Bayesian information theoretic model-averaging stochastic item selection for computer adaptive testing

Anonymous authors
Paper under double-blind review

Abstract

Computer Adaptive Testing (CAT) aims to accurately estimate an individual's ability using only a subset of an Item Response Theory (IRT) instrument. A secondary goal is to ensure diverse item exposure across different testing sessions, preventing any single item from being over or underutilized. In CAT, items are selected sequentially based on a running estimate of a respondent's ability. Prior methods almost universally see item selection through an optimization lens, motivating greedy item selection procedures. While efficient, these methods tend to have poor item exposure. Existing stochastic methods for item selection are ad-hoc, where item sampling weights lack theoretical justification. In this manuscript, we formulate CAT as a Bayesian model averaging problem. At each step, we sample the next item in a manner where the Frequentist item sampling statistics correspond to Bayesian model averaging in the space of next-item ability estimates. This view of the CAT item selection problem also defines the natural criterion of the ability discrepancy: the KL divergence between the unknown next-item ability estimate and the unknown true full item bank ability estimate. We tested our new method on the eight independent IRT models that comprise the Work Disability Functional Assessment Battery, comparing it to prior art. We found that our stochastic methodology had superior item exposure while not compromising in terms of test accuracy and efficiency.

1 Introduction

The combination of Item Response Theory (IRT) and Computer Adaptive Testing (CAT) forms the dominant methodology backing the use of exams for ability assessment. High profile examples of this pairing include the Graduate Management Admission Test (GMAT) (Kingston et al., 1985; Rudner, 2010), the nursing National Council Licensure Examination (NCLEX) (Woo & Dragan, 2012), the National Registry of Emergency Medical Technicians (NREMT) (Ventura et al., 2021), and the Armed Services Vocational Aptitude Battery (ASVAB) (Segall & Moreno, 1999). IRT/CAT also features in many healthcare contexts because of its adaptation in Patient Reported Outcomes Measurement Information System (PROMIS) instruments (Cella et al., 2007; 2010; Segawa et al., 2020) that are widely used in FDA-regulated trials.

1.1 Item Response Theory (IRT)

IRT, a generative latent-variable modeling framework, models how a respondent of a given ability might respond to to each item in a testing bank. In IRT, an ability (canonically denoted θ) is a theoretically continuous valued parameter (Bock et al., 1997; Immekus et al., 2019; Böckenholt & Meiser, 2017). The initial step for developing an IRT model involves creating a large pool of items that are topically grounded in a construct being measured. These items are then administered to a sizable and diverse sample of respondents, producing a dataset of item responses for model calibration. In the process of fitting an IRT model to the set of item responses, each item's specific parameters are determined (Kieftenbeld & Natesan, 2012; Luo & Jiao, 2018; Bürkner, 2021; Lord, 1983; Natesan, 2011; Natesan et al., 2016). Self-consistently, the ability for each of the respondents is also determined. Due to this coupling, the ability statistics for the calibration sample encode into the item-specific parameters, a fact made explicit by their relationship between IRT and probabilistic autoencoders (Converse et al., 2019; Chang et al., 2019; 2023). Fundamentally, IRT maps each respondent's set of discrete item responses to a lower (usually single) dimensional latent space. In this manner, IRT models, like autoencoders, are nonlinear factorization models (Chang et al., 2021).

The goal of IRT is to apply such pre-trained models to new respondents, ranking them relative to the respondents used in model calibration. To do so, item parameters from calibration are held fixed and new responses for a given respondent are scored by solving an associated inverse problem for the ability parameter.

The possibly large item bank developed for the IRT model ideally has content coverage throughout the entire range of possible abilities. Administering a large item bank is burdensome for all parties involved. In the vicinity of any fixed ability parameter, however, the number of items is relatively small. CAT exploits this fact.

1.2 Computer Adaptive Testing (CAT)

The goal of computer adaptive testing (CAT) is to efficiently and accurately estimate a respondent's ability by using only the most relevant questions from a possibly large item battery. This selection is performed sequentially based on a running estimate of the respondent's ability. Selection methods mainly differ on the specific statistical objective being optimized. Generally, individual items are judged based on some measure of the degree to which they may improve the fidelity of the respondent's ability estimate. Most commonly, items are chosen greedily – while efficient, this type of selection procedure has the pitfall of poor item exposure.

Item exposure refers to the rate at which individual items in a testing bank are presented across multiple administrations. When exposure is poor, the effective instrument administered by the CAT is a limited subset of the items in the original bank. In unison with commonly-used improper scoring rules, this condition biases the resulting ability estimates. Having a small number of effective items also implies stereotypical item trajectories, making such instruments easier to game.

CAT methodologies select items based on a running estimate of a test-taker's ability. However, this estimate is unreliable at the beginning of the test, which in turn makes the statistical measures used to compare items noisy. For this reason, simply choosing the item that appears statistically best (a "greedy" approach) may not be ideal. A more effective strategy may be to hedge, selecting items that are useful across a wider range of potential ability levels.

In this manuscript we provide a methodology for hedging that is based on viewing item selection as a model selection problem. Each item implies a different model for the respondent's ability at the next step of the test. As a consequence of viewing the problem through these lens, we both motivate a new item selection criterion based on the information theoretic ability model discrepancy, and a stochastic selection procedure where the Frequentist statistics of item probabilities correspond to Bayesian model averaging statistics of the item-wise implied ability estimates.

1.3 Work Disability Functional Assessment Battery (WD-FAB)

As concrete tests of our methodology we used the eight independent IRT models, and their associated item banks, present in the WD-FAB (Meterko et al., 2015; Marfeo et al., 2016; 2019; Chang et al., 2022; Marfeo et al., 2018; Jette et al., 2019; Porcino et al., 2018). The WD-FAB characterizes whole body and mental function across four physical instruments and four mental instruments. The item banks consist of questions that ask about a range of everyday activities, such as emptying a dishwasher, walking a block, turning a door knob, speaking to someone on the phone, and managing under stress. Accepted responses were graded on either four or five option ordinal Likert scales.

The intended use of this instrument is to provide standardized and reliable information about an individual's functional abilities to help inform 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.

As an application where item exposure is important, the eight independent models that comprise the WD-FAB are an ideal testing ground for our methodology.

2 Preliminaries

Suppose that one has developed a test bank consisting of N_{items} items, of a given ordering, and used a set of responses to these items in order to calibrate an IRT model. The IRT model implies that a person of ability θ is expected to respond to the i-th item according to the probability mass function $p_i(k|\theta)$. For generality, we assume that the IRT model is polytomous so that there are N_{levels} possible responses to each item. If a fully Bayesian approach was used in calibration, then p_i can be the marginal probability mass found by integrating out the posterior item specific parameters. Otherwise, it is the probability mass implied by point estimates of the item specific parameters. For a given individual, knowing all of their responses $\boldsymbol{x} = (x_1, x_2, \ldots, x_{N_{\text{items}}})$, one may estimate the ability of the individual by computing the statistics of the posterior distribution

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta) \prod_{i=1}^{N_{\text{items}}} p_i(x_i|\theta) \tag{1}$$

where the maximum likelihood estimate corresponds to using an uniform $\pi(\theta)$.

The objective of a testing session is to ascertain the ability of a new respondent, efficiently approximating the statistics of Eq. 1. In this sense, Eq. 1 is considered the *true estimate* of a person's ability. In computer adaptive testing (CAT), items are presented sequentially to a respondent. So, it is natural to base the choice of the next item on a combination of the current ability estimate and the properties of the ability estimate implied by the next item. In concrete terms, at step t of a test, the items $\boldsymbol{\alpha}_t = (\alpha_1, \alpha_2, \ldots, \alpha_{t-1}, \alpha_t)$ have been answered by a respondent, from which a running ability estimate is obtained. Commonly, this estimate is based on statistics of the distribution

$$\tilde{\pi}(\theta|\mathbf{x}_t) \propto \pi(\theta) \prod_{s=1}^{t} p_{\alpha_s}(x_{\alpha_s}|\theta), \tag{2}$$

where $\mathbf{x}_t = (x_{\alpha_1}, x_{\alpha_2}, \ldots x_{\alpha_t})$ are the observed responses at step t. Then, the choice of item α_{t+1} is made conditional on this estimate. The item selector, conditional on the ability estimate, computes a given criterion for each of the remaining $N_{\text{items}} - t$ as a basis for making a decision.

2.1 Prior art

The oldest and perhaps most-popular CAT methodology is based on the principle of reducing the asymptotic variance of the ability estimate. This method chooses the item for step $t+1$, conditional on the point ability estimate at step t, $\hat{\theta}_t$ (commonly the expectation of Eq. 2), that has the maximum local item-wise *Fisher information*

$$I_i(\hat{\theta}_t) = -\left.\frac{\partial^2}{\partial\theta^2} \sum_{k=1}^{N_{\text{levels}}} w_{ik} \log p_i(k|\theta)\right|_{\theta=\hat{\theta}_t}. \tag{3}$$

At step t the CAT is not privy to the response to the next item so Eq. 3 requires a weighted sum over the potential responses. Typically one resolves the weights w_{ik} self-consistently using the IRT model by setting them to $w_{ik} = p_i(k|\hat{\theta}_t)$, so that they compute an expectation (Magis, 2015), though sometimes uniform weights $w_{ik} = 1/N_{\text{levels}}$ are used. In this manuscript we will assume that the weights correspond to the former.

The Fisher information method while computationally expedient has several known limitations. First, the method adjudicates items conditional on the current running ability estimate. This quantity is not well-characterized early-on in an exam. A class of slight modifications to this criterion take ability uncertainty into account by computing an expectation of the Fisher information over a distribution of ability values (Owen, 1975; van der Linden, 1998; van der Linden & Ren, 2020; Ueno, 2013; Choi & Swartz, 2009):

$$\text{Bayesian Fisher information} = \int \tilde{\pi}(\theta|\mathbf{x}_t) I_i(\theta) \mathrm{d}\theta. \tag{4}$$

Information theoretic alternatives to the Fisher information are also motivated by resolving this issue, for example the *global information* method of Chang & Ying (1996),

$$\begin{aligned}\text{Global information} &= \mathbb{E}_\theta\left[\sum_k p_i(k|\theta)\log\frac{p_i(k|\theta)}{p_i(k|\hat{\theta}_t)}\right]\\ &= \mathbb{E}_{x_i}\left[\mathcal{D}\left(\tilde{\pi}(\theta|\mathbf{x}_t, x_i = k)\parallel\tilde{\pi}(\theta|\mathbf{x}_t)\right)\right] + \mathcal{D}_{x_i}[\tilde{p}_i^{(t)}\parallel p_i(k|\hat{\theta}_t)],\end{aligned} \tag{5}$$

where $\mathcal{D}(q(\theta)\parallel p(\theta)) = \mathbb{E}_{q(\theta)}\log[q(\theta)/p(\theta)]$, $\mathcal{D}_x(p(x)\parallel q(x)) = \sum_k p(k)\log(p(k)/q(k))$, and $x_i \sim \tilde{p}_i^{(t)}$ for

$$\tilde{p}_i^{(t)}(k) = \int p_i(k|\theta)\tilde{\pi}(\theta|\mathbf{x}_t)\mathrm{d}\theta. \tag{6}$$

Other related information criteria that involve the KL divergence between the next ability estimate and the current ability estimate also exist (Sorrel et al., 2020; Wang & Chang, 2011; Weissman, 2007; Wang et al., 2020).

Second, the Fisher information provides an inaccurate approximation of the estimate precision when the number of observed items is small. Instead, one may directly compute the item-specific conditional variance (van der Linden, 1998)

$$\text{Bayesian variance} = \mathrm{Var}\left[\theta|\mathbf{x}_t, \alpha_{t+1} = i\right]. \tag{7}$$

Third, greedy item selection methods have highly stereotypical item trajectories and poor item exposure. To address this issue, explicit and complex exposure controls exist (Georgiadou et al., 2007; Han, 2018), including by using randomness in the selection procedure Barrada et al. (2008), sampling items according to an ad-hoc function of the item-wise local Fisher information. Implementing these methods is challenging because they require tuning. The stochastic method for instance also relies on adaptive dampening of the sampling probabilities.

While various prior methods target reduction of the posterior estimate variance, they do not consider whether the posterior ability estimate is well-calibrated. Zhuang et al. (2023) introduced a gradient-based method where they select a subset of items that most closely matches the gradient of the likelihood function at an estimate of the true full-bank ability estimate.

2.2 Related methodologies outside of traditional CAT

CAT can be viewed as a particular application of Bayesian Optimal Experimental Design (BOED), which is a broad framework for choosing the next *experiment* or measurement for learning about a system based on maximizing a given utility (Rainforth et al., 2023). Unsurprisingly, many of the methods common to CAT have analogues in BOED, for instance in using Bayesian information theoretic criteria (Sebastiani & Wynn, 2000; Bernardo, 1979) or Frequentist experiment/itemwise Fisher information (Smith, 1918). The most common criterion in modern BOED is the expected information gain (EIG), and many stochastic methods for approximating this quantity exist (Laínez-Aguirre et al., 2015; Foster et al., 2020; Zaballa & Hui, 2023; Goda et al., 2022). However, unlike in CAT, there is not a strong motivation to use stochastic selection in order to improve exposure for BOED experiments.

3 Methods

Our main novel theoretical contribution is that we frame the CAT through the lens of model selection/model averaging, rather than directly as an optimization problem. At a given step t, the choice of the next item α_{t+1} is analogous to choosing among $N_{\text{items}} - t$ choices for the next ability estimate $\tilde{\pi}(\theta|\mathbf{x}_{t+1})$. If the full bank estimate were known then we could compute the item-specific ability model discrepancy measure

$$\mathcal{D}\left(\pi(\theta|\boldsymbol{x})\parallel\tilde{\pi}(\theta|\mathbf{x}_{t+1})\right) = \int\pi(\theta|\boldsymbol{x})\log\frac{\pi(\theta|\boldsymbol{x})}{\tilde{\pi}(\theta|\mathbf{x}_{t+1})}\mathrm{d}\theta \tag{8}$$

and use it as the basis for item selection. In particular, information theoretic model averaging techniques use the negative exponential of the model discrepancy to determine model weights (Akaike, 1978; Bozdogan, 1987; Dormann et al., 2018; Wagenmakers & Farrell, 2004; Yao et al., 2018). Our aim is to perform stochastic selection using these weights. However we first address the challenge of approximating Eq. 8, acknowledging that both distributions in the KL divergence are unknown at time t.

3.1 Plug-in estimation of the expectation of Eq. 8

Like in all CAT methods, we need to resolve our objective (Eq. 8) under incomplete observation. We pursue the usual strategy of computing an expectation. Computing the expectation of Eq. 8 exactly requires specifying $(N_{\text{items}} - t) \times N_{\text{levels}}$ different marginal posterior distributions, each of which is challenging to compute. In order to make the method tractable, we develop a mean field estimate of the expectation of Eq. 8. In this estimate, we ignore the coupling between $\pi(\theta|\boldsymbol{x})$ and the response to the next item, plugging in the expectation of $\pi(\theta|\boldsymbol{x})$, the marginal posterior,

$$\pi(\theta|\mathbf{x}_t) = \mathbb{E}_{\mathbf{z}_t}\pi(\theta, \mathbf{z}_t|\mathbf{x}_t), \tag{9}$$

into Eq. 8. In Eq. 9, $\mathbf{z}_t$ are the responses that have not yet been observed at step t. Still, the expectation in Eq. 9 is intractable. Fortunately, this expectation is amenable to Variational Bayesian Expectation Maximization (VBEM).

VBEM (Bernardo et al., 2003) allows us to iteratively approximate Eq. 9, producing a sequence of estimates $q_\theta^{(0)}, q_\theta^{(1)}, q_\theta^{(2)}, \ldots$ that obey the descent property of the Majorization Minimization (MM) algorithm (Lange et al., 2000; de Leeuw & Lange, 2006; Lange et al., 2021; Wu & Lange, 2010) so that $\mathcal{D}(q_\theta^{(m+1)} \parallel \pi(\theta|\mathbf{x}_t)) \leq \mathcal{D}(q_\theta^{(m)} \parallel \pi(\theta|\mathbf{x}_t))$. Based on $q_\theta^{(m)}$, one can easily compute a corresponding set of response probabilities for all unobserved items. The VBEM update equations have the explicit form

$$\log q_{\mathbf{z}_t,j}^{(m+1)}(k) = \text{const}_j^{(m+1)} + \int \log p_j(k|\theta) q_\theta^{(m)}(\theta) \mathrm{d}\theta \tag{10}$$

$$\log q_\theta^{(m+1)}(\theta) = \text{const}^{(m+1)} + \log \pi(\theta) + \sum_{j \in \boldsymbol{\alpha}_t} \log p_j(x_j|\theta) + \sum_{j \notin \boldsymbol{\alpha}_t} \sum_k q_{\mathbf{z}_t,j}^{(m+1)}(k) \log p_j(k|\theta). \tag{11}$$

Then, after some number of EM iterations M, we can compute the plug-in criterion

$$\Delta_t^{(i)} = \sum_k q_{\mathbf{z}_t,i}^{(M)}(k) \mathcal{D}\left(q_\theta^{(M)}(\theta) \parallel \tilde{\pi}(\theta|\mathbf{x}_t, x_i = k)\right), \tag{12}$$

where $\mathcal{D}$ is the KL divergence. Technically, Eq. 9, rather than the commonly-used Eq. 2, is the best estimate of the ability at step t, an observation that we will save for the Discussion.

3.2 Stochastic item selector

It is our desire to hedge in the choice of the next item with frequency statistics that imply Bayesian model averaging (Hinne et al., 2020; Hoeting et al., 1999) of the corresponding per-item ability estimates. To do so, we draw the next item $i \notin \boldsymbol{\alpha}_t$, according to

$$\alpha_{t+1} \sim \text{Categorical}(\mathbf{w}_t) \qquad\qquad w_t^{(i)} = \frac{\exp\left(-\Delta_t^{(i)}\right)}{\sum_{j \notin \boldsymbol{\alpha}_t} \exp\left(-\Delta_t^{(j)}\right)}, \tag{13}$$

where the categorical distribution is defined over the $N_{\text{items}} - t$ items that have not yet been administered at time step t.

3.3 Relationship to cross validation

We can rewrite the discrepancy (Eq. 8) to remove the explicit dependence on $\tilde{\pi}$,

$$\begin{aligned}\mathcal{D}\left(\pi(\theta|\boldsymbol{x}) \parallel \tilde{\pi}(\theta|\mathbf{x}_{t+1})\right) &= \int \pi(\theta|\mathbf{x}) \log \frac{\tilde{p}_i^{(t)}(x_{t+1})\pi(\theta|\boldsymbol{x})}{p_i(x_{t+1}|\theta)\tilde{\pi}_t(\theta|\mathbf{x}_t)} \mathrm{d}\theta \\ &= \int \pi(\theta|\boldsymbol{x}) \log \frac{\tilde{p}_i^{(t)}(x_{t+1})}{p_i(x_{t+1}|\theta)} \mathrm{d}\theta + \mathcal{D}(\pi(\theta|\boldsymbol{x}) \parallel \tilde{\pi}(\theta|\mathbf{x}_t))\end{aligned} \tag{14}$$

where

$$\tilde{p}_i^{(t)}(k) = \int p_i(k|\theta)\tilde{\pi}(\theta|\mathbf{x}_t)\mathrm{d}\theta,$$

and note that while the second term in the last line of Eq. 14 depends on the response for the next item, it does not depend on the choice of the next item. We can then relate the discrepancy to leave one out (LOO) cross validation, expanding the first term in Eq. 14

$$\begin{aligned}\mathcal{D}\left(\pi(\theta|\boldsymbol{x}) \parallel \tilde{\pi}(\theta|\mathbf{x}_{t+1})\right) &= \mathcal{D}(\pi(\theta|\boldsymbol{x}) \parallel \tilde{\pi}(\theta|\mathbf{x}_t)) + \int \pi(\theta|\boldsymbol{x}) \log \frac{\tilde{p}_i^{(t)}(x_i)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x})p_i(x_i|\theta)} \mathrm{d}\theta \\ &= \mathcal{D}(\pi(\theta|\boldsymbol{x}) \parallel \tilde{\pi}(\theta|\mathbf{x}_t)) + S[\pi(\theta|\boldsymbol{x})] - \mathcal{D}(\pi(\theta|\boldsymbol{x}) \parallel \tilde{\pi}(\theta|\boldsymbol{x} \setminus \{x_i\})) + \log \frac{\tilde{p}_i^{(t)}(x_i)}{\tilde{p}_i^{\mathrm{LOO}}(x_i)}\end{aligned} \tag{15}$$

where, $\tilde{\pi}(\theta|\boldsymbol{x} \setminus \{x_i\})$, the ability estimate when leaving out x_i follows Bayes rule, $p_i(x_i|\theta)\tilde{\pi}(\theta|\boldsymbol{x} \setminus \{x_i\}) = \pi(\theta|\boldsymbol{x})\tilde{p}_i^{\mathrm{LOO}}(x_i)$ and $\tilde{p}_i^{\mathrm{LOO}}(x_i) = \int p(x_i|\theta)\tilde{\pi}(\theta|\boldsymbol{x} \setminus \{x_i\})\mathrm{d}\theta$ is the corresponding LOO mass function for item i. In this representation, only the last two terms in Eq. 15 depend on the item choice. So in minimizing the discrepancy, one is also selecting the item that if left out would yield the biggest discrepancy.

3.4 Numerical implementation

We coded two independent implementations of our methodology as applied to the Graded Response Model: one in Python (redacted) and one in Golang (redacted). Within our implementation we approximated all integrals using trapezoid approximations with 200 equally spaced grid points. We used $M = 5$ iterations to approximate the marginal posterior distributions (Eq. 11).

4 Results

In producing the following results, for each scale, we simulated item responses for 500 respondents for each true underlying ability of $\theta \in \{-3, -2.5, -2, \ldots, 2.5, 3\}$. Then we put each respondent's item responses through each CAT item selection method, obtaining ability estimates at given test lengths. The methods evaluated are greedy selection via the Fisher Information (Eq. 3), Bayesian Fisher information (Eq. 4), Global information (Eq. 5), Bayesian variance (Eq. 7), ability estimate discrepancy (Eq. 12), and our stochastic selection method (Eq. 13). Finally, we also computed ability estimates for each simulated respondent based on all of their item responses. In the main text we report on only the four mental scales of the WD-FAB. Please see the Supplement Results for the corresponding physical scale results.

4.1 Testing error

Figures 1, 2 and 3 provide different measures of ability estimation error in the context of computer adaptive testing. Fig. 1 displays values of the discrepancy (Eq. 8) conditional on the scale, item selection method, test length at stopping (5, 10, 20, 30, 40 items), and true fixed ability used in simulating CAT responses. Using the Fisher information and global information selectors, there are some situations in which the discrepancy increases as the test length increases for an intermediate range of test lengths before dropping. On the other hand, the Bayesian variance and the methods based on our criterion (Eq. 12) reliably decrease the discrepancy as the test length increases. Failure to decrease this discrepancy suggests that a selection

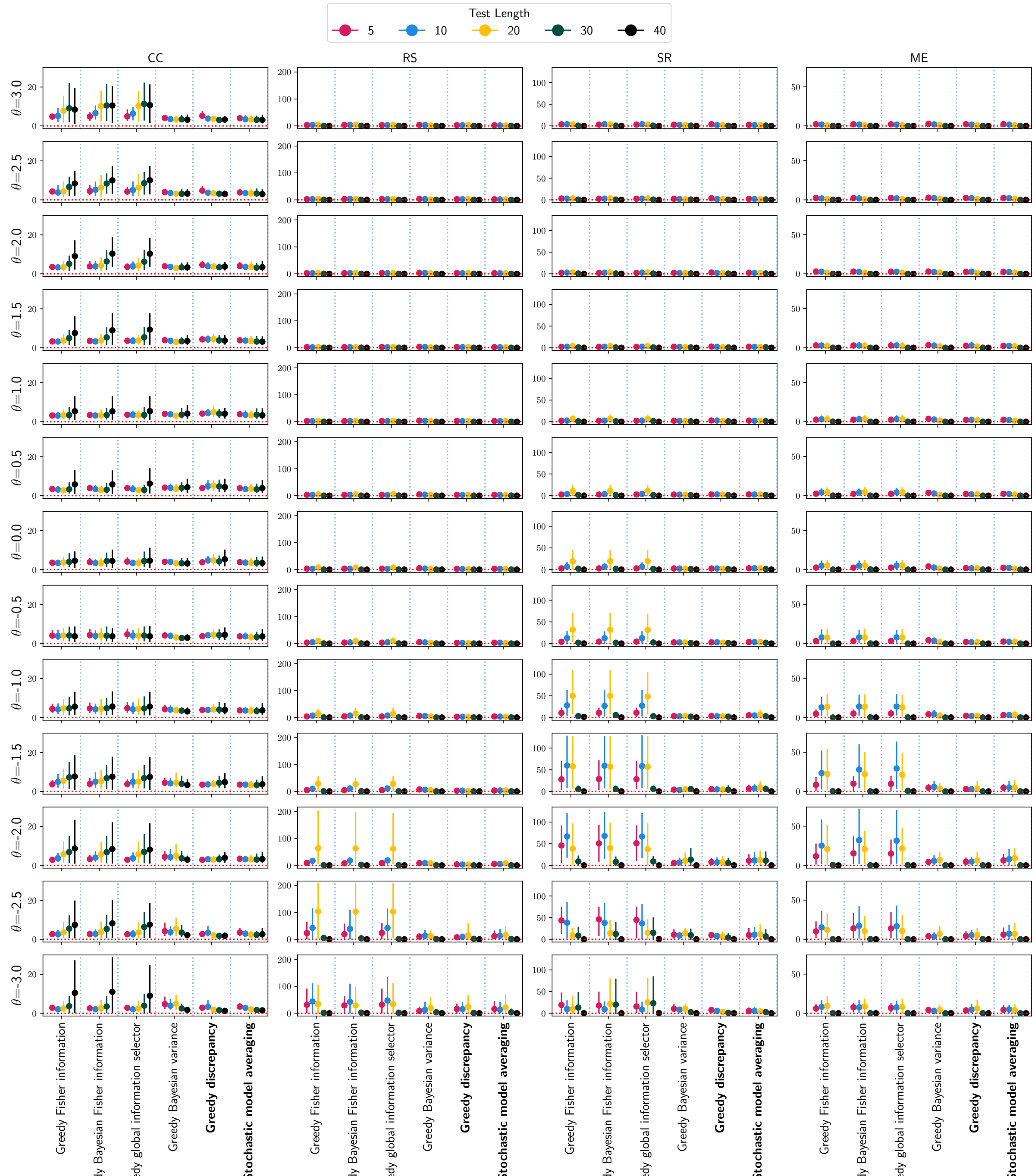

Figure 1: **Ability estimate discrepancy** $\mathcal{D}(\pi(\theta|\mathbf{x}) \parallel \tilde{\pi}(\theta|\mathbf{x}_t))$ (mean and middle 80% interval) conditional on score θ used to generate response sets, by scale, item selection method, and test length t, for mental function scales of the WD-FAB. Lower is better.

procedure generates item subsets that provide inaccurate ability estimates when used as whole-distribution A/B comparisons between individuals because those estimates are ill-calibrated.

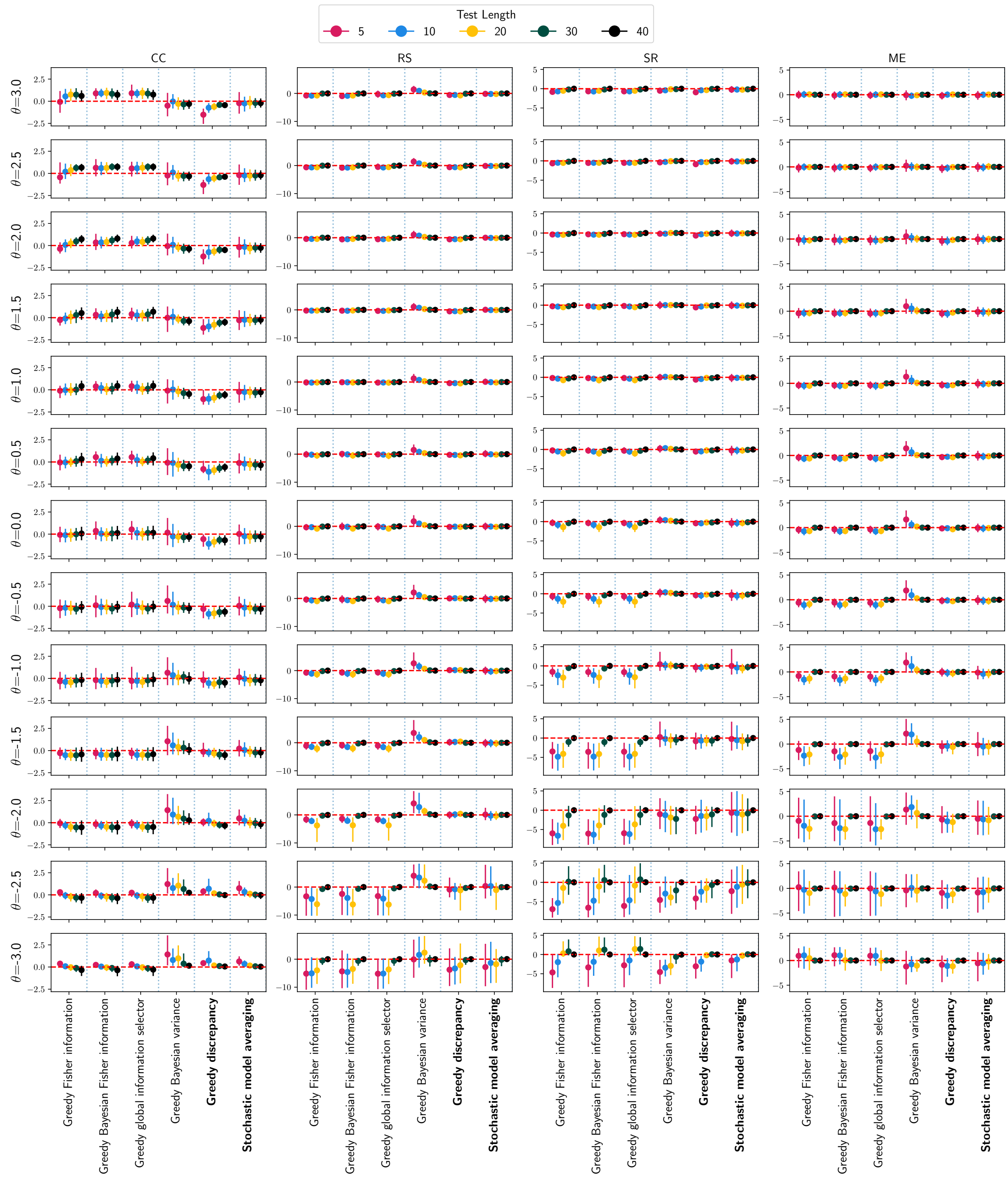

Figure 2: **Error in means** ($\int \theta \tilde{\pi}(\theta|\mathbf{x}_t)\mathrm{d}\theta - \int \theta \pi(\theta|\mathbf{x})\mathrm{d}\theta$) (mean and middle 80% interval) conditional on true score θ by scale, item selection method, and test length t, for mental function scales of the WD-FAB.

In many CAT/IRT based instruments, the mean ability is used in order to characterize a respondent. Fig. 2 presents statistics of the mean ability error (mean and middle 80% coverage) across the different simulation configurations. In Fig. 3, we provide statistics of the absolute value of this error across simulations.

The error distributions are highly variable across these attributes. Generally, the magnitude of the error decreased as the test length increased. For most scales, there is a region of abilities for which all item

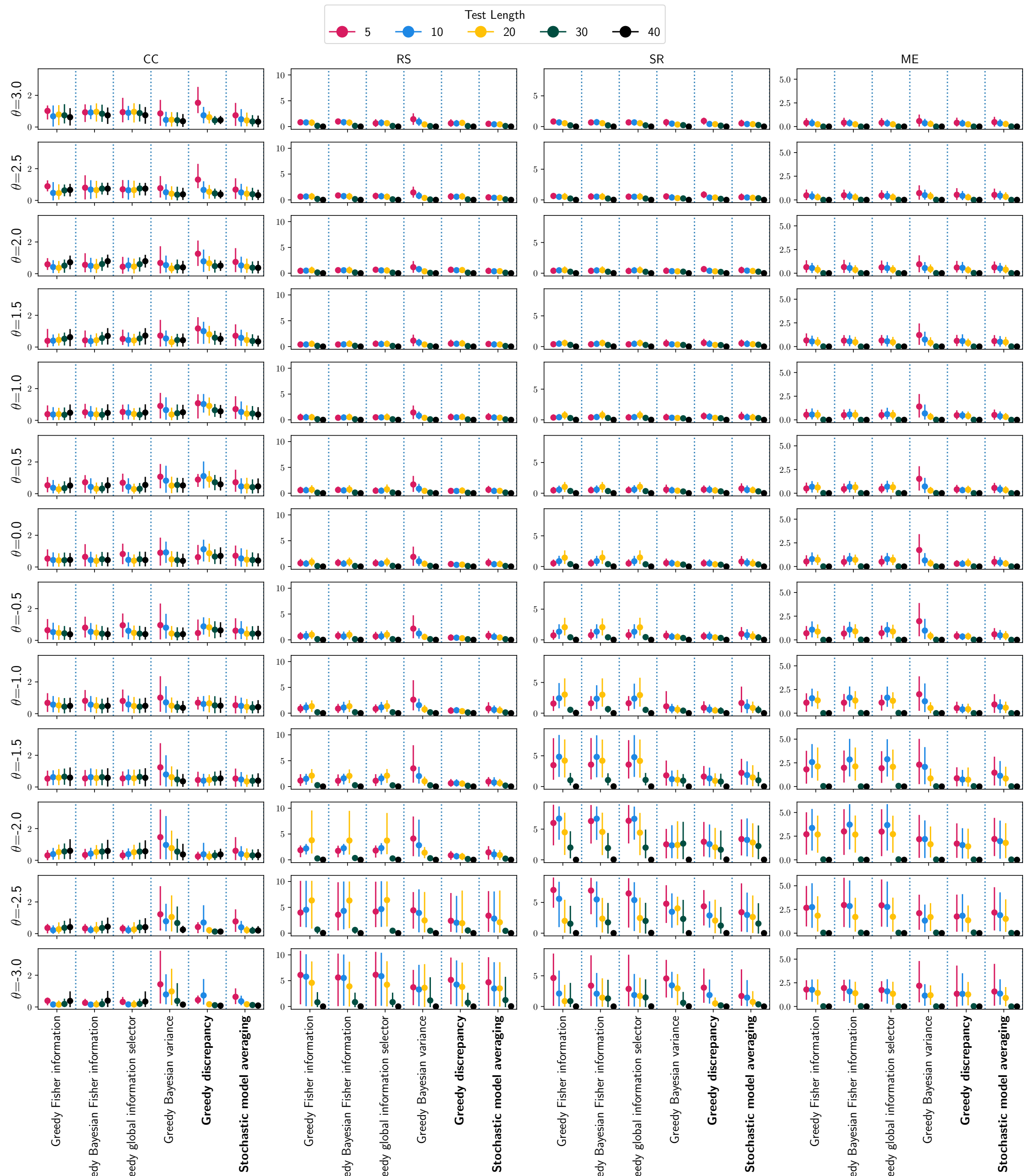

Figure 3: **Absolute error in means** ($|\int \theta\tilde{\pi}(\theta|\mathbf{x}_t)\mathrm{d}\theta - \int \theta\pi(\theta|\mathbf{x})\mathrm{d}\theta|$) (mean and middle 80% interval) conditional on true score θ by scale, item selection method, and test length t, for mental function scales of the WD-FAB. Lower is better.

selectors produced small errors. No single selection method had the lowest errors in all situations, though generally the stochastic selector performed most-consistently well.

Often, the posterior variance is used to define a cutoff for a CAT stopping rule. The standard deviation of the posterior ability estimates is presented in Fig. 4 for the different simulation configurations. In these

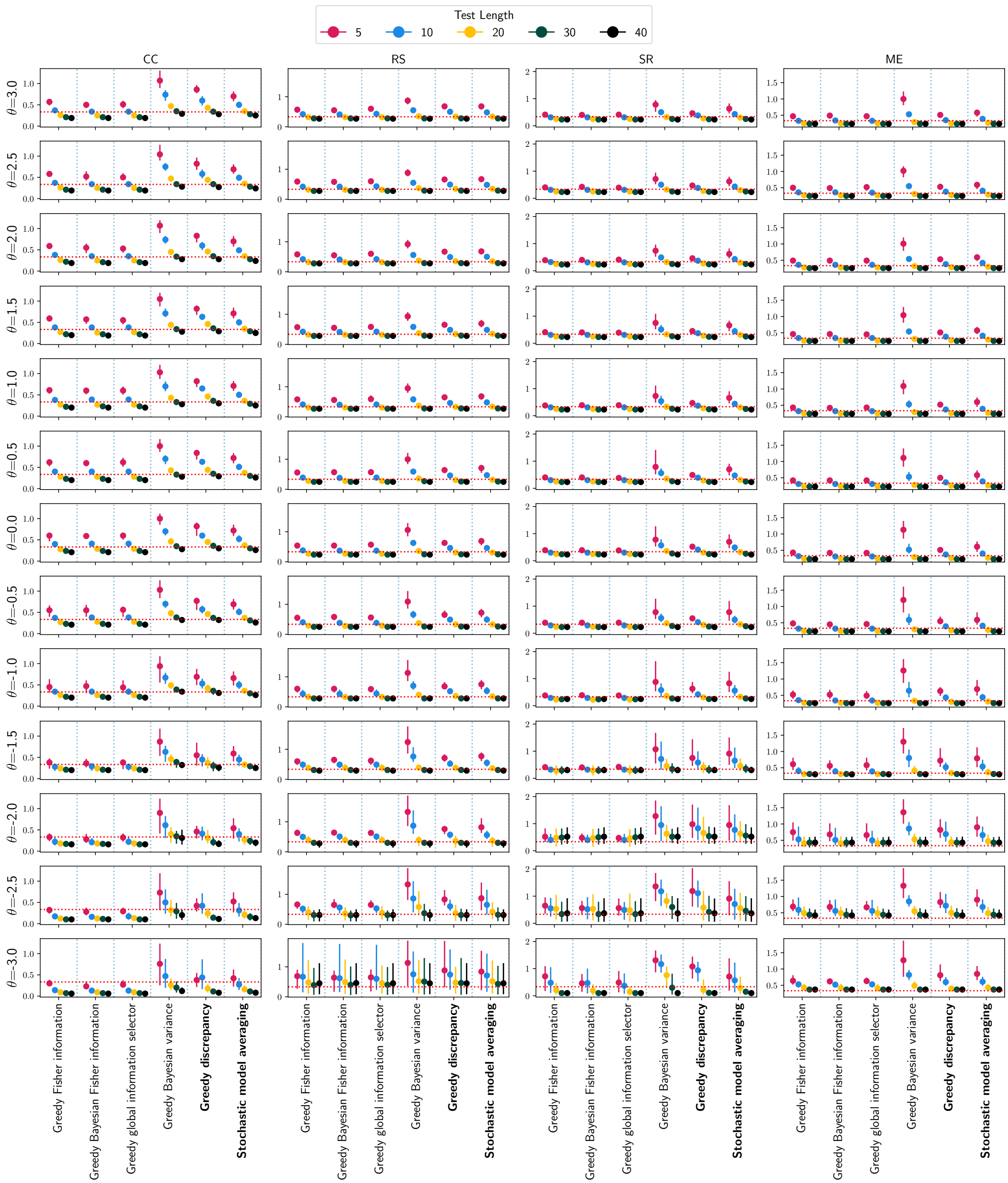

Figure 4: **Standard deviation of ability estimates** ($\sqrt{\mathbf{Var}_t(\theta)}$) (mean and middle 80% percentile) conditional on true score θ by scale and item selection method, for mental function scales of the WD-FAB. Used as stopping criteria for CAT. Lower is better.

simulations, it is clear that the two Fisher methods and the global information method provide the lowest posterior ability standard deviations. However, in light of Figures 1, 2, 3, it is clear that these ability estimates are ill-calibrated. They are terminating quicker than they should and settling on sub-optimal ability estimates.

4.2 Item exposure

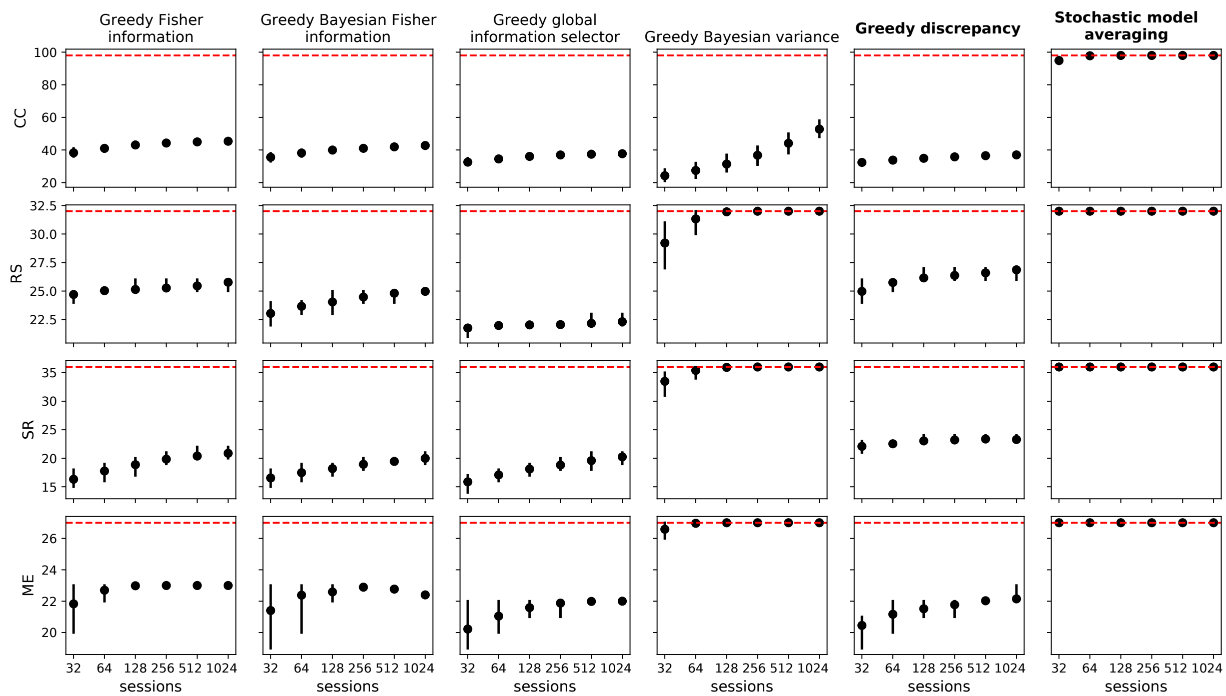

Figure 5: **Item exposure statistics** (mean and middle 80% interval), for each of the given item selection methods across a given number of CAT sessions, for mental function scales of the WD-FAB. The dashed line represents the maximum possible exposure per scale. Higher is better.

Fig. 5 compares the different item selection methods on the basis of item exposure across sessions (with 12 items presented per scale) with randomly distributed abilities. In this figure, for each simulation configuration, we counted the number of unique items seen for each scale across replications of the given number of CAT sessions. For example, for the scale “ME,” we estimate that in each set of 32 sessions approximately 22 items are exposed on average, though with wide variance. As the number of sessions increases, the number of exposed items increases. Of the greedy methods, the Bayesian variance method has the best item exposure. For some scales, the Bayesian variance method performed almost as well as the best selection method, the stochastic selector based on Eq. 13. The stochastic selector successfully exposed all items for all scales in all the scenarios tested.

5 Discussion

In this manuscript we have introduced stochastic selection for CAT where the frequency statistics of the next item correspond to Bayesian model averaging of corresponding discrepancy weighted next item-wise ability estimates. In formulating our method we identified the ability discrepancy (the KL divergence between the next item ability estimate and the full bank true ability estimate) as a selection criterion. We provided a computationally expedient plugin version of our criterion based on variational Bayesian expectation maximization. Using simulations of the new selector (and other selectors for comparison), on the WD-FAB, we found our new stochastic selector to have both superior item exposure properties while not compromising in terms of accuracy. Additionally, the simulations showed that unlike the Fisher information methods, the new selection methods (whether greedy or stochastic) are not over-confident in estimating scoring error. This fact implies that the new methods are less likely to settle on a poor ability estimate. Beyond characterizing a point estimate for ability, using the discrepancy as a criterion optimizes the whole-distribution ability estimate, which implies more-accurate A/B tests when comparing scores between different respondents. Finally, the computationally expensive portion of our overall approach is in computing the marginal posterior ability estimate. As we will discuss, this quantity is the true ability estimate at step t and should be computed and used in all other selection methods. For this reason, our criterion is of similar computational complexity to the other Bayesian criterion mentioned in this manuscript.

5.1 What should the ability estimate be at step t?

In formulating our method, we assume that one is using a scoring methodology similar to what is commonly used in the literature – using the likelihood of the items observed up to step t. Recall that we call the posterior ability estimate obtained by this method $\tilde{\pi}(\theta|\mathbf{x}_t)$, making a distinction between this quantity and $\pi(\theta|\mathbf{x}_t)$, the marginal posterior ability at step t. The latter estimate differs from the former in that it also accounts for the fact that the $N_{\text{items}} - t$ unobserved items at time t will also impact the final ability estimate. The latter is a better estimate of the ability because it is consistent with both the observed and unobserved items being drawn from the same underlying conditional distribution. For this reason, it should also be used in all selection methods in place of $\tilde{\pi}$ when taking expectations over unknown responses and in both the running and final score estimates. In a follow-up to this manuscript, we will elaborate on this point.

5.2 Why ensembling?

Focusing on efficiency, there are reasons to think why randomization in CAT would be sub-optimal. If the objective is to optimize a given criterion, then not always choosing the exact optimal item would seem to result in a less efficient CAT. As we have shown for the WD-FAB, this assumption did not hold. On the other hand, there are at least a couple a-priori explanations in support of our findings. First, in the context of prediction, Le & Clarke (2022) has shown that model averaging is asymptotically better than model selection. Second, each criterion requires resolving unknown future responses. Since the true ability of the respondent is unknown, the statistics of these responses is unknown. However, our method uses the *correct* item response probabilities in computing the expectation in Eq. 12.

5.3 Limitations and extensions

In using the variational Bayesian EM estimates for the marginal item probability mass functions in order to compute the item-wise expectations of Eq. 12, we are using the optimal item probabilities provided by the given IRT model. However, one may also be able to improve the accuracy of this expectation by using different IRT models that are more-tuned to accuracy than interpretability (Chang et al., 2019; 2023), so long as one accounts for unobserved items.

The estimate of the criterion of Eq. 8 in the form of the the mean field plugin estimator in Eq. 12 trades accuracy for computational efficiency. One could more-accurately compute this expectation by developing a version of Eq. 12 that preserves the coupling between $\pi(\theta|\mathbf{x})$ and the response to the next item.

This work was focused on improving the assessment of the WD-FAB, a factorized multidimensional IRT model. We found generally, across all scales (dimensions) that our model ensembling stochastic selector outperformed the other commonly used selection methods that we tested. Your mileage my vary when trying these methods with other instruments.

While we formulate our methodology assuming a multidimensional ability parameter θ, it would likely take additional work in order to adapt this method to non-factorized multidimensional instruments. Additional controls might be needed in order to balance out the administration of the different scales for instance.

References

Hirotugu Akaike. On the Likelihood of a Time Series Model. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 27(3-4):217–235, 1978. ISSN 1467-9884. doi: 10.2307/2988185.

Juan Ramón Barrada, Julio Olea, Vicente Ponsoda, and Francisco José Abad. Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, 61(2):493–513, 2008. ISSN 2044-8317. doi: 10.1348/000711007X230937.

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

Jos M. Bernardo. Expected Information as Expected Utility. *The Annals of Statistics*, 7(3):686–690, 1979. ISSN 0090-5364.

R. Darrell Bock, David Thissen, and Michele F. Zimowski. IRT Estimation of Domain Scores. *Journal of Educational Measurement*, 34(3):197–211, 1997. ISSN 1745-3984. doi: 10.1111/j.1745-3984.1997.tb00515.x.

Ulf Böckenholt and Thorsten Meiser. Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, 70(1):159–181, 2017. ISSN 2044-8317. doi: 10.1111/bmsp.12086.

Hamparsum Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, September 1987. ISSN 1860-0980. doi: 10.1007/BF02294361.

Paul-Christian Bürkner. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, 100:1–54, November 2021. ISSN 1548-7660. doi: 10.18637/jss.v100.i05.

David Cella, Richard Gershon, Jin-Shei Lai, and Seung Choi. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 16 Suppl 1:133–141, 2007. ISSN 0962-9343. doi: 10.1007/s11136-007-9204-6.

David Cella, William Riley, Arthur Stone, Nan Rothrock, Bryce Reeve, Susan Yount, Dagmar Amtmann, Rita Bode, Daniel Buysse, Seung Choi, Karon Cook, Robert Devellis, Darren DeWalt, James F. Fries, Richard Gershon, Elizabeth A. Hahn, Jin-Shei Lai, Paul Pilkonis, Dennis Revicki, Matthias Rose, Kevin Weinfurt, Ron Hays, and PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology*, 63(11):1179–1194, November 2010. ISSN 1878-5921. doi: 10.1016/j.jclinepi.2010.04.011.

Hua-Hua Chang and Zhiliang Ying. A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement*, 20(3):213–229, September 1996. ISSN 0146-6216. doi: 10.1177/014662169602000303.

Joshua C. Chang, Shashaank Vattikuti, and Carson C. Chow. Probabilistically-autoencoded horseshoe-disentangled multidomain item-response theory models. *arXiv:1912.02351 [cs, stat]*, December 2019.

Joshua C. Chang, Patrick Fletcher, Jungmin Han, Ted L. Chang, Shashaank Vattikuti, Bart Desmet, Ayah Zirikly, and Carson C. Chow. Sparse encoding for more-interpretable feature-selecting representations in probabilistic matrix factorization. In *International Conference on Learning Representations*, 2021.

Joshua C. Chang, Julia Porcino, Elizabeth K. Rasch, and Larry Tang. Regularized Bayesian calibration and scoring of the WD-FAB IRT model improves predictive performance over marginal maximum likelihood. *PLOS ONE*, 17(4):e0266350, April 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0266350.

Joshua C. Chang, Carson C. Chow, and Julia Porcino. Autoencoded sparse Bayesian in-IRT factorization, calibration, and amortized inference for the Work Disability Functional Assessment Battery. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 3961–3976. PMLR, April 2023.

Seung W. Choi and Richard J. Swartz. Comparison of CAT Item Selection Criteria for Polytomous Items. *Applied psychological measurement*, 33(6):419–440, September 2009. ISSN 0146-6216. doi: 10.1177/0146621608327801.

Geoffrey Converse, Mariana Curi, and Suely Oliveira. Autoencoders for Educational Assessment. In Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin (eds.), *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pp. 41–45. Springer International Publishing, 2019. ISBN 978-3-030-23207-8.

Jan de Leeuw and Kenneth Lange. Sharp Quadratic Majorization in One Dimension. September 2006.

Carsten F. Dormann, Justin M. Calabrese, Gurutzeta Guillera-Arroita, Eleni Matechou, Volker Bahn, Kamil Bartoń, Colin M. Beale, Simone Ciuti, Jane Elith, Katharina Gerstner, Jérôme Guelat, Petr Keil, José J. Lahoz-Monfort, Laura J. Pollock, Björn Reineking, David R. Roberts, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Simon N. Wood, Rafael O. Wüest, and Florian Hartig. Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88(4):485–504, 2018. ISSN 1557-7015. doi: 10.1002/ecm.1309.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments, February 2020.

Elissavet G. Georgiadou, Evangelos Triantafillou, and Anastasios A. Economides. A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, 5(8), May 2007. ISSN 1540-2525.

Takashi Goda, Tomohiko Hironaka, Wataru Kitade, and Adam Foster. Unbiased MLMC Stochastic Gradient-Based Optimization of Bayesian Experimental Designs. *SIAM Journal on Scientific Computing*, 44(1): A286–A311, February 2022. ISSN 1064-8275. doi: 10.1137/20M1338848.

Kyung (Chris) Tyek Han. Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, 15:7, March 2018. ISSN 1975-5937. doi: 10.3352/jeehp.2018.15.7.

Max Hinne, Quentin F. Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. A Conceptual Introduction to Bayesian Model Averaging. *Advances in Methods and Practices in Psychological Science*, 3(2):200–215, June 2020. ISSN 2515-2459. doi: 10.1177/2515245919898657.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science*, 14(4):382–417, November 1999. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009212519.

Jason C. Immekus, Kate E. Snyder, and Patricia A. Ralston. Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research. *Frontiers in Education*, 4, 2019. ISSN 2504-284X. doi: 10.3389/feduc.2019.00045.

Alan M. Jette, Pengsheng Ni, Elizabeth Rasch, Elizabeth Marfeo, Christine McDonough, Diane Brandt, Lewis Kazis, and Leighton Chan. The Work Disability Functional Assessment Battery (WD-FAB). *Physical Medicine and Rehabilitation Clinics*, 30(3):561–572, August 2019. ISSN 1047-9651, 1558-1381. doi: 10.1016/j.pmr.2019.03.004.

Vincent Kieftenbeld and Prathiba Natesan. Recovery of Graded Response Model Parameters: A Comparison of Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement*, 36(5):399–419, July 2012. ISSN 0146-6216. doi: 10.1177/0146621612446170.

Neal Kingston, Linda Leary, and Larry Wightman. 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, 1985. ISSN 2330-8516. doi: 10.1002/j.2330-8516.1985.tb00119.x.

José M. Laínez-Aguirre, Linas Mockus, and Gintaras V. Reklaitis. A stochastic programming approach for the Bayesian experimental design of nonlinear systems. *Computers & Chemical Engineering*, 72:312–324, January 2015. ISSN 0098-1354. doi: 10.1016/j.compchemeng.2014.06.006.

Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000. ISSN 1061-8600. doi: 10.2307/1390605.

Kenneth Lange, Joong-Ho Won, Alfonso Landeros, and Hua Zhou. Nonconvex Optimization via MM Algorithms: Convergence Theory. pp. 1–22. March 2021. doi: 10.1002/9781118445112.stat08295.

Tri M. Le and Bertrand S. Clarke. Model Averaging Is Asymptotically Better Than Model Selection For Prediction. *Journal of Machine Learning Research*, 23(33):1–53, 2022. ISSN 1533-7928.

Frederic M. Lord. Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, 48(3):477–482, September 1983. ISSN 1860-0980. doi: 10.1007/BF02293689.

Yong Luo and Hong Jiao. Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, 78(3):384–408, 2018.

David Magis. A Note on the Equivalence Between Observed and Expected Information Functions With Polytomous IRT Models. *Journal of Educational and Behavioral Statistics*, 40(1):96–105, February 2015. ISSN 1076-9986. doi: 10.3102/1076998614558122.

Elizabeth Marfeo, Pengsheng Ni, Mark Meterko, Molly Marino, Kara Peterik, Christine McDonough, Elizabeth K. Rasch, Diane Brandt, Leighton Chan, and Alan Jette. 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, August 2016. ISSN 0272-9490. doi: 10.5014/ajot.2016.70S1-RP402B.

Elizabeth E. Marfeo, Pengsheng Ni, Christine McDonough, Kara Peterik, Molly Marino, Mark Meterko, Elizabeth K. Rasch, Leighton Chan, Diane Brandt, and Alan M. Jette. 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, March 2018. ISSN 1573-3688. doi: 10.1007/s10926-017-9710-5.

Elizabeth E. Marfeo, Christine McDonough, Pengsheng Ni, Kara Peterik, Julia Porcino, Mark Meterko, Elizabeth Rasch, Lewis Kazis, and Leighton Chan. 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, March 2019. ISSN 1536-5948. doi: 10.1097/JOM.0000000000001521.

Mark Meterko, Elizabeth E. Marfeo, Christine M. McDonough, Alan M. Jette, Pengsheng Ni, Kara Bogusz, Elizabeth K. Rasch, Diane E. Brandt, and Leighton Chan. Work Disability Functional Assessment Battery: Feasibility and Psychometric Properties. *Archives of Physical Medicine and Rehabilitation*, 96(6):1028–1035, June 2015. ISSN 0003-9993. doi: 10.1016/j.apmr.2014.11.025.

Prathiba Natesan. A Review of Bayesian Item Response Modeling: Theory and Applications. *Journal of Educational and Behavioral Statistics*, 36(4):550–552, 2011. ISSN 1076-9986.

Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D. Rubright. Bayesian Prior Choice in IRT Estimation Using MCMC and Variational Bayes. *Frontiers in Psychology*, 7, September 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.01422.

Roger J. Owen. A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing. *Journal of the American Statistical Association*, 70(350):351–356, June 1975. ISSN 0162-1459. doi: 10.1080/01621459.1975.10479871.

Julia Porcino, Beth Marfeo, Christine McDonough, and Leighton Chan. The Work Disability Functional Assessment Battery (WD-FAB): Development and validation review. *TBV – Tijdschrift voor Bedrijfs- en Verzekeringsgeneeskunde*, 26(7):344–349, September 2018. ISSN 1876-5858. doi: 10.1007/s12498-018-0247-0.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian Experimental Design, November 2023.

Lawrence M. Rudner. Implementing the Graduate Management Admission Test Computerized Adaptive Test. In Wim J. van der Linden and Cees A.W. Glas (eds.), *Elements of Adaptive Testing*, pp. 151–165. Springer, New York, NY, 2010. ISBN 978-0-387-85461-8. doi: 10.1007/978-0-387-85461-8_8.

P. Sebastiani and H. P. Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000. ISSN 1467-9868. doi: 10.1111/1467-9868.00225.

Daniel O. Segall and Kathleen E. Moreno. Development of the Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery *. In *Innovations in Computerized Assessment*. Psychology Press, 1999. ISBN 978-1-4106-0252-7.

Eisuke Segawa, Benjamin Schalet, and David Cella. A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. *Quality of Life Research*, 29(1):213–221, January 2020. ISSN 1573-2649. doi: 10.1007/s11136-019-02312-8.

Kirstine Smith. On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they give Towards a Proper Choice of the Distribution of Observations. *Biometrika*, 12(1/2):1–85, 1918. ISSN 0006-3444. doi: 10.2307/2331929.

Miguel A. Sorrel, Juan R. Barrada, Jimmy de la Torre, and Francisco José Abad. Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. *PLOS ONE*, 15(1):e0227196, January 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0227196.

Maomi Ueno. Adaptive Testing Based on Bayesian Decision Theory. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik (eds.), *Artificial Intelligence in Education*, volume 7926, pp. 712–716, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39111-8 978-3-642-39112-5. doi: 10.1007/978-3-642-39112-5_95.

Wim J. van der Linden. Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2):201–216, June 1998. ISSN 1860-0980. doi: 10.1007/BF02294775.

Wim J. van der Linden and Hao Ren. A Fast and Simple Algorithm for Bayesian Adaptive Testing. *Journal of Educational and Behavioral Statistics*, 45(1):58–85, February 2020. ISSN 1076-9986. doi: 10.3102/1076998619858970.

Christian Ventura, Edward Denton, and Emily Van Court. Taking the NREMT Exam. In Christian Ventura, Edward Denton, and Emily Van Court (eds.), *The Emergency Medical Responder: Training and Succeeding as an EMT/EMR*, pp. 155–157. Springer International Publishing, Cham, 2021. ISBN 978-3-030-64396-6. doi: 10.1007/978-3-030-64396-6_19.

Eric-Jan Wagenmakers and Simon Farrell. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, February 2004. ISSN 1069-9384. doi: 10.3758/bf03206482.

Chun Wang and Hua-Hua Chang. Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. *Psychometrika*, 76(3):363–384, 2011. ISSN 1860-0980. doi: 10.1007/s11336-011-9215-7.

Wenyi Wang, Lihong Song, Teng Wang, Peng Gao, and Jian Xiong. A Note on the Relationship of the Shannon Entropy Procedure and the Jensen–Shannon Divergence in Cognitive Diagnostic Computerized Adaptive Testing. *SAGE Open*, 10(1):2158244019899046, January 2020. ISSN 2158-2440. doi: 10.1177/2158244019899046.

Alexander Weissman. Mutual Information Item Selection in Adaptive Classification Testing. *Educational and Psychological Measurement*, 67(1):41–58, 2007. ISSN 0013-1644. doi: 10.1177/0013164406288164.

Ada Woo and Marijana Dragan. Ensuring Validity of NCLEX® With Differential Item Functioning Analysis. *Journal of Nursing Regulation*, 2(4):29–31, January 2012. ISSN 2155-8256. doi: 10.1016/S2155-8256(15)30252-0.

Tong Tong Wu and Kenneth Lange. The MM Alternative to EM. *Statistical Science*, 25(4), November 2010. ISSN 0883-4237. doi: 10.1214/08-STS264.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, 13(3):917–1007, September 2018. ISSN 1936-0975, 1931-6690. doi: 10.1214/17-BA1091.

Vincent D. Zaballa and Elliot E. Hui. Stochastic Gradient Bayesian Optimal Experimental Designs for Simulation-based Inference, June 2023.

Yan Zhuang, Qi Liu, Guanhao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen, Jinze Wu, and Xin Li. A Bounded Ability Estimation for Computerized Adaptive Testing. *Advances in Neural Information Processing Systems*, 36:2381–2402, December 2023.