ which is bounded away from $\boldsymbol{\theta}_0$. By a well-known application of Jensen's inequality:

$$E_{\boldsymbol{\theta}_0} \log \frac{f(\boldsymbol{\theta}, \mathbf{Y}_i)}{f(\boldsymbol{\theta}_0, \mathbf{Y}_i)} \leq \log E_{\boldsymbol{\theta}_0} \frac{f(\boldsymbol{\theta}, \mathbf{Y}_i)}{f(\boldsymbol{\theta}_0, \mathbf{Y}_i)} = 0. \tag{B.28}$$

In fact, the inequality in Equation B.28 is strict by the model identification assumption (ii) of Theorem 1. Because K_{-0} is compact, there exists a positive number κ such that

$$\sup_{\boldsymbol{\theta} \in K_{-0}} E_{\boldsymbol{\theta}_0} \log \frac{f(\boldsymbol{\theta}, \mathbf{Y}_i)}{f(\boldsymbol{\theta}_0, \mathbf{Y}_i)} < -\kappa, \tag{B.29}$$

by the continuity of the LHS function. Moreover, by the Uniform Law of Large Numbers,

$$\sup_{\boldsymbol{\theta} \in K_{-0}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(\boldsymbol{\theta}, \mathbf{Y}_i)}{f(\boldsymbol{\theta}_0, \mathbf{Y}_i)} - E_{\boldsymbol{\theta}_0} \log \frac{f(\boldsymbol{\theta}, \mathbf{Y}_i)}{f(\boldsymbol{\theta}_0, \mathbf{Y}_i)} \right| \overset{P_{\boldsymbol{\theta}_0}}{\to} 0. \tag{B.30}$$

Therefore, $\sup_{\boldsymbol{\theta} \in K_{-0}} \prod_{i=1}^{n} f(\boldsymbol{\theta}, \mathbf{Y}_i) / \prod_{i=1}^{n} f(\boldsymbol{\theta}_0, \mathbf{Y}_i) \overset{P_{\boldsymbol{\theta}_0}}{\to} 0$, which implies

$$\sup_{\mathbf{h} \in A_{3,n}} \frac{f_{n,\mathbf{h}}}{f_{n,0}} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0, \tag{B.31}$$

because $\mathbf{h} \in A_{3,n}$ implies $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \in [\delta, B']$. It follows that

$$\begin{aligned}
&\int_{A_{3,n}} \left| \frac{b_{n,\mathbf{h}} f_{n,\mathbf{h}}}{f_{n,0}} - b_{n,0} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} \right| d\mathbf{h} \\
&\leq \int_{A_{3,n}} \left| \frac{b_{n,\mathbf{h}} f_{n,\mathbf{h}}}{f_{n,0}} \right| d\mathbf{h} + b_{n,0} \int_{A_{3,n}} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h} \\
&\leq \sup_{\mathbf{h} \in A_{3,n}} \left| \frac{f_{n,\mathbf{h}}}{f_{n,0}} \right| \int_{A_{3,n}} b_{n,\mathbf{h}} d\mathbf{h} + b_{n,0} \int_{A_{3,n}} e^{\mathbf{h}^\top \mathbf{S}_n - \frac{1}{2}\mathbf{h}^\top \boldsymbol{\mathcal{I}}_0 \mathbf{h}} d\mathbf{h}.
\end{aligned} \tag{B.32}$$

Equation B.32 converges in probability to 0 due to the integrability of $b_{n,\mathbf{h}}$, the tail estimates of a multivariate normal distribution, and the tightness of $\mathbf{S}_n$. The proof is now complete.

□

APPENDIX C: NON-UNIQUENESS DUE TO SELECTION RULES

Proof of Theorem 2. Recall that $\mathbf{V} = \mathbf{v}(Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star))$, i.e., an extremal point of the non-empty set inverse function $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)$, has density $g_n(\boldsymbol{\theta}|\mathbf{y})$[1]. Take $\delta > 0$. For each fixed $\mathbf{y}$, $\rho_K(\mathbf{y})$, defined by Equation 2.19, can be bounded by

$$\begin{aligned}
\rho_K(\mathbf{y}) &= P^\star\{\text{diam}Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) > K/n \mid Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\} \\
&= P^\star\{\text{diam}Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) > K/n, \|\mathbf{V} - \boldsymbol{\theta}_0\| \leq \delta \mid Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\} \\
&\quad + P^\star\{\text{diam}Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) > K/n, \|\mathbf{V} - \boldsymbol{\theta}_0\| > \delta \mid Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\} \\
&\leq P^\star\{\text{diam}Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) > K/n \mid \|\mathbf{V} - \boldsymbol{\theta}_0\| \leq \delta\} \\
&\quad + P^\star\{\|\mathbf{V} - \boldsymbol{\theta}_0\| > \delta \mid Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}. \qquad \text{(C.1)}
\end{aligned}$$

Theorem 1 implies that for $\mathbf{Y}$ generated from $P_{\boldsymbol{\theta}_0}$, $P^\star\{\|\mathbf{V} - \boldsymbol{\theta}_0\| > \delta \mid Q(\mathbf{Y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\}$, as a measurable function of $\mathbf{Y}$, converges to 0 in $P_{\boldsymbol{\theta}_0}$-probability: i.e.,

$$P^\star\{\|\mathbf{V} - \boldsymbol{\theta}_0\| > \delta \mid Q(\mathbf{Y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset\} = \int_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \delta} g_n(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0. \qquad \text{(C.2)}$$

---

[1] The definitions of $\mathbf{V}$ and $g_n(\boldsymbol{\theta}|\mathbf{y})$ are conditional on $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset$. In the sequel, when $\mathbf{V}$ appears in the conditioning, the notation automatically implies conditioning on $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star) \neq \emptyset$ as well.

Hence, we focus on the first term in Equation C.1. This term can be further bounded by:

$$\begin{aligned}
&P^\star\{\text{diam}Q(\mathbf{y},\mathbf{A}^\star,\mathbf{Z}^\star) > K/n \mid \|\mathbf{V}-\boldsymbol{\theta}_0\| \le \delta\} \\
=& \sum_I \sum_{\mathbf{k}_I} P^\star\{\text{diam}Q(\mathbf{y},\mathbf{A}^\star,\mathbf{Z}^\star) > K/n, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} \mid \|\mathbf{V}-\boldsymbol{\theta}_0\| \le \delta\} \\
=& \sum_I \sum_{\mathbf{k}_I} P^\star\{\text{diam}Q(\mathbf{y},\mathbf{A}^\star,\mathbf{Z}^\star) > K/n \mid \|\mathbf{V}-\boldsymbol{\theta}_0\| \le \delta, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I}\} \\
&\qquad \cdot P^\star\{\mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} \mid \|\mathbf{V}-\boldsymbol{\theta}_0\| \le \delta\} \\
=& \sum_{\mathbf{y}_I} \sum_{\mathbf{k}_I} P^\star\{\text{diam}Q(\mathbf{y},\mathbf{A}^\star,\mathbf{Z}^\star) > K/n \mid \|\mathbf{V}-\boldsymbol{\theta}_0\| \le \delta, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I}\} \\
&\qquad \cdot \left(\sum_{I':\mathbf{y}_{I'}=\mathbf{y}_I} P^\star\{\mathbf{V} = \mathbf{V}_{I',\mathbf{k}_{I'}} \mid \|\mathbf{V}-\boldsymbol{\theta}_0\| \le \delta\} \right) \\
\le& \sum_{\mathbf{y}_I} \sum_{\mathbf{k}_I} P^\star\{\text{diam}Q(\mathbf{y},\mathbf{A}^\star,\mathbf{Z}^\star) > K/n \mid \|\mathbf{V}_{I,\mathbf{k}_I}-\boldsymbol{\theta}_0\| \le \delta, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I}\} \\
=& \sum_{\mathbf{y}_I} \sum_{\mathbf{k}_I} \int P^\star\{\text{diam}Q(\mathbf{y},\mathbf{A}^\star,\mathbf{Z}^\star) > K/n \\
&\qquad \mid \|\mathbf{V}_{I,\mathbf{k}_I}-\boldsymbol{\theta}_0\| \le \delta, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I}, \mathbf{Z}_I^\star = \mathbf{z}_I\} d\Phi(\mathbf{z}_I)
\end{aligned} \tag{C.3}$$

The first sum over index sets I in the third line of Equation C.3 can be collapsed into a finite sum over all patterns of $\mathbf{y}_I$ (i.e., the fourth equation), for the reason that sub-samples I and I' having the same response pattern $\mathbf{y}_I = \mathbf{y}_{I'}$ are exchangable under our requirement of the selection rules. Note that the event being conditioned on in the integrand of the last line of Equation C.3 happens with a positive probability almost surely under the probability measure of $\mathbf{Z}^\star$; to simplify notation, write $E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I) = \{\|\mathbf{V}_{I,\mathbf{k}_I}-\boldsymbol{\theta}_0\| \le \delta, \mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I}, \mathbf{Z}_I^\star = \mathbf{z}_I\}$ as that event. Because there are only finitely many combinations of $\mathbf{y}_I$ and $\mathbf{k}_I$, it suffices to prove that for each $\varepsilon > 0$ and some $\delta > 0$,

$$P_{\boldsymbol{\theta}_0}\left\{\exists K, N > 0 : \int P^\star\{\text{diam}Q(\mathbf{Y},\mathbf{A}^\star,\mathbf{Z}^\star) > K/n \mid E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)\} d\Phi(\mathbf{z}_I) < \varepsilon,\ \forall n > N\right\} \to 1. \tag{C.4}$$

So fix $\mathbf{y}_I$, $\mathbf{k}_I$, and δ for the rest of the proof. Also note that conditional on $E^{\delta}_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)$, the remaining observations $i \notin I$ are independent.

To proceed, we sequentially project the set inverse $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)$ onto $J = \sum_{j=1}^m K_j - m$ subspaces, each of which spanned by an intercept parameter α_{jk} and the free slopes in $\boldsymbol{\beta}_j$ (for the same j). For each projection, we find a bounding random variable for its diameter; then, the sum of constructed bounds across all projections serves as a upper bound, up to a constant multiplier depending on the dimension of the parameter space, for the diameter of the set inverse. We prove the result stated in equation C.4 with the diameter of $Q(\mathbf{y}, \mathbf{A}^\star, \mathbf{Z}^\star)$ replaced by the constructed bound. In order to establish the desired property for the bounding variables, we allocate the rest observations (i.e., not in I) to each projection, and subsequently use the standard theory for order statistics of i.i.d. random variables. In particular, we rearrange those observations to fill a growing three-dimensional array indexed by a triplet of indices s, j, and k: The last dimension of the array $k = 1, \cdots, K_j - 1$ is filled first[2], then $j = 1, \ldots, m$, and finally s; therefore, only the first dimension indexed by $s = \lfloor n/J \rfloor$ grows as the sample size increases. Notationally, elements corresponding to an observation in the array are denoted by a subscript $[sjk]$[3].

Fix $\mathbf{V} = \mathbf{V}_{I,\mathbf{k}_I} = \boldsymbol{\theta}$ for now. For each item j, let $\tilde{\boldsymbol{\beta}}_j$ be the collection of the r_j free slopes. Also for each $k = 1, \ldots, K_j - 1$, let $\boldsymbol{\theta}_{jk} = (\alpha_{jk}, \tilde{\boldsymbol{\beta}}_j{}^\top)^\top$. $\boldsymbol{\theta}_{jk}$ is uniquely determined by a properly selected size-$(r_j + 1)$ subset of I_j, denoted I_{jk}, which can be determined from the fixed combination of $\mathbf{y}_I$ and $\mathbf{k}_I$[4]. Now intersecting the half-space of a new observation $[sjk]$ in the three-dimensional array with those of observations I_{jk}, the resulting intersection on the subspace of $\boldsymbol{\theta}_{jk}$ can be either bounded (i.e., a simplex) or unbounded. The following lemma provides sufficient and necessary conditions for the (un)bounded case:

---

[2] Each item may have different numbers of response categories, so it is in fact a "ragged" array.

[3] For example, if $I = \{1, \ldots, \sum_{j=1}^m q_j\}$ is the first $\sum_{j=1}^m q_j$ observations in the sample, then $[sjk]$ corresponds to the observation $i = \sum_{j=1}^m q_j + (s-1)J + \sum_{j'=1}^{j-1}(K_{j'} - 1) + k$.

[4] Only one half-space is selected for each $i \in I_{jk}$, which is not reflected in our notation for succinctness.

Lemma 4. *Consider $p+1$ half-spaces: $H_i = \{\mathbf{x} \in \mathbb{R}^p : {\mathbf{n}_i}^\top \mathbf{x} \leq b_i\}$, $i = 1, \ldots, p+1$, in which $\mathbf{n}_i$'s are considered fixed. Then, the following statements are equivalent:*

(i) $\bigcap_{i=1}^{p+1} H_i$ is bounded for all choices of b_i's, $i = 1, \ldots, p+1$, such that the intersection is not empty;

(ii) $\bigcap_{i=1}^{p+1} H_i$ is a bounded simplex for some choices of b_i's, $i = 1, \ldots, p+1$;

(iii) For all $\mathbf{c} \in \mathbb{R}^p$, there exists $i \in \{1, \ldots, p+1\}$ such that ${\mathbf{n}_i}^\top \mathbf{c} > 0$;

(iv) There exists $i \in \{1, \ldots, p+1\}$ such that $\mathbf{n}_j$'s, $j \neq i$, are linear independent, and that $\mathbf{n}_i = -\sum_{j \neq i} \gamma_j \mathbf{n}_j$ with $\gamma_j > 0$ for all $j \neq i$.

Proof. (i) $\Rightarrow$ (ii). We can always make the intersection non-empty by choosing $b_i > 0$ for all $i = 1, \ldots, p+1$. In this case, $\bigcap_{i=1}^{p+1} H_i$ must contain some neighborhood of $\mathbf{0}$. So (i) $\Rightarrow$ (ii) is trivial.

(ii) $\Rightarrow$ (iii). Fix b_i's, $i = 1, \ldots, p+1$, such that $\bigcap_{i=1}^{p+1} H_i$ is a bounded simplex. Take $\mathbf{x}_0 \in \bigcap_{i=1}^{p+1} H_i$; i.e., ${\mathbf{n}_i}^\top \mathbf{x}_0 \leq b_i$ for all $i = 1, \ldots, p+1$. If there exists $\mathbf{c} \in \mathbb{R}^p$ such that $\mathbf{c}^\top \mathbf{n}_i \leq 0$ for all i, then ${\mathbf{n}_i}^\top \mathbf{x}_0 + \lambda {\mathbf{n}_i}^\top \mathbf{c} \leq b_i$ for all i and all $\lambda > 0$. This implies $\mathbf{x}_0 + \lambda \mathbf{c} \in \bigcap_{i=1}^{p+1} H_i$ for all $\lambda > 0$, which contradicts the boundedness.

(iii) $\Rightarrow$ (i). On each direction $\mathbf{c}$, choose i such that ${\mathbf{n}_i}^\top \mathbf{c} > 0$. For every possible value of the corresponding b_i, there exists some $\lambda_0 > 0$ such that for all $\lambda > \lambda_0$, ${\mathbf{n}_i}^\top (\lambda \mathbf{c}) > b_i$, i.e., $\lambda \mathbf{c}_i \notin H_i$. So $\bigcap_{i=1}^{p+1} H_i$ is always bounded.

(iii) $\Rightarrow$ (iv). Let C_i be the convex cone defined by all but the ith normal vectors. (iii) implies $-{\mathbf{n}_i}^\top \mathbf{c} < 0$ for all $\mathbf{c} \in C_i^N = \{\mathbf{c} : {\mathbf{n}_i}^\top \mathbf{c} \leq 0$, for all $j \neq i\}$, i.e., the normal cone (denoted by a superscript N) of C_i. Hence, $-\mathbf{n}_i \in (C_i^N)^N = C_i$.

(iv) $\Rightarrow$ (iii). For $\mathbf{c} \in C_i^N$, (iv) implies ${\mathbf{n}_i}^\top \mathbf{c} > 0$. For $\mathbf{c} \notin C_i^N$, there exists some $j \neq i$ such that ${\mathbf{n}_j}^\top \mathbf{c} > 0$. $\square$

Let $\tilde{\mathbf{z}}_{ij}$ be the elements of $\mathbf{z}_i$ associated with $\tilde{\boldsymbol{\beta}}_j$. For each $i \in I_{jk}$, write $\mathbf{n}_{ijk} = \omega_{ijk}(\delta_{ijk} \;\; {\tilde{\mathbf{z}}_{ij}}^\top)^\top$ as the normal vector of the corresponding (r_j+1)-dimensional half-space, in which $\omega_{ijk} = \pm 1$ and $\delta_{ijk} \in \{0, 1\}$ are determined by the item response y_{ij}. Similar notation is defined for observations in the array: Let $\tilde{\mathbf{Z}}^\star_{[sjk]}$ be the elements of $\mathbf{Z}^\star_{[sjk]}$ associated with

$\tilde{\boldsymbol{\beta}}_j$, and $\mathbf{N}^\star_{[sjk]} = \omega_{[sjk]}(\delta_{[sjk]}\ \ \tilde{\mathbf{Z}}^\star_{[sjk]}{}^\top)^\top$ be the corresponding (random) normal vector; the random variables $\omega_{[sjk]} = \pm 1$ and $\delta_{[sjk]} \in \{0, 1\}$ depend on this observation's response to item j, which is denoted $y_{[sjk]}$ for simplicity. For each j and k, Lemma 4 implies that observation $[sjk]$ produces a bounded intersection if there exist positive real numbers $(\gamma_i)_{i \in I_{jk}}$ such that

$$\omega_{[sjk]} \tilde{\mathbf{Z}}^\star_{[sjk]} = -\sum_{i \in I_{jk}} \gamma_i \omega_{ijk} \tilde{\mathbf{z}}_{ij}, \tag{C.5}$$

and

$$\omega_{[sjk]} \delta_{[sjk]} = -\sum_{i \in I_{jk}} \gamma_i \omega_{ijk} \delta_{ijk}. \tag{C.6}$$

Conditioning on $\mathbf{V}_j = \mathbf{V}_{I_j, \mathbf{k}_{I_j}} = \boldsymbol{\theta}_j$, the intersection cannot be empty, which introduces a truncation to $A^\star_{[sjk]}$, i.e., the associated logistic variate for observation $[sjk]$ and item j:

$$(-1)^{\mathbb{I}\{y_{[sjk]} \geq k'\}} (A^\star_{[sjk]} - \alpha_{jk'} - \tilde{\boldsymbol{\beta}}_j{}^\top \tilde{\mathbf{Z}}^\star_{[sjk]}) \geq 0 \text{ for all } k' = 1, \ldots, K_j - 1. \tag{C.7}$$

Fix j and k. When equations C.5 and C.6 hold, let $\boldsymbol{\theta}^i_{[sjk]} = (\alpha^i_{jk}\ \ \boldsymbol{\beta}^i_{jk}{}^\top)^\top$, $i \in I_{jk}$, be the vertex on the subspace of $\boldsymbol{\theta}_{jk}$ determined by observations $I_{jk} \setminus \{i\}$ together with the new observation $[sjk]$, which is random due to its dependency on $A^\star_{[sjk]}$ and $\mathbf{Z}^\star_{[sjk]}$. Also let $I^i_{jk} = I_{jk} \setminus \{i\}$ for some $i \in I_{jk}$, $\boldsymbol{\delta}_{I^i_{jk}} = (\delta_{ijk})_{i \in I^i_{jk}}$, and treat $\tilde{\mathbf{z}}_{I^i_{jk}} = (\tilde{\mathbf{z}}_{ij})_{i \in I^i_{jk}}$ as an $r_j \times r_j$ matrix throughout this part of derivation. A geometric illustration of these notations for $r = 1$ is shown in Figure C.1.

Applying the formula for inverting a partitioned matrix, we have

$$\begin{pmatrix} \tilde{\mathbf{z}}_{I^i_{jk}} & \boldsymbol{\delta}_{I^i_{jk}} \\ \tilde{\mathbf{Z}}^\star_{[sjk]}{}^\top & \delta_{[sjk]} \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{\mathbf{z}}^{-1}_{I^i_{jk}} + \dfrac{\tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}} \mathbf{Z}^\star_{[sjk]}{}^\top \tilde{\mathbf{z}}^{-1}_{I^i_{jk}}}{\delta_{[sjk]} - \tilde{\mathbf{Z}}^\star_{[sjk]}{}^\top \tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}}} & \dfrac{-\tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}}}{\delta_{[sjk]} - \tilde{\mathbf{Z}}^\star_{[sjk]}{}^\top \tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}}} \\ \dfrac{-\mathbf{Z}^\star_{[sjk]}{}^\top \tilde{\mathbf{z}}^{-1}_{I^i_{jk}}}{\delta_{[sjk]} - \tilde{\mathbf{Z}}^\star_{[sjk]}{}^\top \tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}}} & \dfrac{1}{\delta_{[sjk]} - \tilde{\mathbf{Z}}^\star_{[sjk]}{}^\top \tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}}} \end{pmatrix}. \tag{C.8}$$

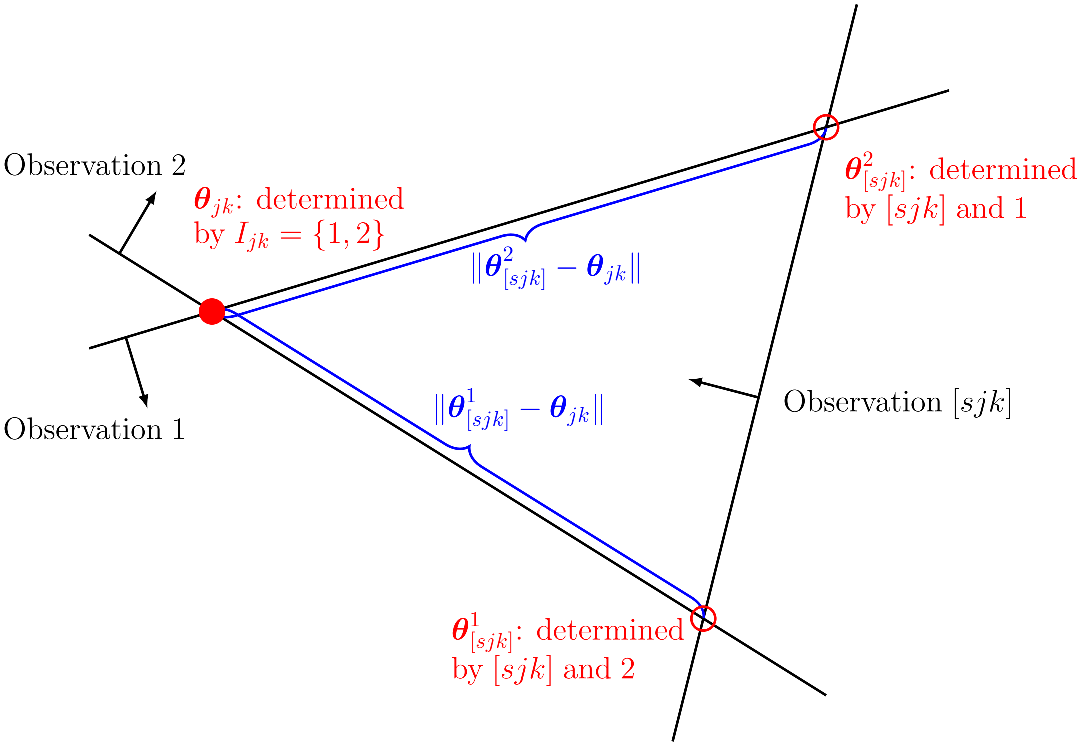
Observation 2
$\boldsymbol{\theta}_{jk}$: determined
by $I_{jk} = \{1, 2\}$
$\boldsymbol{\theta}^2_{[sjk]}$: determined
by $[sjk]$ and 1
$\|\boldsymbol{\theta}^2_{[sjk]} - \boldsymbol{\theta}_{jk}\|$
$\|\boldsymbol{\theta}^1_{[sjk]} - \boldsymbol{\theta}_{jk}\|$
Observation $[sjk]$
Observation 1
$\boldsymbol{\theta}^1_{[sjk]}$: determined
by $[sjk]$ and 2

Figure C.1: Illustration of notation used in the proof of Theorem 2. Here, $r = 1$, and j and k are fixed. $I_{jk} = \{1, 2\}$, which determines the fixed vertex $\boldsymbol{\theta}_{jk}$ (shown as a red dot). The line corresponding to the new observation $[sjk]$ intersects with those of observations 1 and 2, respectively, and produces two new vertices $\boldsymbol{\theta}^2_{[sjk]}$ and $\boldsymbol{\theta}_{[sjk]^1}$ (shown as red circles). The sum of $\|\boldsymbol{\theta}^1_{[sjk]} - \boldsymbol{\theta}_{jk}\|$ and $\|\boldsymbol{\theta}^2_{[sjk]} - \boldsymbol{\theta}_{jk}\|$ (highlighted in blue) gives an upper bound of the diameter of the plotted triangle.

It follows that the elements of $\boldsymbol{\theta}^i_{[sjk]} - \boldsymbol{\theta}_{jk}$ can be expressed as following:

$$\tilde{\boldsymbol{\beta}}^i_{[sjk]} - \tilde{\boldsymbol{\beta}}_j = \frac{\tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}} (A^\star_{[sjk]} - \tilde{\boldsymbol{\beta}}_j{}^\top \tilde{\mathbf{Z}}^\star_{[sjk]} - \alpha_{jk})}{\tilde{\mathbf{Z}}^\star_{[sjk]}{}^\top \tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}} - \delta_{[sjk]}}, \tag{C.9}$$

and

$$\alpha^i_{[sjk]} - \alpha_{jk} = -\tilde{\mathbf{z}}_{ij}{}^\top (\tilde{\boldsymbol{\beta}}^i_{[sjk]} - \tilde{\boldsymbol{\beta}}_j), \text{ for all } i \in I^i_{jk} \text{ such that } \delta_{ijk} = 1. \tag{C.10}$$

Define

$$\bar{U}^{\star}_{[sjk]} = \left[\sum_{i \in I_{jk}} \frac{\left\|\tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}}\right\| \left(1 + \sum_{i' \in I^i_{jk}} \|\tilde{\mathbf{z}}_{i'j}\|^2\right)}{\left|{\tilde{\mathbf{Z}}^{\star}_{[sjk]}}^{\top} \tilde{\mathbf{z}}^{-1}_{I^i_{jk}} \boldsymbol{\delta}_{I^i_{jk}} - \delta_{[sjk]}\right|}\right] |A^{\star}_{[sjk]} - \tilde{\boldsymbol{\beta}}_j^{\top} \tilde{\mathbf{Z}}^{\star}_{[sjk]} - \alpha_{jk}\delta_{[sjk]}| \tag{C.11}$$

If both equations C.5 and C.6 are satisfied, the random variable defined by Equation C.11 gives an upper bound for $\|\boldsymbol{\theta}^i_{[sjk]} - \boldsymbol{\theta}\|$. Also define

$$U^{\star}_{[sjk]} = \begin{cases} \bar{U}^{\star}_{[sjk]}, & \text{if equations C.5) and C.6 hold;} \\ \\ \infty, & \text{otherwise.} \end{cases} \tag{C.12}$$

which is a random variable that is defined on the extended real line.

Pooling across all observations in the array, we have

$$\text{diam} Q(\mathbf{y}, \mathbf{A}^{\star}, \mathbf{Z}^{\star}) \leq C \sum_{j=1}^{m} \sum_{k=1}^{K_j - 1} \min_{t \leq s} U^{\star}_{[tjk]}, \tag{C.13}$$

in which C is a constant determined by the dimension of the parameter space. It follows that

$$\begin{aligned} &\int P^{\star}\{\text{diam} Q(\mathbf{y}, \mathbf{A}^{\star}, \mathbf{Z}^{\star}) > K/n \mid E^{\delta}_{\mathbf{y}_I, \mathbf{k}_I}(\mathbf{z}_I)\} d\Phi(\mathbf{z}_I) \\ &\leq \int P^{\star}\left\{\sum_{j=1}^{m} \sum_{k=1}^{K_j - 1} \min_{t \leq s} U^{\star}_{[tjk]} > \frac{K}{Cn} \;\middle|\; E^{\delta}_{\mathbf{y}_I, \mathbf{k}_I}(\mathbf{z}_I)\right\} d\Phi(\mathbf{z}_I) \\ &\leq \sum_{j=1}^{m} \sum_{k=1}^{K_j - 1} \int P^{\star}\left\{\min_{t \leq s} U^{\star}_{[tjk]} > K'/n \;\middle|\; E^{\delta}_{\mathbf{y}_I, \mathbf{k}_I}(\mathbf{z}_I)\right\} d\Phi(\mathbf{z}_{I_{jk}}), \end{aligned} \tag{C.14}$$

in which $K' = \frac{K}{CJ}$. Now fix $\varepsilon, \delta > 0$. It suffices to prove for each summand of Equation C.14:

$$P_{\boldsymbol{\theta}_0}\left\{\exists K', N > 0 : \int P^\star \left\{\min_{t\leq s} U^\star_{[tjk]} > K'/n \;\middle|\; E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)\right\} d\Phi(\mathbf{z}_I) < \varepsilon,\ \forall n > N\right\} \to 1. \tag{C.15}$$

For each plausible response $y_{[1jk]}$, we assume the existence of a growing sub-collection of $\{1, \ldots, s\}$, $T_s(y_{[1jk]}) = \{t : t \leq s, y_{[tjk]} = y_{[1jk]}\}$, satisfying

$$|T_s(y_{[1jk]})|/n \to \rho(y_{[1jk]}),\ \text{as } n \to \infty \text{ for some } 0 < \rho(y_{[1jk]}) < 1. \tag{C.16}$$

Because there are only finitely many $y_{[1jk]}$ values, we write $\rho_0 = \min_{y_{[1jk]}} \rho(y_{[1jk]})$. Within each sub-collection, $U^\star_{[tjk]}$, $t \in T_s(y_{[1jk]})$, are i.i.d.. Let $\varphi_{jk}(u, \boldsymbol{\theta}_{jk}, \mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[1jk]})$ be the density of $U^\star_{[1jk]}$ conditional on $E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)$. We intend to find a set $B_{jk} \subset \mathbb{R}^{r_j^2}$ such that $P\{\tilde{\mathbf{Z}}^\star_{I_{jk}} \notin B_{jk}\} < \varepsilon/2$, and also a $\kappa > 0$ such that for every $\tilde{\mathbf{z}}_{I_{jk}} \in B_{jk}$, there exists a particular $y_{[1jk]}$ for which

$$\inf\left\{\varphi_{jk}(u, \boldsymbol{\theta}_{jk}, \mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[1jk]}) : 0 \leq u \leq \eta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta\right\} \geq \kappa \tag{C.17}$$

for some $\eta > 0$. Assume for a moment that equations C.16 and C.17 holds. Then we can construct a sequence of i.i.d. non-negative random variables $\{X_n\}$, whose density function is constantly equal to κ within $[0, \eta]$. By the delta method and the standard result for i.i.d. uniform order statistics, $n \min_{i\leq n} X_i \xrightarrow{d} W/\kappa$, in which $W \sim \text{Exp}(1)$. Fix K' such that $P\left(W/\kappa > K'\right) < \varepsilon/4$. By the Portmanteau Lemma, there exists an n_1 such that for all $n > n_1$, $P\{n \min_{i\leq\lfloor\rho_0 n/2\rfloor} X_i > K'\} \leq P\{W/\kappa > K'\} + \varepsilon/4 \leq \varepsilon/2$. Also take n_2 such that $K'/n_2 < \eta$, and n_3 such that $|T_s(y_{[1jk]})|/n > \rho_0/2$ for all $y_{[1jk]}$. Thus, for every $\mathbf{z}_{I_{jk}} \in B_{jk}$, there exists a particular $y_{[1jk]}$ such that along the corresponding subsequence $T_s(y_{[1jk]})$:

$$\begin{aligned}
&P^\star\left\{\min_{t\le s} U^\star_{[tjk]} > K'/n \,\middle|\, E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)\right\}\\
&\le P^\star\left\{\min_{t\in T_s(y_{[1jk]})} U^\star_{[tjk]} > K'/n \,\middle|\, E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)\right\}\\
&\le P\left\{\min_{i\le\lfloor\rho_0 n/2\rfloor} X_n > K'/n\right\}\\
&\le \varepsilon/2 \qquad\qquad (C.18)
\end{aligned}$$

for all $n > \max\{n_1, n_2, n_3\}$. It follows that for all these large n's,

$$\begin{aligned}
&\int P^\star\left\{\min_{t\le s} U^\star_{[tjk]} > K'/n \,\middle|\, E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)\right\} d\Phi(\mathbf{z}_{I_{jk}})\\
\le\ &\int_{B_{jk}} P^\star\left\{\min_{t\le s} U^\star_{[tjk]} > K'/n \,\middle|\, E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)\right\} d\Phi(\mathbf{z}_{I_{jk}}) + \varepsilon/2\\
\le\ &\varepsilon, \qquad\qquad (C.19)
\end{aligned}$$

When $\mathbf{Y}$ is considered random, the probability that Equation C.16 holds for all plausible $y_{[1jk]}$ goes to 1, because the data generating parameter values $\boldsymbol{\theta}_0$ is assumed to be in the interior of the parameter space (and thus $\rho_0 > 0$ is determined solely by $\boldsymbol{\theta}_0$). This implies the intended results (Equation C.15).

Let $\bar\varphi_{jk}(u, \boldsymbol{\theta}_{jk}, \mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[sjk]})$ be the density of $\bar U_{[sjk]}$ conditional on $E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)$, and

$$\begin{aligned}
C_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[sjk]}) = \{&\tilde{\mathbf{Z}}^\star_{[sjk]} = -\sum_{i\in I_{jk}} \gamma_i\omega_{ijk}\tilde{\mathbf{z}}_{ij},\\
&\delta_{[sjk]} = -\sum_{i\in I_{jk}} \gamma_i\omega_{ijk}\delta_{ijk},\\
&\gamma_i > 0 \text{ for all } i \in I_{jk}\}. \qquad\qquad (C.20)
\end{aligned}$$

Then $\varphi_{jk}(u, \boldsymbol{\theta}_{jk}, \mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[sjk]}) = \bar\varphi_{jk}(u, \boldsymbol{\theta}_{jk}, \mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[sjk]}) P\{C_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[sjk]}) | E^\delta_{\mathbf{y}_I,\mathbf{k}_I}(\mathbf{z}_I)\}$. Next, we find lower bounds for the two parts on the RHS, respectively.

First, fix a $y_{[sjk]}$ ensuring equations C.5 and C.6 hold. For easy reference, let

$$\sigma_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, \tilde{\mathbf{Z}}^{\star}_{[sjk]}, y_{[sjk]}) = \sum_{i \in I_{jk}} \frac{\left\| \tilde{\mathbf{z}}_{I^i_{jk}}^{-1} \boldsymbol{\delta}_{I^i_{jk}} \right\| \left(1 + \sum_{i' \in I^i_{jk}} \|\tilde{\mathbf{z}}_{i'j}\|^2 \right)}{\left| \tilde{\mathbf{Z}}^{\star}_{[sjk]}{}^{\top} \tilde{\mathbf{z}}_{I^i_{jk}}^{-1} \boldsymbol{\delta}_{I^i_{jk}} - \delta_{[sjk]} \right|} \tag{C.21}$$

and

$$\mu_{jk}(\boldsymbol{\theta}_{jk}, \tilde{\mathbf{Z}}^{\star}_{[sjk]}, y_{[sjk]}) = \tilde{\boldsymbol{\beta}}_j{}^{\top} \tilde{\mathbf{Z}}^{\star}_{[sjk]} + \alpha_{jk} \delta_{[sjk]}. \tag{C.22}$$

Then we rewrite Equation C.11 as $\bar{U}_{[sjk]} = \sigma_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, \tilde{\mathbf{Z}}^{\star}_{[sjk]}, y_{[sjk]}) | A^{\star}_{[sjk]} - \mu_{jk}(\boldsymbol{\theta}_{jk}, \tilde{\mathbf{Z}}^{\star}_{[sjk]}, y_{[sjk]})|$, whose density function is

$$\begin{aligned} &\bar{\varphi}_{jk}(u, \boldsymbol{\theta}_{jk}, \mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[sjk]}) \\ =& \int \frac{\bar{\psi}(\mu_{jk}(\boldsymbol{\theta}_{jk}, \tilde{\mathbf{z}}_{[sjk]}, y_{[sjk]}) + u / \sigma_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, \tilde{\mathbf{z}}_{[sjk]}, y_{[sjk]}))}{\sigma_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, \tilde{\mathbf{z}}_{[sjk]}, y_{[sjk]})} d\Phi(\tilde{\mathbf{z}}_{[sjk]}) \\ +& \int \frac{\bar{\psi}(\mu_{jk}(\boldsymbol{\theta}_{jk}, \tilde{\mathbf{z}}_{[sjk]}, y_{[sjk]}) - u / \sigma_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, \tilde{\mathbf{z}}_{[sjk]}, y_{[sjk]}))}{\sigma_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, \tilde{\mathbf{z}}_{[sjk]}, y_{[sjk]})} d\Phi(\tilde{\mathbf{z}}_{[sjk]}) \end{aligned} \tag{C.23}$$

in which $\bar{\psi}(\cdot)$ is the standard logistic density conditional on Equation C.7. By the theory of multivariate normal random variables, we can find

$$\begin{aligned} B^1_{jk} = \{ \tilde{\mathbf{z}}_{I_{jk}} \in \mathbb{R}^{r_j^2} : \lambda \leq \|\tilde{\mathbf{z}}_{ij}\| \leq L, \text{ for all } i \in I_{jk}; \\ \lambda' \leq \mathbf{x}^{\top} \tilde{\mathbf{z}}_{I_{jk}}{}^{\top} \tilde{\mathbf{z}}_{I_{jk}} \mathbf{x} \leq L', \text{ for all } \mathbf{x} \in \mathbb{R}^{r_j}, \|x\| = 1 \} \end{aligned} \tag{C.24}$$

with properly defined λ, λ', L, and L' such that $P^{\star}\{\tilde{\mathbf{Z}}^{\star}_{I_{jk}} \in B^1_{jk}\} > 1 - \varepsilon/4$. Also for fixed $D' > 0$ and $D > \delta > 0$, define

$$\begin{aligned} G_{jk}(\tilde{\mathbf{z}}_{I_{jk}}) = \{ \tilde{\mathbf{z}}_{[sjk]} \in \mathbb{R}^{r_j} : \delta' \leq \left| \tilde{\mathbf{z}}_{[sjk]}{}^{\top} \tilde{\mathbf{z}}_{I^i_{jk}}^{-1} \boldsymbol{\delta}_{I^i_{jk}} - \delta_{[sjk]} \right| \leq D' \text{ for all } i \in I_{jk}, \\ \|\tilde{\mathbf{z}}_{[sjk]}\| \leq D \}. \end{aligned} \tag{C.25}$$

Note that $\tilde{\mathbf{Z}}^{\star}_{[sjk]}{}^{\top}\tilde{\mathbf{z}}^{-1}_{I^i_{jk}}\boldsymbol{\delta}_{I^i_{jk}} - \delta_{[sjk]} \sim \mathcal{N}(-\delta_{[sjk]}, \boldsymbol{\delta}_{I^i_{jk}}{}^{\top}\tilde{\mathbf{z}}^{-\top}_{I^i_{jk}}\tilde{\mathbf{z}}^{-1}_{I^i_{jk}}\boldsymbol{\delta}_{I^i_{jk}})$, in which the variance is uniformly bounded from above and below for all $\tilde{\mathbf{z}}_{I_{jk}} \in B^1_{jk}$. It follows that

$$\inf_{\tilde{\mathbf{z}}_{I_{jk}} \in B^1_{jk}} P^{\star}\{G_{jk}(\tilde{\mathbf{z}}_{I_{jk}})\} > 0. \tag{C.26}$$

Thus, by restricting the integrals on the RHS of Equation C.23 to $G_{jk}(\tilde{\mathbf{z}}_{I_{jk}})$, we are able to obtain an uniform lower bound of $\bar{\varphi}_{jk}(u, \boldsymbol{\theta}_{jk}, \mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[sjk]})$ for all $\tilde{\mathbf{z}}_{I_{jk}} \in B^1_{jk}$.

Our final task is to find $B^2_{jk} \subset \mathbb{R}^{r^2_j}$ such that $P\{\tilde{\mathbf{Z}}^{\star}_{I_{jk}} \in B^2_{jk}\} > 1 - \varepsilon/4$, and that $P\{C_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}, y_{[1jk]}) | E^{\delta}_{\mathbf{y}_I, \mathbf{k}_I}(\mathbf{z}_I)\}$ has a uniform lower bound for all $\tilde{\mathbf{z}}_{I_{jk}} \in B^2_{jk}$. Here, we only prove the statement for $r = 1$, and we conjecture that an extended argument can be established for $r > 1$.

When $r = 1$, $|I_{jk}| = 2$; without loss of generality, let I_{jk} be the first two observations. We fix j and k, and for simplicity denote the two normal vectors corresponding to the first two observations by $\mathbf{n}_1 = \omega_1(\delta_1\ z_1)^{\top}$ and $\mathbf{n}_2 = \omega_2(\delta_2\ z_2)^{\top}$, in which $\omega_1, \omega_2 = \pm 1$ and $\delta_1, \delta_2 \in \{0, 1\}$. We now discuss cases for different combinations of ω_1, ω_2, δ_1 and δ_2 values, and establish in each case that the joint probability of $Z^{\star}_{[sjk]} = -\gamma_1\omega_1 z_1 - \gamma_2\omega_2 z_2$ and $\delta_{[sjk]} = -\gamma_1\omega_1\delta_1 - \gamma_2\omega_2\delta_2$, $\gamma_1, \gamma_2 > 0$, is uniformly bounded from below for $(z_1\ z_2)^{\top} \in B^2_{jk} = \{(z_1\ z_2)^{\top} : |z_1 - z_2| \geq \lambda', |z_1| \leq L, |z_2| \leq L\}$ for every $\lambda', L > 0$.

Case 1: $\omega_1 = 1$, $\omega_2 = 1$, $\delta_1 = 1$, *and* $\delta_2 = 1$. Set $\omega_{[sjk]} = -1$ and $\delta_{[sjk]} = 1$, which happens with positive probability provided the data-generating parameter values are in the interior of the parameter space. Then, $\mathbf{N}^{\star}_{[sjk]} = -\gamma_1\mathbf{n}_1 - \gamma_2\mathbf{n}_2$ implies $\gamma_1 + \gamma_2 = 1$ and $Z^{\star}_{[sjk]} = -\gamma_1 z_1 - \gamma_2 z_2$, i.e., $Z^{\star}_{[sjk]}$ falls in the line segment between $-z_1$ and $-z_2$. For all $(z_1\ z_2)^{\top} \in B^2_{jk}$, $P\{\min\{-z_1, -z_2\} \leq Z^{\star}_{[sjk]} \leq \max\{-z_1, -z_2\}\} > \Phi(L) - \Phi(L + \lambda')$.

Case 2: $\omega_1 = 1$, $\omega_2 = -1$, $\delta_1 = 1$, *and* $\delta_2 = 1$. In this case, the constraints are $\omega_{[sjk]}\delta_{[sjk]} = -\gamma_1 + \gamma_2$ and $\omega_{[sjk]}Z^{\star}_{[sjk]} = -\gamma_1 z_1 + \gamma_2 z_2$. When $\omega_{[sjk]} = 1$ and $\delta_{[sjk]} = 1$, $\gamma_2 = 1 + \gamma_1$. It follows that $Z^{\star}_{[sjk]} = \gamma_1(z_2 - z_1) + z_2$, which is greater than z_2 when $z_2 > z_1$ and less than z_2 when $z_2 < z_1$. Then, both $P\{Z^{\star}_{[sjk]} < z_2\}$ and $P\{Z^{\star}_{[sjk]} > z_2\}$ are uniformly greater than $1 - \Phi(L)$ for all $(z_1\ z_2)^{\top} \in B^2_{jk}$. When $\omega_{[sjk]} = 1$ and $\delta_{[sjk]} = 0$, we have $\gamma_2 = \gamma_1$, and thus $Z^{\star}_{[sjk]} = \gamma_1(z_2 - z_1)$; given $(z_1\ z_2)^{\top} \in B^2_{jk}$, it only restricts the sign under a standard normal measure. Other combinations of $\omega_{[sjk]}$ and $\delta_{[sjk]}$ values can be dealt with in a similar fasion.

Case 3: $\omega_1 = 1$, $\omega_2 = 1$, $\delta_1 = 1$, *and* $\delta_2 = 0$. It requires $\omega_{[sjk]} = -1$, $\delta_{[sjk]} = 1$, and $\gamma_1 = 1$. So $Z^{\star}_{[sjk]} = z_1 + \gamma_2 z_2$, which is greater than z_1 when $z_2 > 0$, and smaller than z_1 when $z_2 < 0$. Similar as before, both $P\{Z^{\star}_{[sjk]} < z_1\}$ and $P\{Z^{\star}_{[sjk]} > z_1\}$ are uniformly greater than $1 - \Phi(L)$ for all $(z_1\ z_2)^{\top} \in B^2_{jk}$.

All other combinations of ω_1, ω_2, δ_1 and δ_2 values are reflections of the three cases having been discussed. Also note that δ_1 and δ_2 cannot be both zero, otherwise no vertex is determined from the two observations. Altogether we have shown that $P\{C_{jk}(\mathbf{y}_{I_{jk}}, \tilde{\mathbf{z}}_{I_{jk}}) | E^{\delta}_{\mathbf{y}_I, \mathbf{k}_I}(\mathbf{z}_I)\}$ is uniformly bounded from below for $\tilde{\mathbf{z}}_{I_{jk}} \in B^2_{jk}$. The proof is now complete for $r = 1$.

$\square$

APPENDIX D: PREDICTIVE INFERENCE

Proof of Proposition 2. Fix $\varepsilon > 0$. Using the result expressed in Equation 2.22, the dominated convergence theorem, and the fact that continous functions are uniformly continuous in compacta, we could select $\delta > 0$ such that $\int \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta} |h(\mathbf{t},\boldsymbol{\theta}) - h(\mathbf{t},\boldsymbol{\theta}_0)|d\mu < \varepsilon$. Then

$$\begin{aligned}
\int |h_n(\mathbf{t}|\mathbf{Y}) - h(\mathbf{t},\boldsymbol{\theta}_0)|d\mu &\leq \int \left[\int_\Theta |h(\mathbf{t},\boldsymbol{\theta}) - h(\mathbf{t},\boldsymbol{\theta}_0)|g_n(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta}\right] d\mu \\
&= \int_\Theta \left[\int |h(\mathbf{t},\boldsymbol{\theta}) - h(\mathbf{t},\boldsymbol{\theta}_0)|d\mu\right] g_n(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta} \\
&= \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta} \left[\int |h(\mathbf{t},\boldsymbol{\theta}) - h(\mathbf{t},\boldsymbol{\theta}_0)|d\mu\right] g_n(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta} \\
&\quad + \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \left[\int |h(\mathbf{t},\boldsymbol{\theta}) - h(\mathbf{t},\boldsymbol{\theta}_0)|d\mu\right] g_n(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta} \\
&\leq \varepsilon + 2\int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} g_n(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta}. \qquad \text{(D.1)}
\end{aligned}$$

By Equation 2.21, $P_{\boldsymbol{\theta}_0}\left\{\int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} g_n(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta} \leq \varepsilon\right\} \to 1$ as $n \to \infty$, which concludes the proof. □

APPENDIX E: FIDUCIAL PREDICTIVE CHECK

E.1 Asymptotic covariance with the sample score function

As discussed earlier, the probability for each individual response pattern $\mathbf{y}_i$ is given by $f(\boldsymbol{\theta}, \mathbf{y}_i)$, i.e., Equation 1.2. Denote by $\mathbf{f}(\boldsymbol{\theta})$ be the collection of all $\prod_{j=1}^{m} K_j$ response pattern probabilities, and $\mathbf{p}$ the observed counterpart. Then $n\mathbf{p} \sim \text{Multinomial}(n, \mathbf{f}(\boldsymbol{\theta}))$. Also let $\mathbf{D}(\boldsymbol{\theta}) = \text{diag}(\mathbf{f}(\boldsymbol{\theta}))$ be a diagonal matrix, and $\boldsymbol{\Delta}(\boldsymbol{\theta}) = \partial \mathbf{f}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^\top$ be the Jacobian matrix. With these matrix notations, the sample score function evaluated at the data-generating parameters $\boldsymbol{\theta}$, previously expressed in Equation 2.15 and 2.16, can be rewritten as

$$\mathbf{S}_n = \sqrt{n}\boldsymbol{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{p} = \boldsymbol{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}\sqrt{n}[\mathbf{p} - \mathbf{f}(\boldsymbol{\theta})], \tag{E.1}$$

in which the last equality is because of the fact that $\mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{f}(\boldsymbol{\theta}) = \mathbf{1}$, and $\mathbf{1}^\top \boldsymbol{\Delta}(\boldsymbol{\theta}) = \mathbf{0}$ for any $\boldsymbol{\theta}$. Also write the Fisher information matrix as

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \boldsymbol{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}\boldsymbol{\Delta}(\boldsymbol{\theta}). \tag{E.2}$$

Test statistics of interest in the current work are all linear combinations of the observed proportions in $\mathbf{p}$; we express it generally as $T = \mathbf{b}^\top \mathbf{p}$, in which $\mathbf{b}$ is a constant vector. It follows that $\partial \nu(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^\top = \mathbf{b}^\top \boldsymbol{\Delta}(\boldsymbol{\theta})$. Next, we verify that the asymptotic covariance between $\sqrt{n}T$ and $\mathbf{S}_n$ is given by $\mathbf{b}^\top \boldsymbol{\Delta}(\boldsymbol{\theta})$.

By the multivariate Central Limit Theorem,

$$\sqrt{n}[\mathbf{p} - \mathbf{f}(\boldsymbol{\theta})] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}(\boldsymbol{\theta})) \tag{E.3}$$

under the data-generating model, in which $\mathbf{\Gamma}(\boldsymbol{\theta}) = \mathbf{D}(\boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\theta})\mathbf{f}(\boldsymbol{\theta})^\top$. It follows that

$$\begin{aligned}\begin{pmatrix} \sqrt{n}[T - \nu(\boldsymbol{\theta})] \\ \mathbf{S}_n \end{pmatrix} &= \begin{pmatrix} \mathbf{b}^\top \\ \mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1} \end{pmatrix} \sqrt{n}[\mathbf{p} - \mathbf{f}(\boldsymbol{\theta})] \\ &\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{b}^\top \mathbf{\Gamma}(\boldsymbol{\theta})\mathbf{b} & \mathbf{b}^\top \mathbf{\Gamma}(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{\Delta}(\boldsymbol{\theta}) \\ \mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{\Gamma}(\boldsymbol{\theta})\mathbf{b} & \mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{\Gamma}(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{\Delta}(\boldsymbol{\theta}) \end{pmatrix}\right),\end{aligned} \tag{E.4}$$

in which the asymptotic covariance is $\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{\Gamma}(\boldsymbol{\theta})\mathbf{b} = \mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{b} = \partial \nu(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

The foregoing conclusion holds for centered statistics $T - \nu(\tilde{\boldsymbol{\theta}})$ as well, provided the estimator $\tilde{\boldsymbol{\theta}}$ satisfies

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = [\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{W}\mathbf{\Delta}(\boldsymbol{\theta})]^{-1}\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{W}[\mathbf{p} - \mathbf{f}(\boldsymbol{\theta})] + o_p(1), \tag{E.5}$$

in which $\mathbf{W}$ is some constant weighted matrix. Equation E.5 can be established for a general class of weighted least square estimators (see Maydeu-Olivares, 2006), including the ML estimator (with $\mathbf{W} = \mathbf{D}(\boldsymbol{\theta})^{-1}$). It follows that

$$\begin{aligned}\sqrt{n}[T - \nu(\tilde{\boldsymbol{\theta}})] &= \sqrt{n}[T - \nu(\boldsymbol{\theta})] - \sqrt{n}[\nu(\tilde{\boldsymbol{\theta}}) - \nu(\boldsymbol{\theta})] \\ &= \sqrt{n}\mathbf{b}^\top \left\{\mathbf{I} - [\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{W}\mathbf{\Delta}(\boldsymbol{\theta})]^{-1}\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{W}\right\}[\mathbf{p} - \mathbf{f}(\boldsymbol{\theta})] + o_p(1).\end{aligned} \tag{E.6}$$

The asymptotic covariance between $\sqrt{n}[T - \nu(\tilde{\boldsymbol{\theta}})]$ and $\mathbf{S}_n$ is then

$$\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}\mathbf{\Gamma}(\boldsymbol{\theta})\left\{\mathbf{I} - [\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{W}\mathbf{\Delta}(\boldsymbol{\theta})]^{-1}\mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{W}\right\}^\top \mathbf{b} = \mathbf{\Delta}(\boldsymbol{\theta})^\top \mathbf{b},$$

obtained via a similar computation as before.

E.2 Normal approximation of the likelihood

By the asymptotic normality of the observed response-pattern proportions (Equation E.3),

$$\sqrt{n}\mathbf{b}^\top[\mathbf{p}-\mathbf{f}(\boldsymbol{\theta})] \overset{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{b}^\top\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{b}). \tag{E.7}$$

The likelihood of $T = \mathbf{b}^\top\mathbf{p}$ can then be approximated by the density function of $\mathbf{N}(\mathbf{b}^\top\mathbf{f}(\boldsymbol{\theta}), \mathbf{b}^\top\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{b}/n)$.

REFERENCES

Aitchison, J. and Dunsmore, I. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, 17(3):251–269.

Asparouhov, T. and Muthén, B. (2012). Comparison of computational methods for high dimensional item factor analysis. Unpublished manuscript retrieved from *www.statmodel.com*.

Bartlett, M. S. (1965). R. A. fisher and the last fifty years of statistical methodology. *Journal of the American Statistical Association*, 60(310):395–409.

Bayarri, M. and Berger, J. O. (1999). Quantifying surprise in the data and model verification. *Bayesian statistics*, 6:53–82.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Bernaards, C. A. and Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, 65:676–696.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical theories of mental test scores*, pages 395–479. Addison-Wesley, Reading, MA.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35(2):179–197.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64(2):153–168.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36(1):111–150.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, 61(2):309–329.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75(1):33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3):307–335.

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

Cai, L. and Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66(2):245–276.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., and Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, 59(1):173–194.

Casella, G. and Berger, R. L. (2002). *Statistical inference (2nd Ed.)*, volume 2. Duxbury, Pacific Grove, CA.

Cisewski, J. and Hannig, J. (2012). Generalized fiducial inference for normal linear mixed models. *The Annals of Statistics*, 40(4):2102–2127.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3):315–345.

Crawford, C. B. and Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, 35(3):321–332.

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, 36(1):1–34.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):205–247.

Dempster, A. P. (2008). The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Doucet, A., De Freitas, N., and Gordon, N. (2001). *An introduction to sequential Monte Carlo methods*. Springer, New York.

Duong, T. (2014). *ks: Kernel smoothing.* R package version 1.9.3. http://CRAN.R-project.org/package=ks.

E, L., Hannig, J., and Iyer, H. K. (2009). Fiducial generalized confidence interval for median lethal dose (LD50). Unpublished manuscript.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3):474–497.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statistical Science*, pages 95–114.

Ferrando, P. J. and Lorenzo-Seva, U. (2001). Checking the appropriateness of item response theory models by predicting the distribution of observed scores: The program EO-Fit. *Educational and psychological measurement*, 61(5):895–902.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1):69–78.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, Series A*, 222:309–368.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.

Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London. Series A*, 139(838):343–348.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398.

Fisher, R. A. (1945). The logical inversion of the notion of the random variable. *Sankhyā: The Indian Journal of Statistics*, 7(2):129–132.

Forero, C. G., Maydeu-Olivares, A., and Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16(4):625–641.

Fraser, D. A. S. (1968). *The structure of inference.* John Wiley & Sons, New York.

Geisser, S. (1993). *Predictive inference: An introduction.* Chapman and Hall, New York.

Gelman, A. et al. (2013). Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electronic Journal of Statistics*, 7:2595–2602.

Ghosh, J. and Bickel, P. J. (1990). A decomposition for the likelihood ratio statistic and the bartlett correction: A Bayesian argument. *Annals of Statistics*, 18(3):1070–1090.

Gibbons, R. D. and Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3):423–436.

Glas, C. A. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, 53(4):525–546.

Gunsjö, A. (1994). *Faktoranalys av ordinala variabler.* Studia statistica Upsaliensia. Acta Universitatis Upsaliensis, Stockholm, Sweden.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 29(1):83–100.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. In Clogg, C. C., editor, *Sociological methodology*, pages 193–211. American Sociological Association, Washington, D.C.

Haberman, S. J. (2006). Adaptive quadrature for item response models. *ETS Research Report Series*, 2006(2):1–10.

Haberman, S. J. and Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of the American Statistical Association*, 108(504):1435–1444.

Hambleton, R. and Han, N. (2004). Assessing the fit of irt models: Some approaches and graphical displays. Presentation at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA, USA.

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19(2):491.

Hannig, J. (2013). Generalized fiducial inference via discretization. *Statistica Sinica*, 23(2):489–514.

Hannig, J., Lai, R. C., and Lee, T. C. (2014). Computational issues of generalized fiducial inference. *Computational Statistics & Data Analysis*, 71:849–858.

Hannig, J. and Lee, T. C. (2009). Generalized fiducial inference for wavelet regression. *Biometrika*, 96(4):847–860.

Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J.-S., Varni, J. W., Yeatts, K., and DeWalt, D. A. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, 19(4):595–607.

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika*, 38(4):593–604.

Joe, H. and Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, 75(3):393–419.

Lawless, J. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer-Verlag, New York.

Levy, R., Mislevy, R. J., and Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33(7):519–537.

Li, Z. and Cai, L. (2012). Summed score based fit indices for testing latent variable distribution assumption in IRT. Presentation at the International Meeting of the Psychometric Society, Lincoln, NE, USA.

Liu, Y. and Hannig, J. (2014). Generalized fiducial inference for binary logistic item response models. Under review.

Liu, Y. and Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, 49(4):354–371.

Liu, Y. and Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology*, 67(3):496–513.

Lord, F. M. and Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, 8(4):453–461.

Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501):301–313.

Martin, R., Zhang, J., and Liu, C. (2010). Dempster–Shafer theory and statistical inference with weak beliefs. *Statistical Science*, 25(1):72–87.

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, 71(1):57–77.

Maydeu-Olivares, A. and Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables: a unified framework. *Journal of the American Statistical Association*, 100(471):1009–1020.

Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71(4):713–732.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34(1):100–117.

Meng, X.-L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, 91(435):1254–1267.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43(4):551–560.

Muthén, L. K. and Muthén, B. O. (2012). *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 31(3):705–741.

Neale, M. C. and Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior genetics*, 27(2):113–120.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Pal Majumder, A. and Hannig, J. (2015). On quantile matching fiducial distributions. Manuscript in preparation.

Patz, R. J. and 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.

Plummer, M. (2013a). *JAGS Version 3.4.0 user manual.* http://sourceforge.net/projects/mcmc-jags/files/Manuals/3.x/.

Plummer, M. (2013b). *rjags: Bayesian graphical models using MCMC.* R package version 3-10. http://CRAN.R-project.org/package=rjags.

Ralston, A. (1965). *A First Course in Numerical Analysis.* McGraw-Hill, New York.

Rao, C. (1973). *Linear Statistical Inference and Its Application (2nd Ed.).* Wiley, New York.

Reckase, M. (2009). *Multidimensional item response theory.* Springer, New York.

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, 61(3):509–528.

Robins, J. M., van der Vaart, A. W., and Ventura, V. (2000). Asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association*, 95(452):1143–1156.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications*. Guilford, New York.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph No. 17*. Richmond, VA: Psychometric Society.

Schilling, S. and Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *psychometrika*, 70(3):533–555.

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, 42(4):375–394.

Sinharay, S., Johnson, M. S., and Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4):298–321.

Thissen, D., Pommerich, M., Billeaud, K., and Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, 19(1):39–49.

Thissen, D. and Steinberg, L. (2010). Using item response theory to disentangle constructs at different levels of generality. In Embretson, S., editor, *Measuring psychological constructs: Advances in modelbased approaches*, pages 123–144. American Psychological Association, Washington, DC.

Thissen, D., Steinberg, L., and Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, 27(1):77–83.

Thissen, D. and Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational and Behavioral Statistics*, 15(2):113–128.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press, New York.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Chapman and Hall, London.

Wang, C., Hannig, J., and Iyer, H. K. (2012). Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142(7):1980–1990.

Wirth, R. and Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, 12(1):58.

Xie, M., Liu, R., Chang, K., and Chen, R. (2014). Prediction with confidence: A general and unified framework for prediction. Presentation at [where?].

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81(1):3–39.

Yuan, K.-H., Cheng, Y., and Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, 79(2):232–254.

Zabell, S. L. (1992). R. A. Fisher and fiducial argument. *Statistical Science*, 7(3):369–387.