Autoencoded sparse Bayesian in-IRT factorization, calibration, and amortized inference for the Work Disability Functional Assessment Battery

Joshua C. Chang,[1] **Carson C. Chow,**[2] **Julia Porcino**[1]

[1]Epidemiology and Biostatistics Section, Rehabilitation Medicine, Clinical Center, National Institutes of Health, USA
[2]Mathematical Biology Section, LBM, NIDDK, National Institutes of Health, USA

Abstract

The Work Disability Functional Assessment Battery (WD-FAB) is a multidimensional item response theory (IRT) instrument designed for assessing work-related mental and physical function based on responses to an item bank. In prior iterations it was developed using traditional means – linear factorization and null hypothesis statistical testing for item partitioning/selection, and finally, posthoc calibration of disjoint unidimensional IRT models. As a result, the WD-FAB, like many other IRT instruments, is a posthoc model. Its item partitioning, based on exploratory factor analysis, is blind to the final nonlinear IRT model and is not performed in a manner consistent with goodness of fit to the final model. In this manuscript, we develop a Bayesian hierarchical model for self-consistently performing the following simultaneous tasks: scale factorization, item selection, parameter identification, and response scoring. This method uses sparsity-based shrinkage to obviate the linear factorization and null hypothesis statistical tests that are usually required for developing multidimensional IRT models, so that item partitioning is consistent with the ultimate nonlinear factor model. We also analogize our multidimensional IRT model to probabilistic autoencoders, specifying an encoder function that amortizes the inference of ability parameters from item responses. The encoder function is equivalent to the "VBE" step in a stochastic variational Bayesian expectation maximization (VBEM) procedure that we use for approxiamte Bayesian inference on the entire model. We use the method on a sample of WD-FAB item responses and compare the resulting item discriminations to those obtained using the traditional posthoc method.

Introduction

The United States Social Security Administration (SSA), the administrator of the largest federal disability benefits program in the US, is tasked with determining the eligibility of approximately two million applicants annually for benefits. Determining a person's ability to engage in work is difficult. Additionally, capacity for work in individuals may change over time and tools are needed for assessing these changes, for instance in support of return-to-work programs.

The statutory definition of disability requires determining whether a person's ability to work is limited by the presence of medical conditions. Modern models of disability such as the World Health Organization (WHO)'s International Classification of Functioning, Disability and Health (ICF) view disability as a biopsychosocial construct [Brandt and Smalligan 2019], contextualizing disability as an interaction between the functional capability of individuals and the needs and opportunities of their environment. Assessing disability through this lens is resource-intensive, motivating the development of tools to aid in the adjudication process by objectively characterizing the functional ability of an applicant. The Work Disability Functional Assessment Battery (WD-FAB) is such a tool for understanding work-related physical and mental function of individuals relative to the working adult population based on responses to a battery of items.

Work Disability Functional Assessment Battery

The WD-FAB was developed by researchers at the Boston University Health and Disability Research Institute (BU) in collaboration with the National Institutes of Health (NIH) and with the support of the Social Security Administration (SSA). The intended use of this instrument was to provide more standardized and consistent information about an individual's functional abilities to help improve the efficiency and reliability of SSA's disability adjudication process. The WD-FAB provides eight scores across two domains of physical and mental function that are relevant to a person's ability to work. The ICF is one of the key frameworks for the content of these domains. The ICF includes categories for classifying function at the cellular, organ, and whole person level, referred to as activities and participation. The WD-FAB focuses on measuring activity.

The development of the WD-FAB is detailed in several papers [Marfeo et al. 2018, Meterko et al. 2015, Jette et al. 2019, Porcino et al. 2018]. Subject matter experts used the ICF, discipline-specific frameworks, and existing functional assessment instruments, to develop a bank of approximately 300 physical and 300 mental items that pertain to work-related function. They further divided the physical items into four subcategories (PD - physical demands, PDR - physical demands replenishment, PF - physical function, DA - daily activities) and mental items into three categories (CC - community cognition, II - interpersonal interactions, BH - behavioral health) based on how they relate to ICF content, however, they did not use this categorization in their analyses.

The item banks consist of questions that ask about a range of everyday type activities, such as vacuuming, emptying a dishwasher, painting a room, walking a block, turning a door knob, speaking to someone on the phone, and managing under stress. Valid responses were graded on either four or five option Likert scales with ordinal responses such as agreement (Strongly agree, Agree, Disagree, Strongly disagree), or frequency (Never, Rarely, Sometimes, Often, Always). Overall, these studies collected item responses from a total of 11,901 subjects sampled from claimants for disability benefits as well as working-age adults who represent the general population of the United States.

The developers of the WD-FAB then followed the PROMIS guidelines [Fries et al. 2014, Cella et al. 2007, DeWalt et al. 2007] for measure development. They first performed exploratory factor analysis on the response matrix, the output of which is a collection of linear factors with dense loadings. Then, they extracted the first four factors. For each factor they used stepwise rejection of items based on null hypothesis statistical testing, thresholding to select a subset of items for each dimension. They then assessed validity of unidimensionality of each of the item subsets using confirmatory factor analysis. Finally, they calibrated independent predictive models for how a person may respond to each subset of items. Besides the arbitrariness of the thresholds used for item selection, a major weakness of this procedure is in how the scale factorization is not performed in a way that is mindful of the final nonlinear model. Alternate item factorizations that do not arise from the linear factor analyses are prematurely excluded, uncertainty in the factorization is not propagated, and the IRT model is effectively a posthoc analysis. For this reason, we will refer to the prior WD-FAB instrument as the posthoc WD-FAB.

Item Response Theory

Item response theory (IRT), a generative latent-variable modeling framework, is the dominant statistical paradigm for quantifying assessments. Some applications of IRT include standardized testing including Graduate Record Exam (GRE) [Kingston and Dorans 1982], the Scholastic Aptitude Test (SAT) [Carlson and von Davier 2013] and the Graduate Management Admission Test [Kingston, Leary, and Wightman 1985]. Other applications of IRT include medical/psychological assessments such as activities of daily living [Fieo et al. 2010], quality of life [Bilbao et al. 2014], and personality tests [Goldberg 1992, Bore et al. 2020, Saunders and Ngo 2017, DeYoung et al. 2016, Funke 2005, Spence, Owens, and Goodyer 2012]. IRT also serves as the theoretical basis for the WD-FAB [Meterko et al. 2015, Marfeo et al. 2016, 2019, Chang et al. 2022b].

In item response theory (IRT), a person's test responses are modeled as an interaction between personal traits (also called abilities) and item-specific parameters. The item parameters relate to the difficulty of the item and the discrimination of the item, or the degree to which the question's responses are determined by personal traits. The two types of attributes work together to predict an individual's responses via item response functions. Conversely, a set of responses may be statistically inverted in order to estimate an individual's ability. The central idea behind IRT is to use person-specific abilities in order to make comparisons between people in a population.

Multidimensional instruments: For complex phenomena, such as disability, a single scalar factor cannot adequately describe how a person would respond to a diverse set of items [Yuker 1994]. In these cases, one can develop a multidimensional IRT model (MIRT). Like in the WD-FAB, MIRT models are typically composed of ensembles of unidimensional models, developed using the stepwise procedure of linear factor analyses followed by calibration of disjoint nonlinear unidimensional IRT models. Each step of in these procedures require statistical decisions – in practice these decisions are performed using arbitrary P-value cutoffs. Ultimately, the resulting MIRT model is a post-hoc model, and the initial item partitioning steps are not performed with consideration to how well the final IRT model fits the data. This issue is problematic because abilities are derived from response patterns with the assumption that the model accurately represents the response patterns of the population.

Novelty and relation to prior work

In this manuscript, we re-examine the methodology behind the WD-FAB and highlight how modern statistical techniques can improve it. Specifically, we show that probabilistic autoencoders can serve as a complete pipeline for translating survey responses into a set of interpretable indicators about functional ability, with greater predictive power than existing techniques. Prior work has noted that IRT models are inherently similar to probabilistic autoencoders [Chang, Vattikuti, and Chow 2019, Converse, Curi, and Oliveira 2019, Converse et al. 2021], where an encoder performs amortized inference on person-specific abilities. Viewing IRT models as a specific category of autoencoders motivates extensions to standard IRT methods. Prior work has not constrained the encoder function so that it does not modify the statistics of the decoder. Our main methodological contributions are: 1. the adaptation of Bayesian sparsity methods to perform factorization directly in an IRT model 2. the specification of an encoder function, fully specified by the decoder, that defines the 'VBE"-step of a variational Bayesian expectation maximization algorithm – and in doing so does not modify the statistics of the decoder.

Methods

Notation

The response data takes the form of a $P \times I$ matrix, where P corresponds to the number of people and i corresponds to the number of items. We denote this matrix $\mathbf{X}$. Unless otherwise stated, we will index rows in this matrix using the symbol p and columns of this matrix using the symbol i. Each entry of this matrix is a valid response from the set $\{1, 2, \dots K\}$, where $K = 5$ for the WD-FAB.

Parameters in the model may vary according to person p, item i, and latent dimension d. We generally use bold letters for denoting the collection of all values of a parameter (e.g., $\boldsymbol{\theta}$ denotes the collection of all ability parameters). For specific slices of a parameter we use bold lowercase symbols

– for example, $\boldsymbol{\theta}_p = (\theta_p^{(1)}, \theta_p^{(2)}, \ldots, \theta_p^{(D)})$ corresponds to a vector of all ability parameters for person p.

In this manuscript we will denote the collection of all model parameters as $\boldsymbol{\Gamma}$, the collection of all ability parameters as $\boldsymbol{\theta}$, and the collection of all model parameters except the ability parameters as $\boldsymbol{\Gamma} \setminus \boldsymbol{\theta}$.

Multidimensional IRT as a probabilistic autoencoder

The unidimensional ability scale graded response model (GRM) [Samejima 1969] is an item response theory (IRT) model for ordinal responses. The GRM states that the probability that person p responds to item i with a choice j is

$$\begin{aligned}\Pr(X_{pi} = j|\theta_p, \boldsymbol{\tau}_i, \lambda_i) = \Pr(X_{pi} \geq j|\theta_p, \tau_{ij}, \lambda_i) \\ - \Pr(X_{pi} \geq j+1|\theta_p, \tau_{i,j+1}, \lambda_i),\end{aligned} \quad (1)$$

where we define the GRM in its probit variation, utilizing the complementary cumulative distribution function for the unit normal distribution Φ, so that

$$\Pr(x \geq j|\theta, \tau_j, \lambda) = \begin{cases} \Phi(\lambda(\theta - \tau_j)) & j \in [2, K] \\ 1 & j \leq 1 \\ 0 & j > K \end{cases}. \quad (2)$$

Within the model, $\boldsymbol{\tau}_i = (\tau_{i,1}, \tau_{i,2}, \ldots)$ where $\tau_{i,j+1} \geq \tau_{i,j}$ are item difficulty parameters. The ability parameters θ_p map a person's ability ranking within their population to a real-valued scale. The remaining parameters λ_i are item discrimination parameters – they represent how informative a particular item is to the scale, and visa versa. When the discrimination goes to zero, then an item is effectively decoupled from the scale.

Extending the GRM to multiple ability scale dimensions, we define a discrimination-weighted mixture GRM:

$$\begin{aligned}&\Pr(X_{pi} = j|\{\theta_p^{(d)}\}_d, \{\{\tau_{ij}^{(d)}\}_j\}_d, \{\lambda_i^{(d)}\}_d) \\ &\quad = \sum_{d=1}^{D} w_{id} \Pr(X_{pi} = j|\theta_p^{(d)}, \tau_{i,j}^{(d)}, \tau_{i,j+1}^{(d)}, \lambda_i^{(d)}) \\ &w_{id} = \lambda_i^{(d)} / \sum_{d=1}^{D} \lambda_i^{(d)},\end{aligned} \quad (3)$$

noting that $\lambda_i^{(d)} = 0 \Rightarrow w_{id} = 0$; this form of weighting allows us to extend the GRM to a mixture model without needing to introduce any new free parameters.

This multidimensional IRT model assumes that each person's ability consists of D scales. The parameter $\theta_p^{(d)}$ is the ability for person p on scale d and $\lambda_i^{(d)}$ is the discrimination of item i with respect to scale d. It strongly resembles probabilistic matrix factorization and other probabilistic autoencoders. When trained on a sample of individuals and their responses, the model in Eq. 3 defines a total likelihood

$$\begin{aligned}&\pi(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\tau}) = \\ &\quad \prod_p \prod_i \Pr(X_{pi} = j|\{\theta_p^{(d)}\}_d, \{\{\tau_{ij}^{(d)}\}_j\}_d, \{\lambda_i^{(d)}\}_d)^{\delta_{X_{pij}}}\end{aligned} \quad (4)$$

that takes as input a high-dimensional response matrix $\mathbf{X} = (X_{pi})$ and derives a lower dimensional representation matrix $\boldsymbol{\theta} = (\theta_p^{(d)})_{pd}$, where the $p-th$ row in the representation matrix corresponds to the multidimensional ability for person p. The weight matrix $\mathbf{W} = (w_{id})_{id}$ decodes the ability components for an individual into probability masses for their item responses. This matrix serves the same purpose as a factor loading matrix in principle components analysis. Our objective is to obtain this matrix in-unison with other model parameters that directly relate to how individuals might respond to a given item battery.

Sparse factorization: By determining the matrix $\mathbf{W}$, we factor the items into multiple scales. For improving the interpretability of these factorizations, we seek sparse factors, as in sparse probabilistic matrix factorization [Gopalan et al. 2014, Mnih and Salakhutdinov 2008, Chang, Vattikuti, and Chow 2019, Chang et al. 2020]. We accomplish this goal by using the horseshoe priors [Carvalho, Polson, and Scottt 2010, Bhadra et al. 2015, 2019] on the discrimination parameters on a scale-by-scale basis. Our overall hierarchical probabilistic model for simultaneous factorization and calibration of the multidimensional GRM is specified:

$$\begin{aligned}-\log \pi\left(\boldsymbol{\lambda}|\boldsymbol{\xi}_i, \boldsymbol{\kappa}, \boldsymbol{\eta}\right) = \sum_{i,d}\left[\frac{\lambda_i^{(d)}}{2(\xi_i^{(d)}\kappa^{(d)})^2} + \log(\xi_i^{(d)}\kappa^{(d)})\right] \\ + \sum_{i,d} \log \pi(\mathbf{w}_i|\eta_i) + \text{const}\end{aligned} \quad (5a)$$

$$\pi(\mathbf{w}_i|\eta_i) \propto \exp\left(\eta_i^{-1} \sum_d w_i^{(d)} \log w_i^{(d)}\right) \quad (5b)$$

$$\sigma_i^{(d)} = \xi_i^{(d)}\kappa^{(d)} \quad \xi_i^{(d)} \sim \text{cauchy}^+(0, 1) \quad (5c)$$

$$\eta_i \sim \text{normal}^+(0, \eta_0) \quad \kappa^{(d)} \sim \text{cauchy}^+(0, \kappa_0^{(d)}) \quad (5d)$$

$$\tau_{i,2}^{(d)} \sim \text{normal}(\mu_i^{(d)}, 1) \quad \tau_{i,j}^{(d)}|\tau_{i,j-1}^{(d)} \sim \text{normal}^+(\tau_{i,j-1}^{(d)}, 1) \quad (5e)$$

$$\mu_i^{(d)} \sim \text{normal}(0, 1) \quad \theta_p^{(d)} \sim \text{normal}(0, 1) \quad (5f)$$

where the discrimination parameters $\lambda_i^{(d)}$ are each constrained to non-negativity and we define a per-item entropy penalty in Eq. 5b.

The dimension-wise horseshoe priors on the discrimination parameters encourage scale sparsity, and the item-wise entropy priors encourage items too load into a small number of scales.

Hyperparameter scaling: If the apriori expectation is that the dominant scale (on a per-item basis) holds weight $q \approx 1$ and the other weights are uniform, then $\eta_i = -q\log(q) - (1-q)\log((1-q)/(D-1))$ is an appropriate value for the scaling factor η_i. In this manuscript we use $q = 0.8$.

The parameters $\kappa^{(d)}$ control the overall amount of sparsity in each scale dimension. For partitioning a set of I items into D dimensions, we expect each dimension to have approximately I/D nonzero terms. As in Piironen and Vehtari [2017b] and van der Pas, Kleijn, and van der Vaart [2014], we derived an approximate scaling on $\kappa^{(d)}$ based on asymptotic approximation of the bias in the posterior mode. This approximation suggests the scaling $\kappa_0^{(d)} = \sqrt{\Delta(D, K, I)/P}$ where $\Delta(D, K, I)$ is a constant derived in

the Supplemental Methods.

Autoencoded amortized inference

The intended use of item response models like the WD-FAB is to use them to score new response patterns, effectively reducing high-dimensional response vectors to low-dimensional ability representations. In probabilistic autoencoders, the mapping is known as the *encoder*. As part of training the generative hierarchical Bayesian model of Eq. 5 (the decoder), we also learn the encoder function $\text{encoder}(\mathbf{X}_p) = q_{\boldsymbol{\theta}_p} : \mathbb{R}^D \to \mathbb{R}^+$ where $q_{\boldsymbol{\theta}^p}$ is an approximation of the marginal density $\pi(\boldsymbol{\theta}_p|\mathbf{X}_p)$. This surrogate density can then be used for approximating posterior expectations

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}|\mathbf{X}}\left(g(\boldsymbol{\theta}_p)|\mathbf{X}_p, \mathbf{X}\right) \\
&= \int g(\boldsymbol{\theta}_p) \left(\iint \pi(\boldsymbol{\theta}_p|\boldsymbol{\lambda}, \boldsymbol{\tau}, \mathbf{X}_p)\pi(\boldsymbol{\lambda}, \boldsymbol{\tau}|\mathbf{X}) \mathrm{d}\boldsymbol{\lambda} \mathrm{d}\boldsymbol{\tau} \right) \mathrm{d}\boldsymbol{\theta}_p \\
&\approx \int g(\boldsymbol{\theta}_p) \text{encoder}(\mathbf{X}_p) \mathrm{d}\boldsymbol{\theta}_p. \qquad (6)
\end{aligned}$$

We note that the model defined in Eq. 5, without mention of an encoder function, is already sufficiently defined for Bayesian inference. For this reason, one needs to take care so that the encoder function that does not modify the statistics of the model. We do so by defining a variational Bayesian expectation maximization (VBEM) algorithm for inferring the model and solving for the implied encoder function, which ends up obeying the integral relationship in Eq. 6.

Variational Bayesian EM

In the WD-FAB, the high dimensionality of the item bank makes Markov-Chain Monte-Carlo based inference of the model in Eq. 5 computationally impractical. Instead, we developed an efficient variational Bayesian expectation maximization (VBEM) [Bernardo et al. 2003] procedure that resembles common training techniques used for learning variational probabilistic autoencoders [Higgins et al. 2016, Ainsworth et al. 2018, Ansari and Soh 2018, Kingma and Welling 2013, Doersch 2016]. Additionally, this algorithm specifies an encoder function that consistently estimates the posterior statistics of the overall model.

The objective of variational inference is to find a surrogate distribution Q maximizing the evidence lower bound (ELBO), which takes the form

$$\mathcal{F}(\boldsymbol{\chi}) = \mathbb{E}_Q \log \pi(\mathbf{X}|\boldsymbol{\Gamma}) - D_{\text{KL}}\left(Q(\boldsymbol{\Gamma}|\boldsymbol{\chi})||\pi(\boldsymbol{\Gamma})\right), \qquad (7)$$

where $\boldsymbol{\Gamma}$ is a catch-all to represent all parameters from Eq. 5, and $\boldsymbol{\chi}$ represents the parameters that define the surrogate distribution Q.

In practice, we are only able to approximately optimize Eq. 7 because its exact maximization generally requires the computations of integrals that do not admit exact closed-form solutions. Instead, we utilize the strategy from ADVI [Blei, Kucukelbir, and McAuliffe 2017, Kucukelbir et al. 2017], seeking a factorized joint distribution Q that consists of a product of independent transformed Gaussian distributions $q_{(\cdot)}$

$$\boldsymbol{\alpha} \sim q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \Leftrightarrow f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \sim \text{normal}(\boldsymbol{\chi}_{\mu_{\boldsymbol{\alpha}}}, \boldsymbol{\chi}_{\sigma_{\boldsymbol{\alpha}}}), \qquad (8)$$

where $f_{\boldsymbol{\alpha}}$ is an invertible function such that $\text{supp}(\pi_{\boldsymbol{\alpha}}) = \text{codomain}(f_{\boldsymbol{\alpha}})$, and $\boldsymbol{\chi}_{\mu_{\boldsymbol{\alpha}}}, \boldsymbol{\chi}_{\sigma_{\boldsymbol{\alpha}}}$ are surrogate parameters. In our manuscript, we utilize the identify function for parameters that are valid on the reals, and the softplus function for positively supported functions. For the discrimination parameters, which are horseshoe-regulated, we utilize the normal inverse-gamma parameterization [Wand et al. 2011].

We define an iterative stochastic VBEM algorithm as follows. In VBE step $t+1$, we update the surrogate densities for parameters other than $\boldsymbol{\theta}$ (denoted $\boldsymbol{\Gamma} \setminus \boldsymbol{\theta}$) by taking a gradient ascent update against the expected ELBO,

$$\boldsymbol{\chi}^{(t+1)}_{\boldsymbol{\Gamma}\setminus\boldsymbol{\theta}} = \boldsymbol{\chi}^{(t)}_{\boldsymbol{\Gamma}\setminus\boldsymbol{\theta}} + \Delta_t \nabla_{\boldsymbol{\Gamma}\setminus\boldsymbol{\theta}} \mathbb{E}_{Q^{(t)}_{\boldsymbol{\theta}}}[\mathcal{F}(\boldsymbol{\chi}^{(t)})], \qquad (9)$$

where Δ_t is an adaptive step size – we utilized the Adam optimizer. Then we update $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, to a product of independent normal distributions that approximately satisfy

$$q^{(t+1)}_{\boldsymbol{\theta}_p} \propto \exp\left[\mathbb{E}_{Q^{(t+1)}_{\boldsymbol{\Gamma}\setminus\boldsymbol{\theta}}} \log \pi(\mathbf{X}|\boldsymbol{\Gamma}) \right]. \qquad (10)$$

The expectations in these expressions are not available in closed form. We approximate them using Monte-Carlo, by sampling parameters $\{\boldsymbol{\Gamma}^{(t)}_s\}_{s=1}^S$ drawn from $Q^{(t)}$, and computing the Monte-Carlo integral

$$\mathbb{E}_{Q_{\boldsymbol{\Gamma}\setminus\boldsymbol{\theta}}} \log \pi(\mathbf{X}|\boldsymbol{\Gamma}) \approx \frac{1}{S} \sum_{s=1}^{S} \log \pi(X|\boldsymbol{\Gamma}^{(t)}_s). \qquad (11)$$

For approximating the density in Eq. 10, we used moment matching, by parameterizing the independent Gaussian approximation using the mean and variance of the density – by computing first and second moments using the corresponding integral in Eq 6. Conditional on a sample of the item-level model parameters, this integral can be approximated to arbitrary numerical precision as a matrix product. The overall computations that go into evaluating the integrals constitute the learned encoder function. Note that the encoder function gracefully handles missingness in the item responses so long as responses are missing at random. The sum within the likelihood function of Eq. 4 excludes any unanswered items.

We implemented our method in Tensorflow Probability. Our implementation can be found publicly at `github:CC-RMD-EpiBio/autoencirt`.

Metrics

The instrument is intended to be used in comparing members in the population. The validity of these comparisons depends on the ability for a small set of latent factors to predict a multitude of items corresponding to functional ability. On this basis, we wish to evaluate the predictive accuracy of a candidate model, as performed in Chang et al. [2022b].

As a measure of predictive accuracy, we consider the Pareto-smoothed importance sampling-based leave-one-out (LOO) cross validation metric [Vehtari, Gelman, and Gabry 2015, Vehtari et al. 2019, Gelman, Hwang, and Vehtari 2014, Vehtari, Gelman, and Gabry 2017], which approximates the total log likelihood of left out data when fitting the model

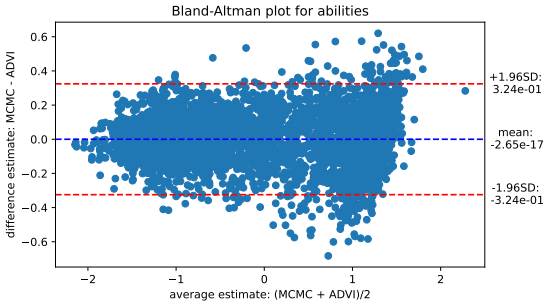

Figure 1: **Bland Altman plot for comparison of MCMC versus stochastic variational EM.** All scales simultaneously shown. Individual ability estimates are posterior means.

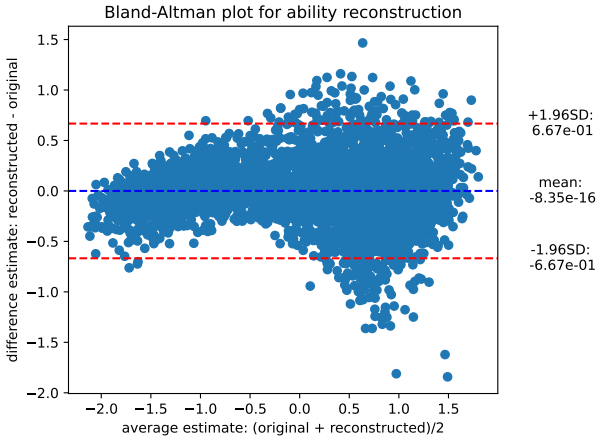

Figure 2: **Bland Altman plot for comparison of reconstructed abilities based on simulated responses.** All scales simultaneously shown. Individual ability estimates are posterior means.

using n-fold cross validation. Crucially, we respect the statistical dependencies in implementing LOO, by defining datapoints on a per-person basis rather than a per-item response basis. The prior literature has evaluated cross-validation and approximation metrics such as the LOO in IRT and similar contexts [Luo and Al-Harbi 2017, Chang 2019].

Results

We utilized a starting learning rate of 0.01 for the Adam optimizer, with batch sizes of 1190, shuffling the dataset and rebatching every full epoch. We used a maximum of 150 epochs, stopping training early if there was no improvement in the mean batch loss for three consecutive epochs. Generally, our models converged based on this criteria in approximately 80 to 120 epochs. For our dataset with $P = 11,901$ and $I \approx 300$, it took between 40 to 60 minutes to train each of the models mentioned in this section on an Apple M1 Pro Macbook Pro in CPU mode with 16GB of system memory.

Like in the posthoc WD-FAB, we separate the physical items and mental items, training two separate models. We will refer to the set of physical scales as the physical domain and the set of mental scales as the mental domain.

To validate our implementation of the stochastic variational Bayesian EM (VBEM) algorithm, we compare the abilities obtained using this method against ability estimates obtained via Hamiltonian Monte Carlo. Fig. 1 is a Bland-Altman plot that compares the mean estimate using MCMC against the mean estimate using our method for a $D = 3$ model calibrated using physical items. The standard deviation of the difference between these two estimates was approximately 0.16. By contrast, the average posterior standard deviation for the VBEM ability estimates was approximately 0.11.

We also evaluated the reconstruction of ability estimates from simulated responses. In this case, we took the fitted model where $D = 3$, and used it to simulate a set of new responses. We then fit a new model to the simulated responses. Fig. 2 compares the ability estimates reconstructed from the simulations to the original ability estimates. The standard deviation of the difference in these estimates was approximately 0.33.

Model selection

In formulating the instrument, one needs to resolve choices such as the scale dimension. Additionally, we also evaluated whether we should provide an additional mechanism to exclude items from the latent factor structure by uncoupling one of the dimensions in the instrument so that it is statistically independent of person-specific abilities. Finally, we wished to evaluate whether a local perturbation of the preexisting posthoc WD-FAB can perform as well as random initialization.

In Fig. 3 we use predictive accuracy by presenting the leave one out cross-validation information criterion computed over model variations. Specifically, the metric provides an approximation of the out-of-sample total log likelihood. Larger values of this metric are better.

The best-performing models did not have an independent dimension that was uncoupled to personalized abilities. Another consistent trend in these results is that the random initialization of the model appears to yield better-performing models, with the exception of high-dimensional models that perform worse overall. Initialization of the variational inference algorithm in the vicinity of the posthoc WD-FAB led to convergence to a local minimum obeying the posthoc WD-FAB factorization structure that did not predict item responses as accurately. Note that the prior instrument is four dimensional for each of physical and mental factorizations. For $D < 4$, we initialized to the first D scales. For $D > 4$, we initialized the extra scales using white noise.

For the physical items, the optimal dimension appears to be three, however, the four dimensional factorization metric falls within a standard error of that of the best three dimensional model. For reference, the dimension of the preexisting posthoc WD-FAB scales is four. For this reason, it is reasonable to utilize the four dimensional factorizations for each set of items.

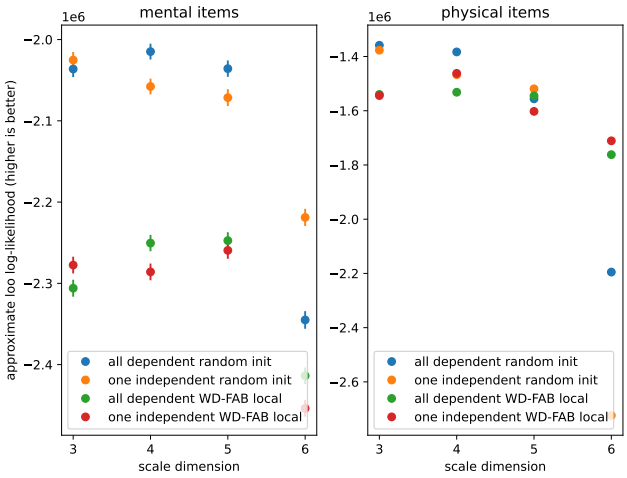

Figure 3: **Empirical predictive model evaluation** for physical and mental items using approximate leave one out cross validation (higher is better). Dimension, dependence of all scales on personal ability estimates, and local initialization in the vicinity of the prior posthoc model are evaluated.

Item factorizations

The original posthoc WD-FAB used four scale dimensions for each of the mental and physical domains. For this reason, in conjunction with the dimensionality analysis results of Fig. 3, we compare the four-dimensional in-IRT factorizations obtained using our method against the original posthoc WD-FAB factorizations. Fig. 4 provides the discrimination parameters for mental items and Fig. 5 provides discrimination parameters for physical parameters. Note that these parameters are proportional equivalent to weight matrices – an entry of zero means that an item does not load into a given instrument dimension. In each of the two figures, we display the top 20 items per dimension, as determined by ordering the discrimination parameters of the new factorization method (left) and ordering the discrimination of the posthoc WD-FAB discrimination parameters (right). Along the y-axis we denote the original ICF subcategorization for each item. Items with the same subcategorization were judged by disability experts to be more related than otherwise in terms of content matter as relates to the ICF. Since the SSA is exploring the use of the WD-FAB in its disability determination processes, the individual items are not published to prevent potential unfair advantages to applicants or beneficiaries. Notably, the in-IRT method yields factorizations that are distinct from the posthoc WD-FAB.

Mental factorization: For the mental items (Fig. 4), we see that the top items in the first in-IRT factorization dimension consist of a mixture of CC, II, and BH items. The second dimension consists almost entirely of II items, similar to the ME scale in the posthoc WD-FAB, detecting structure similar to the linear factorization used in the posthoc WD-FAB. The third and forth dimensions consist of mainly BH and CC items, corresponding largely to the CC scale of the posthoc WD-FAB. Ordering the items by their contribution to the

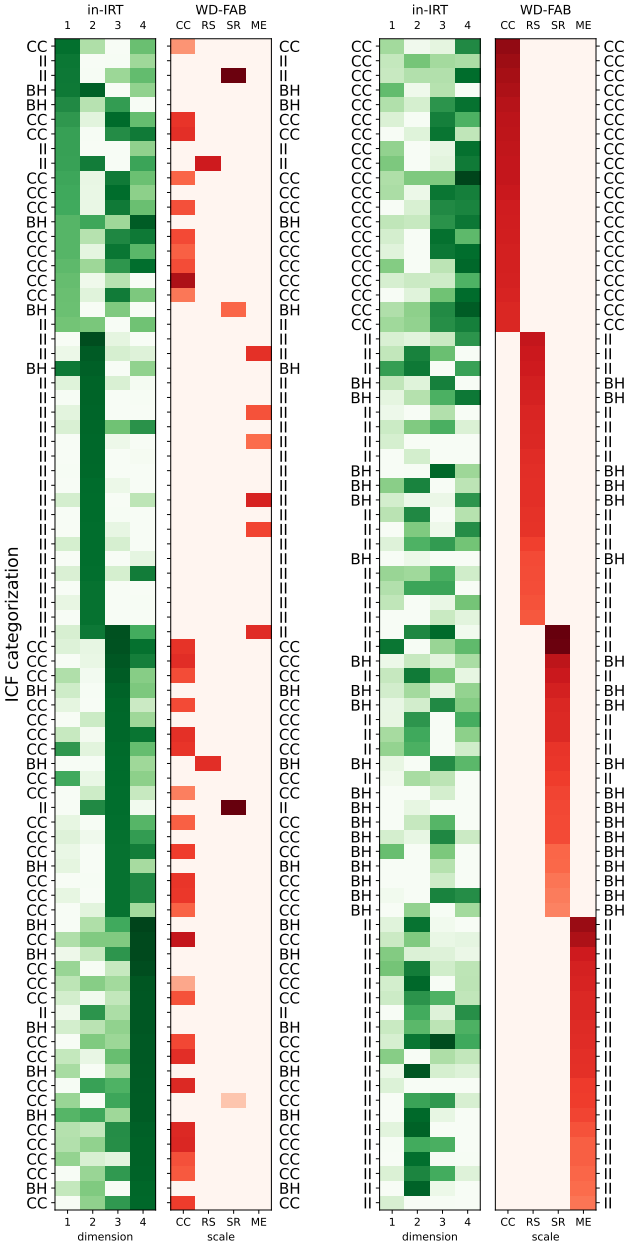

Figure 4: **Comparing mental item discrimination parameters** between the in-IRT factorization and the posthoc WD-FAB. Shown for each scale are the top 20 items as defined by discrimination for each of the in-IRT factorization (left) and the original posthoc WD-FAB factorization (right). Item discrimination parameters colored green for our method and red for the posthoc WD-FAB. Darker shades mean larger discrimination. Items identified by ICF subcategorization (CC, II, BH) along y-axis. Instrument dimension shown on x-axis.

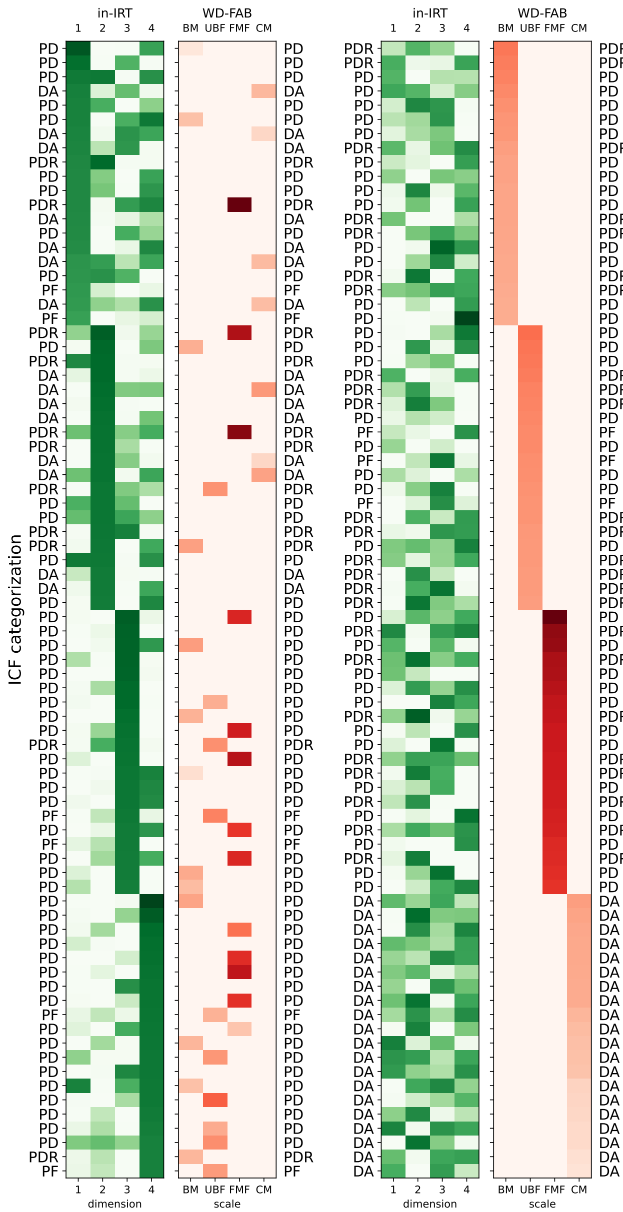

Figure 5: **Comparing physical item discrimination parameters** between the in-IRT factorization and the posthoc WD-FAB.Shown for each scale are the top 20 items as defined by discrimination for each of the in-IRT factorization (left) and the original posthoc WD-FAB factorization (right). Item discrimination parameters colored green for our method and red for the posthoc WD-FAB. Darker shades mean larger discrimination. Items identified by ICF subcategorization (PDR, PD, PF, DA) along y-axis. Instrument dimension shown on x-axis.

posthoc WD-FAB, we see that the top items in this instrument appear fairly randomly in the new factorization. The only notable trends are the posthoc CC items appearing the most-strongly in the third/forth dimensions of the new instrument, and the strongest ME items appearing largely in the second dimension of the new instrument.

Physical factorization: For the physical items (Fig. 5), we see that the top items in the first in-IRT dimension do not appear strongly in any of the posthoc WD-FAB scales. These items are a mixture between PD, DA, PDR, and PF ICF items. Only the forth scale in the new factorization has items that appear strongly within the posthoc WD-FAB, within the posthoc UBF and FMF subscales. Conversely, the strongest posthoc WD-FAB items do appear to have influence in the new factorization, though that influence is diffused within all four scales.

Discussion

We introduced a probabilistic autoencoder where the decoder is a multidimensional item response theory model and the encoder both helps define a variational EM procedure for Bayesian inference and amortizes the scoring of new responses. The key feature of this method is that it performs item factorization, selection, and model calibration coherently in a single self-consistent step. Hence, the development of the final model does not require subjective cutoffs that are typically used for setting either the structure or the dimensionality of the final model. Additionally, all model choices can be evaluated in unison using contemporary predictive metrics – as we did in choosing the scale dimension and model structure. It is seen in Fig. 3 that the in-IRT method consistently outperforms the posthoc WD-FAB in terms of predictive accuracy. Dividing the total out of sample likelihood by the number of responses, and exponentiating, we arrive at an estimate of the geometric mean of the per-response out-of-sample model likelihood. For the mental items, the geometric mean likelihood is 0.60 versus 0.56 for the posthoc WD-FAB. For the physical items, the geometric mean of the is 0.78 versus 0.75 for the WD-FAB.

Interpretability

By construction, our new factorization method yields an inherently interpretable model where each of the parameters have concrete explanations. The decoder is a multidimensional IRT model, the latent factors are person-specific ability parameters, and the encoder performs the relevant a-posteriori integral for mapping responses to scores. In this sense, the resulting model is inherently computationally interpretable [Chang et al. 2022a] but does not necessarily have attributes that make it comprehensible [Sudjianto and Zhang 2021]. In our case, for a disability instrument to be sensible requires that each ability parameter can map to an understandable attribute of function. The strength of the posthoc WD-FAB is in how each of the scales represents a concrete aspect of function. A-priori, the expectation that empirical patterns of responses would correspond to conceptually valid divisions in function may be unreasonable. Nonetheless, empirical factorization-based methods such as

ours and exploratory factor analysis all operate on this expectation. We note that our method yielded a solution that shares some of the factor structure of of the posthoc WD-FAB. In particular, the second dimension in the mental factorization is composed mostly of ME items from the posthoc WD-FAB, which are themselves mostly a subset of items that were categorized as II under the ICF. The posthoc WD-FAB benefited from a collaborative development iterative process where items were accepted or rejected based on subject matter cohesiveness. Future work will focus on how to incorporate such a process into developing such instruments.

Future directions

Our methodology analogizes multidimensional IRT and probabilistic autoencoders. Consistent with the generative Bayesian IRT model, we identified an appropriate encoder function that was completely determined by the decoder. In probabilistic (and other) autoencoders, this constraint is not generally true. An open question is to what extent an unconstrained encoder function would in-effect alter the statistics of the generative decoder model.

Since the likelihood function for our generative model can be expressed in elementary matrix operations common to artificial neural networks (using nonlinear activation functions), our overall method is also an interpretable neural network. This type of model may serve as a useful test bed for better-understanding the properties of neural networks in general, and probabilistic autoencoders in particular. Additionally, the way we have formulated the encoder function allows it to easily deal with missingness in the data – this aspect of the methodology could extend to autoencoders in general.

Our factorization method requires the pre-setting of dimension D, choosing the dimension based on model comparison using cross-validation. It may be possible to perform this dimensionality selection within a single model by putting a prior on the dimension and using posterior projective inference techniques [Piironen and Vehtari 2017b].

Acknowledgements

This work is supported by the Intramural Research Programs of the National Institutes of Health Clinical Center (CC) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), and the US Social Security Administration. The authors thank Beth Rasch, Elizabeth Marfeo, Christine McDonough, and Howard Goldman for their helpful feedback.

References

Ainsworth, S.; Foti, N.; Lee, A. K.; and Fox, E. 2018. Interpretable VAEs for Nonlinear Group Factor Analysis. *arXiv:1802.06765 [cs, stat]*.

Ansari, A. F.; and Soh, H. 2018. Hyperprior Induced Unsupervised Disentanglement of Latent Representations. *arXiv:1809.04497 [cs, stat]*.

Bernardo, J. M.; Bayarri, M. J.; Berger, J. O.; Dawid, A. P.; Heckerman, D.; Smith, A. F. M.; West (eds, M.; Beal, M. J.; and Ghahramani, Z. 2003. The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring Graphical Model Structures.

Bhadra, A.; Datta, J.; Polson, N. G.; and Willard, B. 2015. The Horseshoe+ Estimator of Ultra-Sparse Signals. *arXiv:1502.00560 [math, stat]*.

Bhadra, A.; Datta, J.; Polson, N. G.; and Willard, B. 2019. Lasso Meets Horseshoe: A Survey. *Statistical Science*, 34(3): 405–427.

Bilbao, A.; Las Hayas, C.; Forero, C. G.; Padierna, A.; Martin, J.; and Quintana, J. M. 2014. Cross-Validation Study Using Item Response Theory: The Health-Related Quality of Life for Eating Disorders Questionnaire–Short Version. *Assessment*, 21(4): 477–493.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518): 859–877.

Bore, M.; Laurens, K. R.; Hobbs, M. J.; Green, M. J.; Tzoumakis, S.; Harris, F.; and Carr, V. J. 2020. Item Response Theory Analysis of the Big Five Questionnaire for Children–Short Form (BFC-SF): A Self-Report Measure of Personality in Children Aged 11–12 Years. *Journal of Personality Disorders*, 34(1): 40–63.

Brandt, D.; and Smalligan, J. 2019. A New Approach to Examining Disability: How the WD-FAB Could Improve SSA's Processes and Help People with Disabilities Stay Employed.

Carlson, J. E.; and von Davier, M. 2013. Item Response Theory. *ETS Research Report Series*, 2013(2): i–69.

Carvalho, C. M.; Polson, N. G.; and Scottt, J. G. 2010. The Horseshoe Estimator for Sparse Signals. *Biometrika*, 97(2): 465–480.

Cella, D.; Yount, S.; Rothrock, N.; Gershon, R.; Cook, K.; Reeve, B.; Ader, D.; Fries, J. F.; Bruce, B.; Rose, M.; and PROMIS Cooperative Group. 2007. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group during Its First Two Years. *Medical Care*, 45(5 Suppl 1): S3–S11.

Chang, J. C. 2019. Predictive Bayesian Selection of Multistep Markov Chains, Applied to the Detection of the Hot Hand and Other Statistical Dependencies in Free Throws. *Royal Society Open Science*, 6(3): 182174.

Chang, J. C.; Chang, T. L.; Chow, C. C.; Mahajan, R.; Mahajan, S.; Maisog, J.; Vattikuti, S.; and Xia, H. 2022a. Interpretable (Not Just Posthoc-Explainable) Medical Claims Modeling for Discharge Placement to Prevent Avoidable All-Cause Readmissions or Death. arXiv:2208.12814.

Chang, J. C.; Fletcher, P.; Han, J.; Chang, T. L.; Vattikuti, S.; Desmet, B.; Zirikly, A.; and Chow, C. C. 2020. Sparse Encoding for More-Interpretable Feature-Selecting Representations in Probabilistic Matrix Factorization. *arXiv:2012.04171 [cs, q-bio, stat]*.

Chang, J. C.; Porcino, J.; Rasch, E. K.; and Tang, L. 2022b. Regularized Bayesian Calibration and Scoring of the WD-FAB IRT Model Improves Predictive Performance over Marginal Maximum Likelihood. *PLOS ONE*, 17(4): e0266350.

Chang, J. C.; Vattikuti, S.; and Chow, C. C. 2019. Probabilistically-Autoencoded Horseshoe-Disentangled Multidomain Item-Response Theory Models. *arXiv:1912.02351 [cs, stat]*.

Converse, G.; Curi, M.; and Oliveira, S. 2019. Autoencoders for Educational Assessment. In Isotani, S.; Millán, E.; Ogan, A.; Hastings, P.; McLaren, B.; and Luckin, R., eds., *Artificial Intelligence in Education*, Lecture Notes in Computer Science, 41–45. Springer International Publishing. ISBN 978-3-030-23207-8.

Converse, G.; Curi, M.; Oliveira, S.; and Templin, J. 2021. Estimation of Multidimensional Item Response Theory Models with Correlated Latent Variables Using Variational Autoencoders. *Machine Learning*, 110(6): 1463–1480.

DeWalt, D. A.; Rothrock, N.; Yount, S.; and Stone, A. A. 2007. Evaluation of Item Candidates: The PROMIS Qualitative Item Review. *Medical care*, 45(5 Suppl 1): S12–S21.

DeYoung, C. G.; Carey, B. E.; Krueger, R. F.; and Ross, S. R. 2016. 10 Aspects of the Big Five in the Personality Inventory for DSM-5. *Personality disorders*, 7(2): 113–123.

Doersch, C. 2016. Tutorial on Variational Autoencoders. *arXiv:1606.05908 [cs, stat]*.

Fieo, R.; Watson, R.; Deary, I. J.; and Starr, J. M. 2010. A Revised Activities of Daily Living/Instrumental Activities of Daily Living Instrument Increases Interpretive Power: Theoretical Application for Functional Tasks Exercise. *Gerontology*, 56(5): 483–490.

Fries, J. F.; Witter, J.; Rose, M.; Cella, D.; Khanna, D.; and Morgan-DeWitt, E. 2014. Item Response Theory, Computerized Adaptive Testing, and PROMIS: Assessment of Physical Function. *The Journal of Rheumatology*, 41(1): 153–158.

Funke, F. 2005. The Dimensionality of Right-Wing Authoritarianism: Lessons from the Dilemma between Theory and Measurement. *Political Psychology*, 26(2): 195–218.

Gelman, A.; Hwang, J.; and Vehtari, A. 2014. Understanding Predictive Information Criteria for Bayesian Models. *Statistics and Computing*, 24(6): 997–1016.

Goldberg, L. R. 1992. The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment*, 4(1): 26–42.

Gopalan, P.; Ruiz, F. J.; Ranganath, R.; and Blei, D. 2014. Bayesian Nonparametric Poisson Factorization for Recommendation Systems. In *Artificial Intelligence and Statistics*, 275–283.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.

Jette, A. M.; Ni, P.; Rasch, E.; Marfeo, E.; McDonough, C.; Brandt, D.; Kazis, L.; and Chan, L. 2019. The Work Disability Functional Assessment Battery (WD-FAB). *Physical Medicine and Rehabilitation Clinics*, 30(3): 561–572.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*.

Kingston, N.; Leary, L.; and Wightman, L. 1985. An Exploratory Study of the Applicability of Item Response Theory Methods to the Graduate Management Admission Test1. *ETS Research Report Series*, 1985(2): i–56.

Kingston, N. M.; and Dorans, N. J. 1982. The Feasibility of Using Item Response Theory as a Psychometric Model for the Gre Aptitude Test. *ETS Research Report Series*, 1982(1): i–148.

Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; and Blei, D. M. 2017. Automatic Differentiation Variational Inference. *The Journal of Machine Learning Research*, 18(1): 430–474.

Luo, Y.; and Al-Harbi, K. 2017. Performances of LOO and WAIC as IRT Model Selection Methods. *Psychological Test and Assessment Modeling*, 59(2): 183.

Marfeo, E.; Ni, P.; Meterko, M.; Marino, M.; Peterik, K.; McDonough, C.; Rasch, E. K.; Brandt, D.; Chan, L.; and Jette, A. 2016. Development of a New Instrument to Assess Work-Related Function: Work Disability Functional Assessment Battery (WD-FAB). *American Journal of Occupational Therapy*, 70(4_Supplement_1): 7011500012p1–7011500012p1.

Marfeo, E. E.; McDonough, C.; Ni, P.; Peterik, K.; Porcino, J.; Meterko, M.; Rasch, E.; Kazis, L.; and Chan, L. 2019. Measuring Work Related Physical and Mental Health Function: Updating the Work Disability Functional Assessment Battery (WD-FAB) Using Item Response Theory. *Journal of Occupational and Environmental Medicine*, 61(3): 219–224.

Marfeo, E. E.; Ni, P.; McDonough, C.; Peterik, K.; Marino, M.; Meterko, M.; Rasch, E. K.; Chan, L.; Brandt, D.; and Jette, A. M. 2018. Improving Assessment of Work Related Mental Health Function Using the Work Disability Functional Assessment Battery (WD-FAB). *Journal of Occupational Rehabilitation*, 28(1): 190–199.

Meterko, M.; Marfeo, E. E.; McDonough, C. M.; Jette, A. M.; Ni, P.; Bogusz, K.; Rasch, E. K.; Brandt, D. E.; and Chan, L. 2015. Work Disability Functional Assessment Battery: Feasibility and Psychometric Properties. *Archives of Physical Medicine and Rehabilitation*, 96(6): 1028–1035.

Mnih, A.; and Salakhutdinov, R. R. 2008. Probabilistic Matrix Factorization. In Platt, J. C.; Koller, D.; Singer, Y.; and Roweis, S. T., eds., *Advances in Neural Information Processing Systems 20*, 1257–1264. Curran Associates, Inc.

Piironen, J.; and Vehtari, A. 2017a. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In *AISTATS*.

Piironen, J.; and Vehtari, A. 2017b. Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors. *Electronic Journal of Statistics*, 11(2): 5018–5051.

Porcino, J.; Marfeo, B.; McDonough, C.; and Chan, L. 2018. The Work Disability Functional Assessment Battery (WD-FAB): Development and validation review. *TBV – Tijdschrift voor Bedrijfs- en Verzekeringsgeneeskunde*, 26(7): 344–349.

Samejima, F. 1969. Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2): 100–100.

Saunders, B. A.; and Ngo, J. 2017. The Right-Wing Authoritarianism Scale. In Zeigler-Hill, V.; and Shackelford, T. K., eds., *Encyclopedia of Personality and Individual Differences*, 1–4. Cham: Springer International Publishing. ISBN 978-3-319-28099-8.

Spence, R.; Owens, M.; and Goodyer, I. 2012. Item Response Theory and Validity of the NEO-FFI in Adolescents. *Personality and Individual Differences*, 53(6): 801–807.

Sudjianto, A.; and Zhang, A. 2021. Designing Inherently Interpretable Machine Learning Models. arXiv:2111.01743.

van der Pas, S. L.; Kleijn, B. J. K.; and van der Vaart, A. W. 2014. The Horseshoe Estimator: Posterior Concentration around Nearly Black Vectors. *Electronic Journal of Statistics*, 8(2).

Vehtari, A.; Gelman, A.; and Gabry, J. 2015. Pareto Smoothed Importance Sampling. *arXiv:1507.02646 [stat]*.

Vehtari, A.; Gelman, A.; and Gabry, J. 2017. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing*, 27(5): 1413–1432.

Vehtari, A.; Simpson, D.; Gelman, A.; Yao, Y.; and Gabry, J. 2019. Pareto Smoothed Importance Sampling. *arXiv:1507.02646 [stat]*.

Wand, M. P.; Ormerod, J. T.; Padoan, S. A.; and Frühwirth, R. 2011. Mean Field Variational Bayes for Elaborate Distributions. *Bayesian Analysis*, 6(4): 847–900.

Yuker, H. E. 1994. Variables That Influence Attitudes toward People with Disabilities: Conclusions from the Data. *Journal of Social Behavior & Personality*, 9: 3–22.

Shrinkage scaling

The global shinkage parameters $\kappa^{(d)}$ control the overall amount of sparsity in each dimension. Apriori, for partitioning a set of I items into D dimensions, we would expect each dimension to have approximately $\bar{I}/D$ nonzero terms, where $\bar{I} \leq I$. Our objective is to find a consistent scaling for the global shrinkage parameters $\kappa^{(d)}$. Ignoring the entropy penalization, consider the conditional posterior density of the discrimination parameters $\lambda_i^{(d)}$,

$$\begin{aligned}
&\pi(\boldsymbol{\lambda}|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\kappa}) \propto \\
&\prod_{i,p,d,k} \left[\Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{ik}^{(d)})\right) - \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)\right]^{\delta_{x_{pi}k}} \\
&\times \exp\left[-\frac{1}{2}\sum_{i,d}\left(\frac{\lambda_i^{(d)}}{\xi_i^{(d)}\kappa^{(d)}}\right)^2\right].
\end{aligned} \tag{12}$$

For notational convenience, $\tau_{i1} = -\infty$ and $\tau_{i,K+1} = \infty$. We analyze this density to examine how $\kappa^{(d)}$ influences the mode of this posterior marginal distribution for the discriminations.

Suppose that

$$\hat{\lambda}_i^{(d)}(\kappa^{(d)}) = \hat{\lambda}_{i,\infty}^{(d)}\left(1 - \hat{\Delta}_i^{(d)}(\kappa^{(d)})\right)$$

conditionally maximizes Eq. 12 for a given value of $\kappa^{(d)}$. Then, as in Piironen and Vehtari [2017a,b], we define the expected number of non-zero discrimination parameters in a single scale d,

$$m_{\text{eff}}^{(d)} = I\left(1 - \mathbb{E}(\hat{\Delta}_i^{(d)}(\kappa^{(d)}))\right) \approx \frac{\bar{I}}{D}.$$

So, we would like to find $\kappa^{(d)}$ such that

$$\mathbb{E}\left(\hat{\Delta}_i^{(d)}(\kappa^{(d)})\right) \approx 1 - \frac{\bar{I}}{ID}. \tag{13}$$

Approximation of the shrinkage

We will approximate the expectation of the shrinkage factor $\hat{\Delta}_i^{(d)}(\kappa^{(d)})$ to leading order in P^{-1}. We note that $\hat{\lambda}_i^{(d)}(\kappa^{(d)}) \to \hat{\lambda}_i^{(d)}(\infty) \equiv \hat{\lambda}_{i,\infty}^{(d)}$ as $\kappa^{(d)}/P \to 0$, and make the ansatz

$$\hat{\Delta}_i^{(d)}(\kappa^{(d)}) = \frac{1}{P}\hat{\Delta}_i^{(d,0)}(\kappa^{(d)}) + \mathcal{O}(P^{-2}). \tag{14}$$

By definition, $\hat{\lambda}_i^{(d)}(\kappa^{(d)})$ is a root of the equation

$$\begin{aligned}
0 &= \frac{\partial}{\partial \lambda_i^{(d)}} \log \pi(\boldsymbol{\lambda}|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\kappa}) \\
&= \sum_p \left[\sum_k \delta_{x_{pi}k} \frac{\partial_{\lambda_i^{(d)}} F_{ipkd}}{F_{ipkd}} - \frac{1}{P}\frac{\lambda_i^{(d)}}{(\xi_i^{(d)}\kappa^{(d)})^2}\right]
\end{aligned} \tag{15}$$

where

$$F_{ipkd} = \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{ik}^{(d)})\right) - \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right). \tag{16}$$

Denoting the standard normal density

$$\phi(x) = \Phi'(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \tag{17}$$

we differentiate Eq. 16 with respect to $\lambda_i^{(d)}$ in each of three cases. If $k \in \{2, 3, \ldots, K-1\}$, then

$$\begin{aligned}
\partial_{\lambda_i^{(d)}} F_{ipkd} = &(\theta_p^{(d)} - \tau_{i,k}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) \\
&- (\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right),
\end{aligned} \tag{18}$$

otherwise if $k = 1$,

$$\partial_{\lambda_i^{(d)}} F_{ipkd} = -(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right), \tag{19}$$

lastly when $k = K$

$$\partial_{\lambda_i^{(d)}} F_{ipkd} = (\theta_p^{(d)} - \tau_{i,k}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right). \tag{20}$$

We then substitute Eqs. 18–20 into Eq. 15 so that

$$\begin{aligned}
0 = \sum_p \Bigg\{ & \delta_{x_{pi}K} \frac{(\theta_p^{(d)} - \tau_{i,K}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,K}^{(d)})\right)}{\Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} - \delta_{x_{pi}1} \frac{(\theta_p^{(d)} - \tau_{i,2}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,2}^{(d)})\right)}{1 - \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,2}^{(d)})\right)} \\
& + \sum_{k=2}^{K-1} \delta_{x_{pi}k} \left[\frac{(\theta_p^{(d)} - \tau_{i,k}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right)}{\Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{ik}^{(d)})\right) - \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} - \frac{(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)}{\Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \right] \\
& - \frac{\lambda_i^{(d)}}{P(\xi_i^{(d)}\kappa^{(d)})^2} \Bigg\}.
\end{aligned} \tag{21}$$

We now wish to perturb Eq 21. To begin, we expand some constituent terms about $\lambda_i^{(d)} \approx \hat{\lambda}_{i,\infty}^{(d)}$, in powers of the shrinkage factor,

$$\begin{aligned}
\frac{1}{\Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} &= \frac{1}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} \\
&\quad \times \left[1 + \hat{\lambda}_{i,\infty}^{(d)}\hat{\Delta}_i^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)}) \frac{\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} + \mathcal{O}((\hat{\Delta}_i^{(d)})^2)\right] && \text{(22a)} \\
\frac{1}{1 - \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)} &= \frac{1}{1 - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)} \\
&\quad \times \left[1 - \hat{\lambda}_{i,\infty}^{(d)}\hat{\Delta}_i^{(d)} \frac{(\theta_p^{(d)} - \tau_{iK}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)}{1 - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)} + \mathcal{O}((\hat{\Delta}_i^{(d)})^2)\right] && \text{(22b)} \\
\frac{1}{\Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} &= \\
\frac{1}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} & \\
\times \Bigg[1 + \hat{\lambda}_{i,\infty}^{(d)}\hat{\Delta}_i^{(d)} \Bigg(& \frac{(\theta_p^{(d)} - \tau_{ik}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{ik}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \\
& - \frac{(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \Bigg) + \mathcal{O}((\hat{\Delta}_i^{(d)})^2)\Bigg] && \text{(22c)} \\
\phi\left(\lambda_i^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) &= \phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \hat{\lambda}_{i,\infty}^{(d)}\hat{\Delta}_i^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\phi'\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) + \mathcal{O}((\hat{\Delta}_i^{(d)})^2) \\
&= \phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) + \hat{\Delta}_i^{(d)}(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)}))^2\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) + \mathcal{O}((\hat{\Delta}_i^{(d)})^2). && \text{(22d)}
\end{aligned}$$

Now we groups terms from Eq. 21 in terms of P. To order P,

$$
\begin{aligned}
0 = \sum_p \Bigg\{ & \delta_{x_{pi}K} \frac{(\theta_p^{(d)} - \tau_{i,K}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,K}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} - \delta_{x_{pi}1} \frac{(\theta_p^{(d)} - \tau_{i,2}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,2}^{(d)})\right)}{1 - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,2}^{(d)})\right)} \\
& + \sum_{k=2}^{K-1} \delta_{x_{pi}k} \Bigg[\frac{(\theta_p^{(d)} - \tau_{i,k}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{ik}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \\
& \quad - \frac{(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \Bigg] \Bigg\}
\end{aligned} \tag{23}
$$

.

To order P^0,

$$
\begin{aligned}
& \frac{\hat{\lambda}_{i,\infty}^{(d)}}{(\xi_i^{(d)}\kappa^{(d)})^2} = \\
& \hat{\Delta}_i^{(d)} \sum_p \Bigg\{ \frac{\delta_{x_{pi}K}(\theta_p^{(d)} - \tau_{iK}^{(d)})}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} \\
& \qquad \times \left[(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)}))^2 \phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) + \hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)}) \frac{\phi^2\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} \right] \\
& + \frac{\delta_{x_{pi}1}(\theta_p^{(d)} - \tau_{i2}^{(d)})}{1 - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)} \Bigg[(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,2}^{(d)}))^2 \phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,2}^{(d)})\right) \\
& \qquad - \frac{\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\phi^2\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)}{1 - \Phi\left(\lambda_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)} \Bigg] \\
& \quad + \sum_{k=2}^{K-1} \frac{\delta_{x_{pi}k}}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \\
& \qquad \times \Bigg[(\hat{\lambda}_{i,\infty}^{(d)})^2(\theta_p^{(d)} - \tau_{i,k}^{(d)}))^3 \phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - (\hat{\lambda}_{i,\infty}^{(d)})^2(\theta_p^{(d)} - \tau_{i,k+1}^{(d)}))^3 \phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right) \\
& + \hat{\lambda}_{i,\infty}^{(d)} \frac{\left((\theta_p^{(d)} - \tau_{i,k}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - (\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)\right)^2}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \Bigg] \Bigg\} \\
& \equiv \hat{\Delta}_i^{(d)} \sum_p R_p \\
& \equiv \hat{\Delta}_i^{(d)} P \bar{R},
\end{aligned} \tag{24}
$$

for some value $\bar{R}$.

Approximation of $\bar{R}$

First we note that

$$
\mathbb{E}(\delta_{x_{pi}k}) = \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\hat{\theta}_{p,\infty}^{(d)} - \hat{\tau}_{i,k,\infty}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\hat{\theta}_{p,\infty}^{(d)} - \hat{\tau}_{i,k+1,\infty}^{(d)})\right) + \mathcal{O}(1/P). \tag{25}
$$

for parameters $\hat{\tau}_{i,k,\infty}^{(d)}, \hat{\theta}_{p,\infty}^{(d)}$ corresponding to the posterior mode of the model in Eq. 5. In the large P limit, the marginal distributions for these parameters becomes tightly-centered around their posterior modes. So, we approximate Eq 24 by directly

substituting in Eq. 25, approximating each of $\hat{\tau}_{i,k,\infty}^{(d)}, \hat{\theta}_{p,\infty}^{(d)}$ about $\hat{\tau}_{i,k}^{(d)}, \hat{\theta}_{p}^{(d)}$, and discarding higher order terms, leading to the following expression

$$
\begin{aligned}
R_p &\approx \hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})^2 \frac{\phi^2\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)} - \frac{\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})^2\phi^2\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)}{1 - \Phi\left(\lambda_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i2}^{(d)})\right)} \\
&\quad + \sum_{k=2}^{K-1} \hat{\lambda}_{i,\infty}^{(d)} \frac{\left((\theta_p^{(d)} - \tau_{i,k}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - (\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)\right)^2}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k}^{(d)})\right) - \Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{i,k+1}^{(d)})\right)} \\
&\approx K\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})^2 \frac{\phi^2\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)}{\Phi\left(\hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})\right)},
\end{aligned} \tag{26}
$$

where we have also assumed approximate symmetry in the empirical response distribution, retaining only the terms corresponding to $k = K$.

Now we intend to take the expectation of Eq. 24 with respect to the remaining free parameters $\boldsymbol{\theta}_p^{(d)}, \boldsymbol{\tau}_i^{(d)}$. We note first that $\theta_p - \tau_{i2}^{(d)} \sim \mathcal{N}(0, \sqrt{3})$, invoking the central limit theorem to approximate the statistics of $\tau_{i,k}^{(d)}$, for $k \geq 2$, as

$$
\pi(\theta_p^{(d)} - \tau_{i,k}^{(d)}) \approx \text{normal}\left(\underbrace{(k-2)\sqrt{\frac{2}{\pi}}}_{M_k}, \underbrace{\sqrt{3 + (k-2)\left(1 + \frac{2}{\pi}\right)}}_{S_k}\right). \tag{27}
$$

Using the substitution $z = \hat{\lambda}_{i,\infty}^{(d)}(\theta_p^{(d)} - \tau_{iK}^{(d)})$ we write the expectation of R_p with respect to the density in Eq. 27,

$$
\bar{R} \approx \frac{K}{\hat{\lambda}_{i,\infty}^{(d)} S_k} \int_{-\infty}^{\infty} z^2 \phi(z) \frac{\phi(z)}{\Phi(z)} \phi\left(\frac{z - \hat{\lambda}_{i,\infty}^{(d)} M_K}{S_K \hat{\lambda}_{i,\infty}^{(d)}}\right) \mathrm{d}z \tag{28}
$$

which is the expectation of the function

$$
g(z) = z^2\phi^2(z)/\Phi(z)
$$

relative to Gaussian density with mean $\hat{\lambda}_{i,\infty}^{(d)} M_K > 0$ and standard deviation $\hat{\lambda}_{i,\infty}^{(d)} S_K$. One can easily approximate Eq. 28 using numerical techniques. We provide a cheap estimate by expanding $g(z)$ in a power series around $z = \hat{\lambda}_{i,\infty}^{(d)} M_K$,

$$
\begin{aligned}
\hat{R} &\approx K \sum_{n=0}^{\infty} \frac{g^{(2n)}(\hat{\lambda}_{i,\infty}^{(d)} M_K)}{(2n)!} \int_{-\infty}^{\infty} \frac{\left(z - \hat{\lambda}_{i,\infty}^{(d)} M_K\right)^{2n}}{\hat{\lambda}_{i,\infty}^{(d)} S_k} \phi\left(\frac{z - \hat{\lambda}_{i,\infty}^{(d)} M_K}{S_K \hat{\lambda}_{i,\infty}^{(d)}}\right) \mathrm{d}z \\
&= K \sum_{n=0}^{\infty} \frac{g^{(2n)}(\hat{\lambda}_{i,\infty}^{(d)} M_K)}{(2n)!} (2n-1)!! (\hat{\lambda}_{i,\infty}^{(d)} M_K)^{2n}.
\end{aligned} \tag{29}
$$

Putting it all together

From Eq. 24 and Eq. 13 we have

$$
\xi_i^{(d)} \kappa^{(d)} = \sqrt{\frac{\hat{\lambda}_{i,\infty}^{(d)} ID}{(ID - \bar{I})\, P\bar{R}}}. \tag{30}
$$

Assuming $\hat{\lambda}_i^{(d)}$ and $\xi_i^{(d)}$ are both unit scale, then

$$
\kappa_0^{(d)} = \sqrt{\frac{\Delta(D, K, I)}{P}} \tag{31}
$$

where

$$
\Delta(D, K, I) = \frac{ID}{(ID - \bar{I})\, \bar{R}} \tag{32}
$$

is an appropriate scaling term for the global shrinkage parameters.