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

Anonymous authors
Paper under double-blind review

Abstract

The goal of Computer Adaptive Testing (CAT) is to reliably estimate an individual's ability as modeled by an item response theory (IRT) instrument using only a subset of the instrument's items. A secondary goal is to vary the items presented across different testing sessions so that the sequence of items does not become overly stereotypical – we want all items to have an exposure rate sufficiently far from zero. We formulate the optimization problem for CAT in terms of Bayesian information theory, where one chooses the item at each step based on the criterion of the ability model discrepancy – the statistical distance between the ability estimate at the next step and the full-test ability estimate. This viewpoint of CAT naturally motivates a stochastic selection procedure that equates sampling the next item to Bayesian model averaging in the space of ability estimates. Using the NIH Work Disability Functional Assessment Battery (WD-FAB), we evaluate our new methods in comparison to pre-existing methods found in the literature. We find that our stochastic selector has superior properties in terms of both item exposure and test accuracy/efficiency.

1 Introduction

Computer Adaptive Testing (CAT), coupled with Item Response Theory (IRT) is the dominant statistical paradigm behind assessment. Examples in high-stakes standardized testing alone 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 such as in the Patient Reported Outcomes Measurement Information System (PROMIS) instruments (Cella et al., 2010; Segawa et al., 2020). The objective of computer adaptive testing (CAT) is to tailor the administration of an item battery to the ability on the respondent so that one can obtain a precise estimate for the respondent's ability in a shorter amount of time than that required to administer the entire battery. In CAT, items are selected sequentially, conditional on a running estimate of a respondent's aptitude. Given an ability estimate, item selection is based on maximizing a given utility related to the information gain provided by an item.

One of the oldest and most popular CAT methodology selects items based on their specific contribution to the overall Fisher information $I = \sum_i J_i$. For a given step of the test $t+1$, conditional on a point estimate of the respondent's ability conditional on previously answered items, $\hat{\theta}_t$, one chooses the item i that has the maximum local information

$$J_i(\theta)|_{\theta=\hat{\theta}_t} = -\left.\frac{\partial^2}{\partial\theta^2}\sum_{k=1}^{K} w_{ik}\log p_i(k|\theta)\right|_{\theta=\hat{\theta}_t}, \tag{1}$$

based on model point estimates, where $p_i(k|\theta)$ is the probability mass function for item i, and w_{ik} is a weighting parameter associated with choosing k as a response to item i for a person with ability estimate $\hat{\theta}_t$. The rationale for this criteria derives from asymptotic normality of the sampling distribution of the estimate,

$$\hat{\theta}_t \xrightarrow{d} \text{normal}\left(\hat{\theta}_t, I_t^{-1}\right) \qquad \text{as} \quad t \to \infty,$$

where

$$I_t = \sum_{s=1}^{t} J_s(\hat{\theta}_t) \tag{2}$$

is the *Fisher information.* In choosing an item that maximizes the Fisher information, one is seeking the maximal reduction to the estimator variance.

Typically one resolves the weights w_{ik} self-consistently using the IRT model by setting them to $w_{ik} = p_i(k|\boldsymbol{\theta})$, though sometimes uniform weights $w_{ik} = 1/K$ are used. In this manuscript we will assume that one sets the weights in accordance with taking a conditional (on θ) expectation over the probability mass function for item i.

1.1 Newer CAT item selection criteria

The Fisher information method is simple and computationally expedient. Although widely used, it 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. Second, the Fisher information of Eq. 2 is an asymptotic approximation of the ability estimate precision; it is inaccurate when the number of observed items is small. Third, the greedy nature of most item selection methods, in conjunction with pre-set initialization, leads to highly stereotypical item trajectories and poor item exposure. Optimizing strictly for ability estimate variance ignores other concerns such as item exposure.

To address the first issue, item selection criteria that take ability uncertainty into account exist, taking the 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). To address the second issue, some item selection methods directly target the posterior variance of the ability estimate (van der Linden, 1998). To address the third 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). The third issue motivates our proposed selection method.

1.2 Item Response Theory (IRT)

IRT, a generative latent-variable modeling framework, is the dominant statistical paradigm for using assessments in order to evaluate the ability of respondents. In addition to its use in pretty much every high-stakes standardized assessment, applications of IRT are also widespread in health applications such as activities of daily living (Fieo et al., 2010), quality of life (Bilbao et al., 2014), depression (Carlo et al., 2021), and in personality tests (Goldberg, 1992; Bore et al., 2020; Saunders & Ngo, 2017; DeYoung et al., 2016; Funke, 2005; Spence et al., 2012).

In IRT, an ability parameter (canonically denoted θ) is associated with a person, placing that individual into a percentile rank relative to the population. In multidimensional IRT models, the ability of an individual $\boldsymbol{\theta}$ is a vector. Many types of multidimensional IRT models exist; however, our primary target is factorized models where each scale (dimension) can be isolated and treated separately.

In IRT models, a person's response to an item depends on the person's latent ability and the item's difficulty. Additionally, each item has a degree to which it is informed by the scale called the item discrimination. At calibration, the item-specific scale and discrimination parameters are fitted to a sample of responses collected from multiple respondents. The model's item parameters are then frozen and the model is used to score new respondents. In doing so, one is placing ability for new respondents within the context of the sample used for calibration.

1.3 Work Disability Functional Assessment Battery (WD-FAB)

IRT also serves as the theoretical basis for the WD-FAB (Meterko et al., 2015; Marfeo et al., 2016; 2019; Chang et al., 2022). Boston University and the National Institutes of Health, with the support of the Social Security Administration (SSA) (Marfeo et al., 2018; Meterko et al., 2015; Jette et al., 2019; Porcino et al., 2018), developed this multidimensional instrument for characterizing whole body and mental function. The

IRT model itself is fully factorized, with four mental scales and four physical scales, each scale using an independent uni-dimensional graded response model. Initial versions of it were developed using empirical Bayesian methods and later versions employed full Bayesian inference (Chang et al., 2022).

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.

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. Overall, these studies collected item responses from approximately $12,000$ subjects sampled from claimants for disability benefits as well as working-age adults who represent the general population of the United States. The underlying parameters for these models were rescaled so that a unit of one point corresponds to one standard deviation in estimated scores from a representative control sample of working age adults in the United States.

It is our objective to improve the CAT administration of the WD-FAB in terms of accuracy, efficiency, and item exposure. In the current assessment, where item selection is based on optimizing the Fisher information, a minimum of five and maximum of twelve items are administered per scale. These limits are designed in order to achieve at least a minimum degree of scoring convergence while being mindful of respondent burden. Subject to these limits, a scale is also considered converged if its posterior score estimate has standard deviation of about a third of a point.

The assessment is often administered to physically and mentally impaired individuals who are applying for disability benefits. These individuals commonly need assistance, so it is important to reduce administration times. However, the prior methodology has poor item exposure, and we would like to improve this aspect without compromising accuracy and efficiency.

Item exposure in the context of the WD-FAB is important for two main reasons. First, by increasing the diversity of the exposed items (and real-world item trajectories), we are able to reduce the possibility of gaming the instrument and by proxy the disability determination process. Second, having poor item exposure can bias the instrument in focusing on a small subset of physical or mental impairments while ignoring others. While the underlying IRT model theoretically is invariant to which exact items are provided to a respondent in terms of producing a score, in reality no IRT model perfectly describes the response data. The necessarily finite nature of the CAT means that finite-size effects can bias scores.

2 Methods

2.1 Notation

Suppose that one has modeled responses to a test bank using an item response model so that a person of ability $\boldsymbol{\theta}$ is expected to respond to item i according to the probability mass function $p_i(k|\boldsymbol{\theta})$. For a given individual, knowing all of his responses $\mathbf{x} = (x_1, x_2, \ldots, x_I)$, one may estimate the ability of the individual by computing the statistics of the posterior distribution,

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{I} p_i(x_i|\boldsymbol{\theta}) \tag{3}$$

where the maximum likelihood estimate corresponds to using a flat prior for $\pi(\boldsymbol{\theta})$.

The objective of a testing session is to use this model to ascertain the ability of a given individual relative to that of the calibration sample, approximating the statistics of Eq. 3 in as efficient a manner as possible. In this sense, Eq. 3 is considered the *true* estimate of a person's ability. Let the vector $\boldsymbol{\alpha}_t = (\alpha_1, \alpha_2, \ldots, \alpha_t) \in P(I, t)$ represent the particular permutation of items presented to a respondent by step t, and $\mathbf{x}_t$ be the responses to those particular items. In CAT applications it is common to estimate the ability at step t according to

the likelihood of the previously observed items so that

$$\tilde{\pi}_t(\boldsymbol{\theta}|\mathbf{x}_t) \propto \pi(\boldsymbol{\theta}) \prod_{s=1}^{t} p_{\alpha_s}(x_{\alpha_s}|\boldsymbol{\theta}). \tag{4}$$

Then, the choice of item α_{t+1} is made conditional on this running estimate. At a given step t, the choice of the next item α_{t+1} is analogous to choosing among $I-t$ choices for the next ability estimate $\tilde{\pi}_{t+1}(\boldsymbol{\theta}|\mathbf{x}_{t+1})$. By computing the KL divergence between this next estimate and the true estimate, we obtain the information theoretic discrepancy measure

$$\mathcal{D}\left(\pi(\boldsymbol{\theta}|\mathbf{x}), \tilde{\pi}_{t+1}(\boldsymbol{\theta}|\mathbf{x}_{t+1})\right) = \int \pi(\boldsymbol{\theta}|\mathbf{x}) \log \frac{\pi(\boldsymbol{\theta}|\mathbf{x})}{\tilde{\pi}_{t+1}(\boldsymbol{\theta}|\mathbf{x}_{t+1})} \mathrm{d}\boldsymbol{\theta}. \tag{5}$$

2.2 Computer adaptive testing

In computer adaptive testing (CAT), items are presented sequentially to a respondent. Our focus is on addressing the optimization of Eq. 5 at each step in this setting. Conditional on the new response, one may update ability estimates by application of Bayes rule,

$$\tilde{\pi}_{t+1}(\boldsymbol{\theta}|\mathbf{x}_{t+1}) = \frac{p_{\alpha_{t+1}}(x_{t+1}|\boldsymbol{\theta})\tilde{\pi}_t(\boldsymbol{\theta}|\mathbf{x}_t)}{\int p_{\alpha_{t+1}}(x_{t+1}|\boldsymbol{\phi})\tilde{\pi}_t(\boldsymbol{\phi}|\mathbf{x}_t)\mathrm{d}\boldsymbol{\phi}}. \tag{6}$$

We wish to adjudicate item choice based on Eq. 5, but a major difficulty remains: we do not know the future responses so $\pi(\boldsymbol{\theta}|\mathbf{x})$ is unknown. This issue is not unique to our methodology, and is usually resolved by taking an expectation using a given mass function (typically using the current ability estimate).

Each different set of responses $\{x_{\alpha_{t+1}}\}_{s=t+1}^{I}$ yields a different value for the discrepancy; in particular, it is inconvenient that each possible response to the next item yields a different $\pi(\boldsymbol{\theta}|\mathbf{x})$. Computing the expectation of Eq. 5 exactly requires specifying $(I-t)\times K$ 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. 5. In this estimate, we ignore the coupling between $\pi(\boldsymbol{\theta}|\mathbf{x})$ and the response to the next item, plugging in the expectation of $\pi(\boldsymbol{\theta}|\mathbf{x})$ into Eq. 5.

2.3 Variational Bayesian Expectation Maximization (VBEM)

In order to decouple the full bank posterior $\pi(\boldsymbol{\theta}|\mathbf{x})$ term, we wish to marginalize it against the unobserved items,

$$\pi_t(\boldsymbol{\theta}|\mathbf{x}_t) = \mathbb{E}_{\mathbf{z}_t}\pi(\boldsymbol{\theta}, \mathbf{z}_t|\mathbf{x}_t), \tag{7}$$

where $\mathbf{z}_t$ are the responses that have not yet been observed at step t. VBEM (Bernardo et al., 2003) allows us to approximate Eq. 7 through the following iterative procedure, at step m,

$$\log q_{\mathbf{z}_t,j}^{(m+1)}(k) = \mathrm{const}_j^{(m+1)} + \int \log p_j(k|\boldsymbol{\theta}) q_{\boldsymbol{\theta}}^{(m)}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{8}$$

$$\log q_{\boldsymbol{\theta}}^{(m+1)}(\boldsymbol{\theta}) = \mathrm{const}^{(m+1)} + \log \pi(\boldsymbol{\theta}) + \sum_{j\in\boldsymbol{\alpha}_t} \log p_j(x_j|\boldsymbol{\theta}) + \sum_{j\notin\boldsymbol{\alpha}_t}\sum_{k} q_{\mathbf{z}_t,j}^{(m+1)}(k) \log p_j(k|\boldsymbol{\theta}) \tag{9}$$

where each iteration is guaranteed to not increase $\mathcal{D}(\pi_t(\boldsymbol{\theta}|\mathbf{x}_t), q_{\boldsymbol{\theta}}^{(m)})$ by the principle of majorization-minimization (Wu, 1983; Wu & Lange, 2010; Lange, 2016; Lange & Zhou, 2022). Then, after some number of EM iterations M, we can compute the plug-in criterion

$$\Delta_t^{(i)} = \sum_k q_{\mathbf{z}_t,i}^{(M)}(k)\mathcal{D}\left(q_{\boldsymbol{\theta}}^{(M)}(\boldsymbol{\theta}), \tilde{\pi}(\boldsymbol{\theta}|\mathbf{x}_t, x_i = k)\right). \tag{10}$$

Technically, Eq. 7, rather than the commonly-used Eq. 4, is the best estimate of the ability at step t, an observation that we will save for the Discussion.

2.4 Stochastic item selection

At step t there remain $I-t$ un-selected items to present to the respondent. Eq. 10 provides a relative measure for judging each item, with the goal of closing the gap between the running ability estimate and the final ability estimate. Typically, one uses criteria such as Eq. 10 as part of a stepwise greedy algorithm where at each step one chooses the single item that optimizes the given criterion. However, with the aim of reducing stereotypical item paths and improving item exposure, we motivate a stochastic selection method.

Each item choice implies a different posterior ability distribution at the next step. The act of choosing between these potential distributions is effectively a form of model selection. Computing Eq. 10 is equivalent to estimating the information theoretic model discrepancy, relative to a true ability estimate and conditional on next item choice. This fact motivates the creation of a meta distribution in ability estimate space where each ability model (implied by item choice) has the relative weight of $\exp(-\Delta_t^{(i)})$ (Akaike, 1978; Bozdogan, 1987; Dormann et al., 2018; Wagenmakers & Farrell, 2004; Yao et al., 2018). Corresponding to this ensemble ability model we introduce an item sampling scheme. We draw the next item $i \notin \boldsymbol{\alpha}_t$, according to

$$\alpha_{t+1} \sim \text{Categorical}(\mathbf{w}_t) \qquad w_t^{(i)} = \frac{\exp\left(-\Delta_t^{(i)}\right)}{\sum_{j \notin \boldsymbol{\alpha}_t} \exp\left(-\Delta_t^{(j)}\right)}, \tag{11}$$

where the categorical distribution is defined over the $I-t$ items that have not yet been administered at time step t. In effect, the frequency statistics of items in Eq. 11 correspond to Bayesian model averaging (Hinne et al., 2020; Hoeting et al., 1999) of the corresponding per-item ability estimates.

2.5 Alternate formulations

We can rewrite the discrepancy (Eq. 5) to remove the explicit dependence on $\tilde{\pi}_{t+1}$,

$$\begin{aligned}\mathcal{D}\left(\pi(\boldsymbol{\theta}|\mathbf{x}), \tilde{\pi}_{t+1}(\boldsymbol{\theta}|\mathbf{x}_{t+1})\right) &= \int \pi(\boldsymbol{\theta}|\mathbf{x}) \log \frac{\tilde{p}_i^{(t)}(x_{t+1})\pi(\boldsymbol{\theta}|\mathbf{x})}{p_i(x_{t+1}|\boldsymbol{\theta})\tilde{\pi}_t(\boldsymbol{\theta}|\mathbf{x}_t)} \mathrm{d}\boldsymbol{\theta} \\ &= \int \pi(\boldsymbol{\theta}|\mathbf{x}) \log \frac{\tilde{p}_i^{(t)}(x_{t+1})}{p_i(x_{t+1}|\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} + \mathcal{D}(\pi(\boldsymbol{\theta}|\mathbf{x})||\tilde{\pi}_t(\boldsymbol{\theta}|\mathbf{x}_t))\end{aligned} \tag{12}$$

where

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

and note that while the second term in the last line of Eq. 12 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. 12

$$\begin{aligned}\mathcal{D}\left(\pi(\boldsymbol{\theta}|\mathbf{x}), \tilde{\pi}_{t+1}(\boldsymbol{\theta}|\mathbf{x}_{t+1})\right) &= \mathcal{D}(\pi(\boldsymbol{\theta}|\mathbf{x})||\tilde{\pi}_t(\boldsymbol{\theta}|\mathbf{x}_t)) + \int \pi(\boldsymbol{\theta}|\mathbf{x}) \log \frac{\tilde{p}_i^{(t)}(x_i)\pi(\boldsymbol{\theta}|\mathbf{x})}{\pi(\boldsymbol{\theta}|\mathbf{x})p_i(x_i|\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \\ &= \mathcal{D}(\pi(\boldsymbol{\theta}|\mathbf{x})||\tilde{\pi}_t(\boldsymbol{\theta}|\mathbf{x}_t)) + S[\pi(\boldsymbol{\theta}|\mathbf{x})] - \mathcal{D}(\pi(\boldsymbol{\theta}|\mathbf{x})||\tilde{\pi}_{I-1}(\boldsymbol{\theta}|\mathbf{x} \setminus \{x_i\})) + \log \frac{\tilde{p}_i^{(t)}(x_i)}{\tilde{p}_i^{\text{LOO}}(x_i)}\end{aligned} \tag{13}$$

where, $\tilde{\pi}_{I-1}(\boldsymbol{\theta}|\mathbf{x} \setminus \{x_i\})$, the ability estimate when leaving out x_i follows Bayes rule,

$$\frac{p_i(x_i|\boldsymbol{\theta})\tilde{\pi}_{I-1}(\boldsymbol{\theta}|\mathbf{x} \setminus \{x_i\})}{\tilde{p}_i^{\text{LOO}}(x_i)} = \pi(\boldsymbol{\theta}|\mathbf{x})$$

and $\tilde{p}_i^{\text{LOO}}(x_i) = \int p(x_i|\boldsymbol{\theta})\tilde{\pi}_{I-1}(\boldsymbol{\theta}|\mathbf{x} \setminus \{x_i\})\mathrm{d}\boldsymbol{\theta}$ is the corresponding LOO mass function for item i. In this representation, only the last two terms in Eq. 13 depend on the item choice. In minimizing the discrepancy, one is also selecting the item that if left out would yield the biggest discrepancy.

2.6 Relationship to prior methods

Eq. 5 is not in the class of variance minimizing criteria, whether it be the Fisher information (Eq. 2), global variants of the Fisher information

$$\text{Bayesian Fisher information} = \int \tilde{\pi}_t(\theta) J_i(\theta) \mathrm{d}\theta, \tag{14}$$

or any criteria that directly approximates the quantity

$$\text{Bayesian variance} = \text{Var}\left[\theta | \mathbf{x}_t, \alpha_{t+1} = i\right]. \tag{15}$$

The LOO version of the discrepancy (Eq. 13) relates to the "global information" method of Chang & Ying (1996),

$$\begin{aligned}
\text{Global information} &= \mathbb{E}_\theta \left[\sum_k p_i(k|\theta) \log \frac{p_i(k|\theta)}{p_i(k|\hat{\theta}_t)} \right] \\
&= \sum_k \int \tilde{\pi}_t(\theta|\mathbf{x}_t) p_i(k|\theta) \log \left[\frac{p_i(k|\theta)}{p_i(k|\hat{\theta}_t)} \frac{\tilde{\pi}_t(\theta|\mathbf{x}_t)}{\tilde{\pi}_t(\theta|\mathbf{x}_t)} \right] \mathrm{d}\theta \\
&= \sum_k \tilde{p}_i^{(t)}(k) \left[\mathcal{D}\left(\tilde{\pi}_{t+1}(\theta|\mathbf{x}_t, x_i = k) || \tilde{\pi}_t\right) - \log\ p_i(k|\hat{\theta}_t) + \log \tilde{p}_i^{(t)}(k) \right] \\
&= \mathbb{E}_{x_i}\left[\mathcal{D}\left(\tilde{\pi}_{t+1}(\theta|\mathbf{x}_t, x_i = k) || \tilde{\pi}_t(\theta|\mathbf{x}_t)\right)\right] + \mathcal{D}_{x_i}[\tilde{p}_i^{(t)} || p_i(k|\hat{\theta}_t)],
\end{aligned} \tag{16}$$

for $x_i \sim \tilde{p}_i^{(t)}$. Other information-theoretic methods that are based on comparing the statistical distance relative to the current ability estimate also exist (Sorrel et al., 2020; Wang & Chang, 2011; Weissman, 2007; Wang et al., 2020). The main difference between our method and these prior methods is that we evaluate the item choice against the *true* ability estimate rather than the current ability estimate – and thereby motivate a model-averaging stochastic selector.

2.7 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. 9).

3 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. 2), Bayesian Fisher information (Eq. 14), Global information (Eq. 16), Bayesian variance (Eq. 15), ability estimate discrepancy (Eq. 10), and our stochastic selection method (Eq. 11). 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.

3.1 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. 5) 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

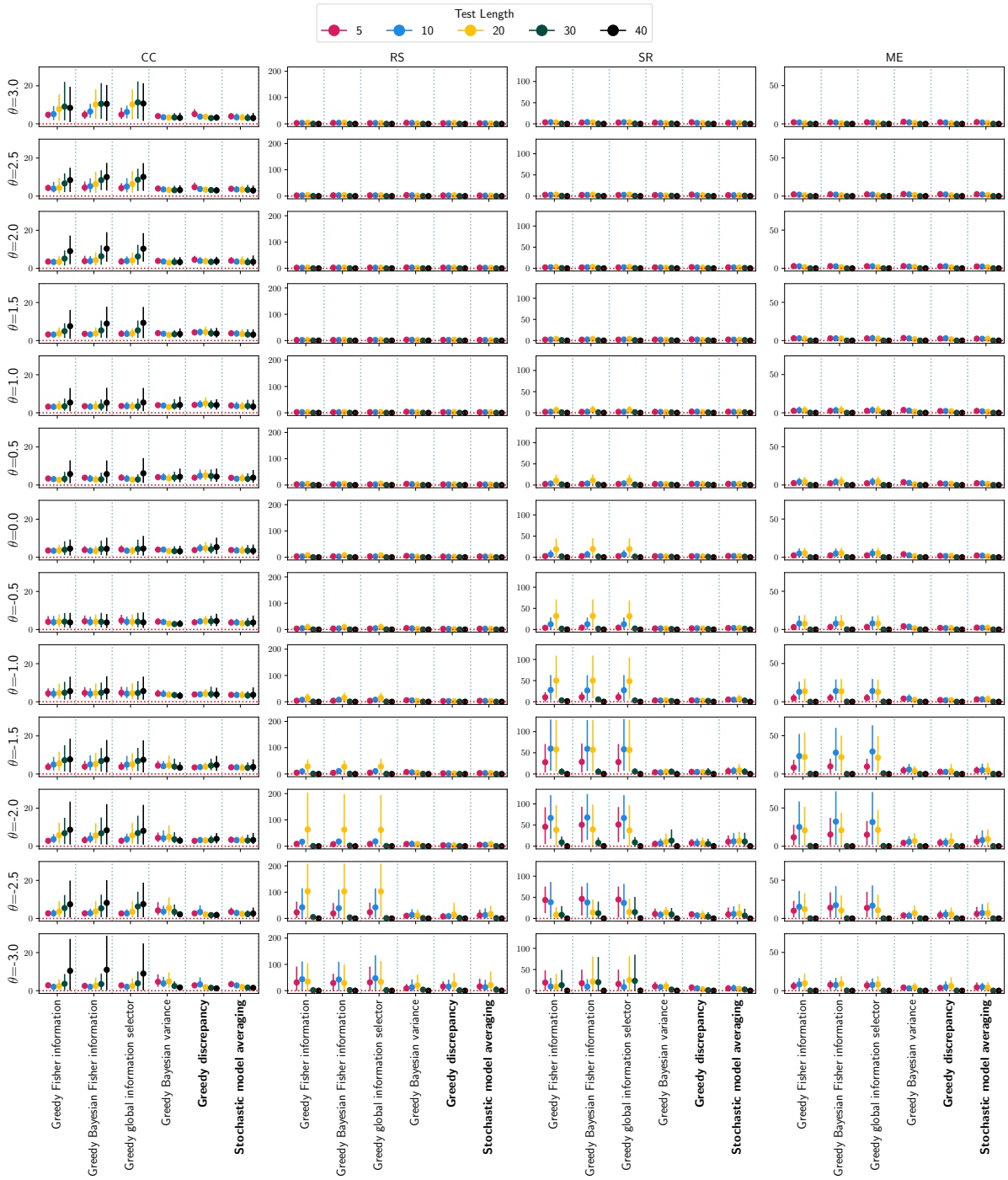

Figure 1: **Ability estimate discrepancy** $\mathcal{D}(\pi(\boldsymbol{\theta}|\mathbf{x})||\tilde{\pi}(\boldsymbol{\theta}|\mathbf{x}_t))$ (mean and middle 80% interval) conditional on score θ used to generate response sets, by scale, item selection method, and test length t, for mental function scales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

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. 10) reliably decrease the discrepancy as the test length increases. Failure to decrease this discrepancy suggests that a selection

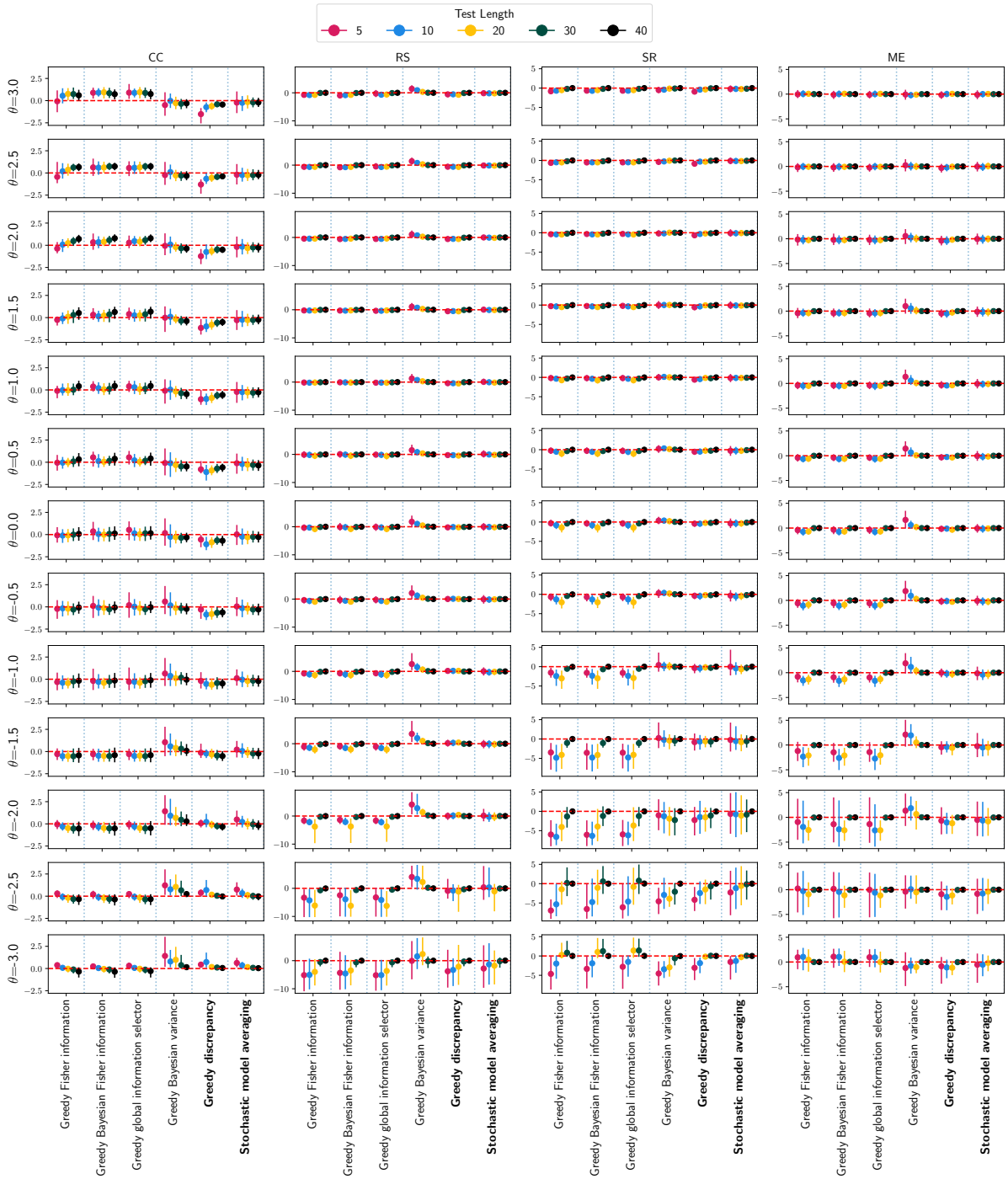

Figure 2: **Error in means** ($\int \theta\tilde{\pi}(\theta|\mathbf{x}_t)\mathrm{d}\theta - \int \theta\pi(\theta|\mathbf{x})\mathrm{d}\theta$) (mean and middle 80% interval) conditional on true score θ by scale, item selection method, and test length t, for mental function scales of the WD-FAB.

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

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.

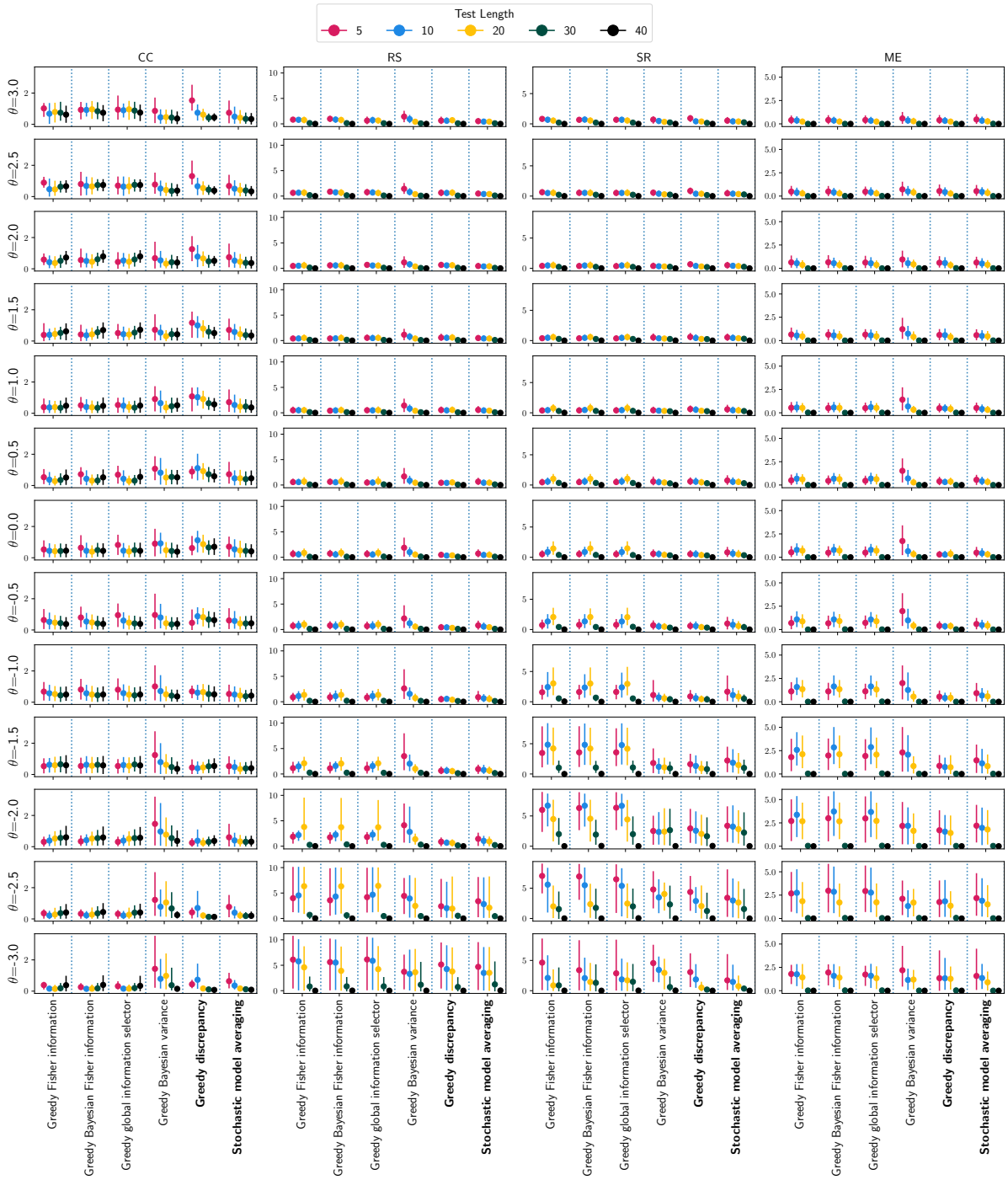

Figure 3: **Absolute error in means** ($|\int \theta\tilde{\pi}(\theta|\mathbf{x}_t)\mathrm{d}\theta - \int \theta\pi(\theta|\mathbf{x})\mathrm{d}\theta|$) (mean and middle 80% interval) conditional on true score θ by scale, item selection method, and test length t, for mental function scales of the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

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

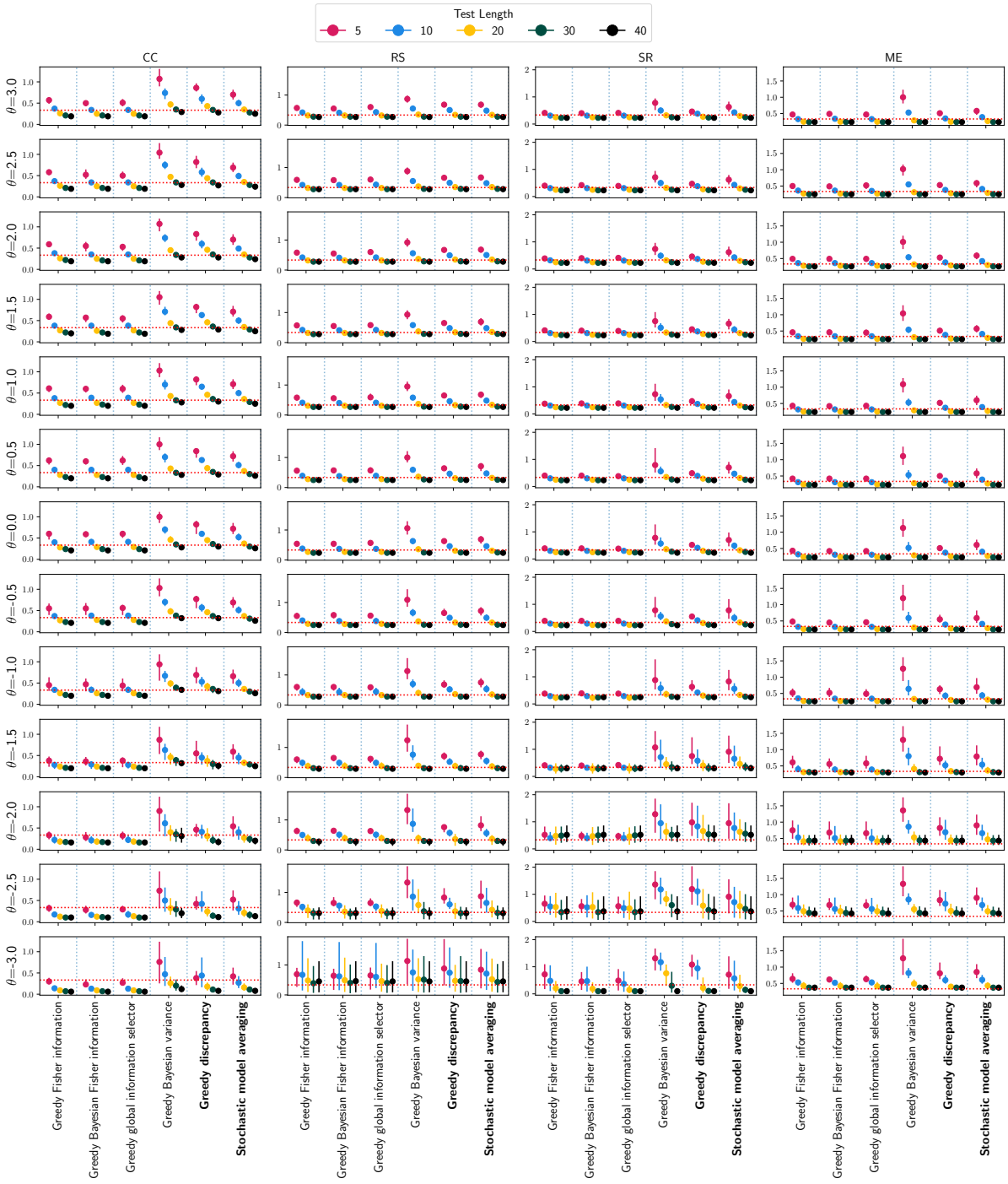

Figure 4: **Standard deviation of ability estimates** ($\sqrt{\mathbf{Var}_t(\theta)}$ (mean and middle 80% percentile) conditional on true score θ by scale and item selection method, for mental function scales of the WD-FAB. Used as stopping criteria for CAT. Lower is better.

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 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 methods are under-estimating the error of their ability estimates. In doing so, they are terminating quicker than they should and settling on sub-optimal ability estimates.

3.2 Item exposure

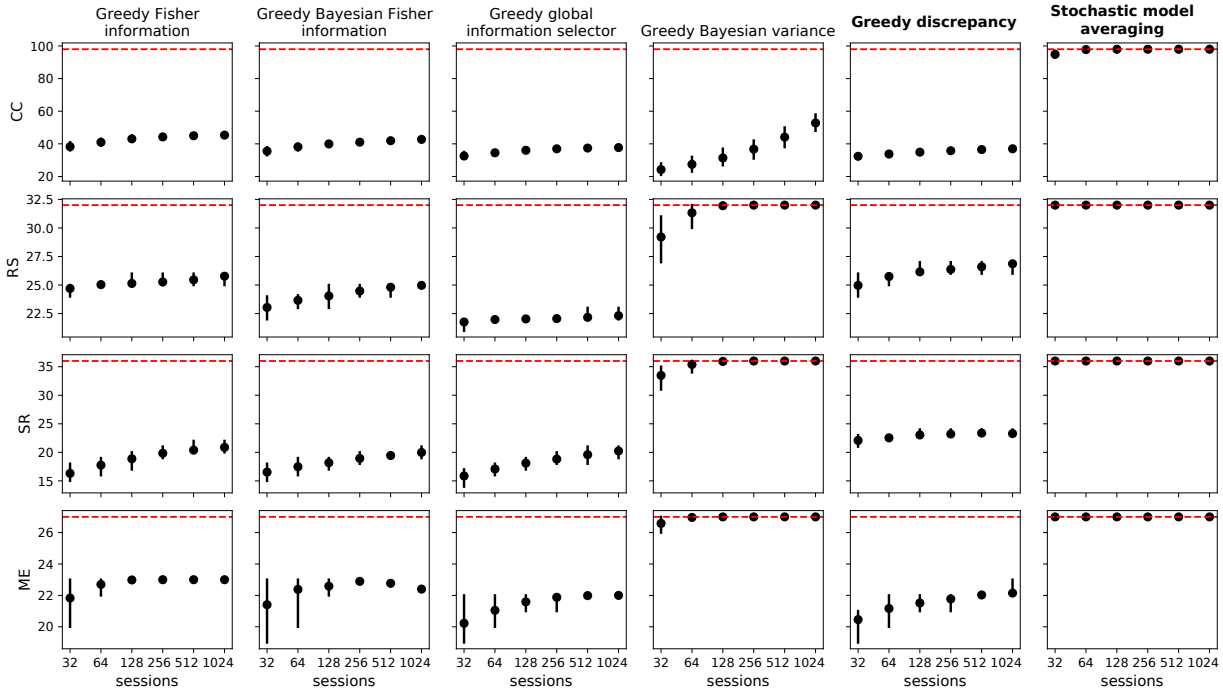

Figure 5: **Item exposure statistics** (mean and middle 80% interval), for each of the given item selection methods across a given number of CAT sessions, for mental function scales of the WD-FAB. The dashed line represents the maximum possible exposure per scale. Higher is better.

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. 11. The stochastic selector successfully exposed all items for all scales in all the scenarios tested.

4 Discussion

In this manuscript we have introduced a new item selection criterion for CAT based on Bayesian information theory and motivated its use in defining a stochastic selector that samples from the hypothetical ensemble of ability estimates conditional on the next item. 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.

4.1 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 $\tilde{\pi}_t(\boldsymbol{\theta}|\mathbf{x}_t)$, making a distinction between this quantity and $\pi(\boldsymbol{\theta}|\mathbf{x}_t)$, the marginal posterior ability at step t. The latter estimate differs from the former in that it also accounts for the fact that the $I-t$ 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 $\tilde{\pi}_t$ 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.

4.2 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. 10.

4.3 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. 10, 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. 5 in the form of the the mean field plugin estimator in Eq. 10 trades accuracy for computational efficiency. One could more-accurately compute this expectation by developing a version of Eq. 10 that preserves the coupling between $\pi(\boldsymbol{\theta}|\mathbf{x})$ 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 $\boldsymbol{\theta}$, 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.

Amaia Bilbao, Carlota Las Hayas, Carlos G. Forero, Angel Padierna, Josune Martin, and José M. Quintana. Cross-Validation Study Using Item Response Theory: The Health-Related Quality of Life for Eating Disorders Questionnaire–Short Version. *Assessment*, 21(4):477–493, August 2014. ISSN 1073-1911. doi: 10.1177/1073191113509004.

Miles Bore, Kristin R. Laurens, Megan J. Hobbs, Melissa J. Green, Stacy Tzoumakis, Felicity Harris, and Vaughan J. Carr. 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, February 2020. ISSN 0885-579X. doi: 10.1521/pedi_2018_32_380.

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.

Andrew D. Carlo, Brian S. Barnett, and David Cella. Computerized Adaptive Testing (CAT) and the Future of Measurement-Based Mental Health Care. *Administration and Policy in Mental Health*, 48(5):729–731, 2021. ISSN 0894-587X. doi: 10.1007/s10488-021-01123-9.

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, 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.

Colin. G. DeYoung, Bridget E. Carey, Robert F. Krueger, and Scott R. Ross. 10 Aspects of the Big Five in the Personality Inventory for DSM-5. *Personality disorders*, 7(2):113–123, April 2016. ISSN 1949-2715. doi: 10.1037/per0000170.

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.

Robert Fieo, Roger Watson, Ian J. Deary, and John M. Starr. 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, 2010. ISSN 0304-324X, 1423-0003. doi: 10.1159/000271603.

Friedrich Funke. The Dimensionality of Right-Wing Authoritarianism: Lessons from the Dilemma between Theory and Measurement. *Political Psychology*, 26(2):195–218, 2005. ISSN 0162-895X.

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.

Lewis R. Goldberg. The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1):26–42, 1992. ISSN 1939-134X(Electronic),1040-3590(Print). doi: 10.1037/1040-3590.4.1.26.

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.

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.

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.

Kenneth Lange. Chapter 7: Convergence and Acceleration. In *MM Optimization Algorithms*, Other Titles in Applied Mathematics, pp. 173–194. Society for Industrial and Applied Mathematics, July 2016. ISBN 978-1-61197-439-3. doi: 10.1137/1.9781611974409.ch7.

Kenneth Lange and Hua Zhou. A Legacy of EM Algorithms. *International statistical review = Revue internationale de statistique*, 90(Suppl 1):S52–S66, December 2022. ISSN 0306-7734. doi: 10.1111/insr.12526.

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.

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.

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.

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.

Benjamin A. Saunders and Josephine Ngo. The Right-Wing Authoritarianism Scale. In Virgil Zeigler-Hill and Todd K. Shackelford (eds.), *Encyclopedia of Personality and Individual Differences*, pp. 1–4. Springer International Publishing, Cham, 2017. ISBN 978-3-319-28099-8. doi: 10.1007/978-3-319-28099-8_1262-1.

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.

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.

Ruth Spence, Matthew Owens, and Ian Goodyer. Item response theory and validity of the NEO-FFI in adolescents. *Personality and Individual Differences*, 53(6):801–807, October 2012. ISSN 0191-8869. doi: 10.1016/j.paid.2012.06.002.

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.

C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, March 1983. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176346060.

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.