Graduate
Management
Admission
Council®

A Burdened CAT: Incorporating Response Burden With Maximum Fisher's Information for Item Selection

Richard J. Swartz
The University of Texas M.D. Anderson Cancer Center

Seung W. Choi
Northwestern University Feinberg School of Medicine

Presented at the Item Selection Paper Session, June 2, 2009

2009 GMAC® Conference on Computerized Adaptive Testing

Abstract

Widely used in various educational and vocational assessment applications, computerized adaptive testing (CAT) has recently begun to be used to measure patient-reported outcomes Although successful in reducing respondent burden, most current CAT algorithms do not formally consider it as part of the item selection process. This study used a loss function approach motivated by decision theory to develop an item selection method that incorporates respondent burden into the item selection process based on maximum Fisher information item selection. Several different loss functions placing varying degrees of importance on respondent burden were compared, using an item bank of 62 polytomous items measuring depressive symptoms. One dataset consisted of the real responses from the 730 subjects who responded to all the items. A second dataset consisted of simulated responses to all the items based on a grid of latent trait scores with replicates at each grid point. The algorithm enables a CAT administrator to more efficiently control the respondent burden without severely affecting the measurement precision than when using MFI alone. In particular, the loss function incorporating respondent burden protected respondents from receiving longer tests when their estimated trait score fell in a region where there were few informative items.

Acknowledgment

This research was supported in part by an NCI Career Development Award (K07-CA-113641; P.I. Richard Swartz) for the first author and in part by an NIH PROMIS Network grant (U-01 AR 052177-04; P.I. David Cella) for the second author. Collection of the data used in this study was funded in part through a core facility of a Cancer Center Support Grant (NCI CA16672; P.I. John Mendelsohn). Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Authors

Citation

Swartz, R. J., & Choi, S. W. (2009). A burdened CAT: Incorporating response burden with maximum Fisher's information for item selection. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

**Richard J. Swartz, Rice University, P.O. Box 2932,
Houston TX, 77252, U.S.A. Email: rswartz@rice.edu**

A Burdened CAT: Incorporating Response Burden with Maximum Fisher's Information for Item Selection

Widely used in various assessment applications, computerized adaptive testing (CAT) has begun to infiltrate the patient-reported outcomes (PRO) arena (Bjorner, Chang, Thissen, & Reeve, 2007; Reeve, 2006). PROs such as depression and fatigue are represented as latent traits (similar to mathematical achievement) so CATs in conjunction with IRT are natural considerations for PRO measurement (Bjorner, et al., 2007; Cella & Chang, 2000; McHorney, 2003; Reeve, 2006). A typical CAT design uses a mathematical algorithm to sequentially select items that are in some sense "best" from a pool of pertinent items (called an item bank) until an estimate of the latent trait is achieved with a certain precision. As a result, CAT assessments are typically shorter and at least as precise as traditional static instruments (Jenkinson, Fitzpatrick, Garratt, Peto, & Stewart-Brown, 2001; Meijer & Nering, 1999; Revicki & Cella, 1997). This is particularly enticing to researchers interested in PRO measurement, where respondent burden is a concern (Bjorner, et al., 2007; Science Advisory Committee of the Medical Outcomes Trust, 2002).

Respondent burden is defined as the demands and expectations placed on the people responding to the instrument or questionnaire (Science Advisory Committee of the Medical Outcomes Trust, 2002). It includes the time needed to complete the instrument or questionnaire, special requests or requirements placed on the respondent (i.e., remembering medical record information), physical or emotional stress placed on the individual, and overall suitability of the instrument for the respondents. In longitudinal assessments, the frequency with which subjects must respond to the instrument also contributes to respondent burden. This burden is a concern when measuring PROs because these measures are often administered repeatedly during a short time interval (Bjorner, et al., 2007**.**

With a few exceptions, many current item selection methods to date perform similarly for polytomous items (Choi & Swartz, 2009; Penfield, 2006). However, most item selection methods developed to date do not explicitly consider the cost of administering items (Swartz, Choi, & Herrick, 2009; van der Linden & Pashley, 2000). The idea of considering respondent burden has been encountered in mastery testing (Lewis & Sheehan, 1990; Vos, 1999, 2000), but the structure of the problem is different. In mastery testing, items are selected to minimize the cost of administering items and the cost of a wrong decision, i.e., master vs. non-master (Kingsbury & Weiss, 1983; Lewis & Sheehan, 1990). However, in PRO assessment research, the focus centers on estimation of the individual score on the latent trait rather than on classification. CAT algorithms frequently used in PRO assessments select items based on their information (typically Fisher's information based on the provisional estimate). To control respondent burden, these algorithms typically rely on ad hoc rules imposed to control respondent burden, such as imposing a maximum number of items to be administered (Bjorner, et al., 2007; Fliege, et al., 2005; Hart, Wang, Stratford, & Mioduski, 2008a, 2008b; Walter, et al., 2007).

The trade-off between the precision of a PRO measurement and the burden associated with that measurement is unequivocal: Other characteristics being equal, administering more items produces more precise estimates, but administering more items also increases the respondent burden. Motivated by a Bayesian decision theoretic approach, this study proposes an adaptation to the maximum Fisher's information (MFI) selection criterion to incorporate respondent burden. First, we derive a Bayesian decision theoretic approach to motivate the process taken to adapt the MFI. Next we introduce a loss function to generalize the MFI to incorporate the respondent burden. Finally we compare the performance of the MFI selection criterion that includes burden, called MFI-*b*, to the standard MFI selection method. We hypothesized that the MFI-*b* will control burden by not administering items that in some sense are not "worth" asking.

Incorporating Respondent Burden

To formally incorporate respondent burden, we start with establishing some notation. Adopting the notation from van der Linden and Pashley (van der Linden & Pashley, 2000), let θ be the latent trait of interest, and u_{i_k} represent the response to item i_k administered at the k^{th} stage in the sequence. Then, let m_k be the number of response categories for item i_k (m_k =4 for all k in this study). Let A_k denote the set of items administered up to and including stage k, and R_k be the set of items that are available for selection after administering k items (in other words, R_k is the complement of A_k). $\mathbf{u}_{A_k}$ denotes a response vector associated with items administered in a sequence having k stages: $(u_{i_1}, \ldots, u_{i_k})'$. Let $g(\theta \mid \mathbf{u}_{A_k})$ denote the posterior distribution after administering k items (the probability distribution of θ, given the previous k item responses and the prior). Let $p_j(u \mid \mathbf{u}_{A_k})$ denote the posterior predictive distribution (the probability of giving response u to item j given the previous response history). For more details about these distributions, see van der Linden and Pashley (van der Linden & Pashley, 2000), Carlin and Lewis (Carlin & Louis, 2000), or Gelman, Carlin, Stern, and Rubin (Gelman, Carlin, Stern, & Rubin, 2004).

Bayesian decision theory approach to item selection. Formally, a CAT item selection and stopping process can be viewed as a Bayesian decision problem. A Bayesian decision theoretic approach requires specifying (1) decision rules that dictate the action taken once data are observed, (2) a loss function to assign or model the consequences associated with each decision, and (3) a probability model to describe the uncertainty in the decision problem (Berger, 1985; Bernardo & Smith, 1994; DeGroot, 1969; Ferguson, 1967). Under a Bayesian decision theoretic framework, the optimal decision is the decision that minimizes the expectation of the loss function with respect to the posterior distribution. This is called minimizing the posterior expected loss (Berger, 1985). A fully Bayesian adaptive sequential item selection (BASIS) method has been proposed (Swartz & Choi, 2008; Swartz, et al., 2009).

For the BASIS method, a general loss function is proposed:

$$l(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}) = G\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right) + \alpha(\theta) C\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right). \tag{1}$$

The component $G\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right)$ models a loss associated with using the observed responses $(\mathbf{u}_{A_{k-1}}, u_{i_k})$ at the current stage k to estimate a value of the true but unknown latent score, θ. The component $C\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right)$ models a loss associated with the respondent burden. The component $\alpha(\theta)$ serves two purposes. First, it transforms the loss associated with the respondent burden onto the same scale as the loss arising from estimating θ. Second, it characterizes the importance of respondent burden relative to the estimation of θ. More details concerning this loss function and the BASIS method can be found in Swartz, Choi, and Herrick (2009).

Under certain conditions, the BASIS method reduces to the minimum expected posterior variance (MEPV) item selection criterion, making the MEPV a special case of the BASIS approach. We next briefly discuss this relationship to motivate our later adaptation to the maximum Fisher's information (MFI) selection criterion. More details are given in Appendix A.

The MEPV has been defined for dichotomous items (Owen, 1975; Thissen & Mislevy, 2000; van der Linden & Glas, 2000) and extended to polytomous items as follows (Choi & Swartz, 2009):

$$i_k^* \equiv \arg\min_{j \in R_{k-1}} \left\{ \sum_{u=1}^{m_j} p_j(u \mid \mathbf{u}_{A_{K-1}^*}) \operatorname{Var}(\theta \mid \mathbf{u}_{A_{k-1}^*}, U_j = u) \right\}, \tag{2}$$

where $\operatorname{Var}(\theta \mid \mathbf{u}_{A_{k-1}^*}, U_j = u)$ is the posterior variance for item *j* with predicted response category *u*, and the * on A_{k-1}^* indicates that the set of previously administered items was selected according to the current selection criterion (i.e., the MEPV). The MEPV is a special case of the BASIS method when (1) we restrict the decision to first selecting an item at the current stage, and then deciding to either stop the test, or continue the test and administer an additional item;(2) we place no importance on burden $\left[\alpha(\theta) \equiv 0\right]$; (3) we use squared error loss (that is, the square of the difference between the true value and the estimate) to model the loss associated with estimating θ using the data; and (4) we estimate θ with the posterior mean $E(\theta \mid \mathbf{u}_{A_{k-1}}, u_{i_k})$, or the expected a posteriori (EAP) estimator, which is a commonly used estimator for the θ in CAT instruments. In other words:

$$G\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right) = \left(\theta - E(\theta \mid \mathbf{u}_{A_{k-1}}, u_{i_k})\right)^2. \tag{3}$$

A more important benefit: if we define the loss associated with respondent burden to be the number of items administered at each stage, i.e., $C\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right) = k$, and assume that the trade-off between the respondent burden and precision of the estimate is constant $\left[\alpha(\theta) = \alpha\right]$, Appendix A shows that the BASIS method extends the MEPV to include respondent burden. Specifically a Bayesian decision theoretic approach identifies a stopping criterion that complements the standard MEPV selection criterion:

$$\begin{aligned}\text{Stop } &\Leftrightarrow \operatorname{var}(\theta \mid \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}) - E_{U_{i^*_{k+1}} \mid \mathrm{u}_{A^*_k}}\left[\operatorname{var}(\theta \mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}})\right] \le \alpha, \\ \text{Continue } &\Leftrightarrow \operatorname{var}(\theta \mid \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}) - E_{U_{i^*_{k+1}} \mid \mathrm{u}_{A^*_k}}\left[\operatorname{var}(\theta \mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}})\right] > \alpha.\end{aligned} \tag{4}$$

where i^*_{k+1} is the item selected by the MEPV method to be administered at the next stage. Thus, incorporating the burden using the loss function results in terminating the CAT instrument when the expected decrease in variance that occurs by administering the next item is less than α. That is, the CAT instrument ceases to administer items when the expected decrease in measurement error resulting from the administration of any additional item remaining in the bank is below a predetermined criterion. We will call this method MEPV-*b* to indicate the item selection procedure that incorporates burden.

Generalizing maximum Fisher information (MFI). In this section, we demonstrate how this loss function approach can facilitate generalizing the MFI selection criterion to include respondent burden. Let $I_{\mathbf{u}_{A_k}}(\theta)$ represent Fisher's information function for the items administered up to stage *k*, and $\hat{\theta}_{\mathbf{u}_{A_k}}$ be some estimate of θ based on the previous *k* item responses. MFI is then defined as:

$$i^*_{k-1} \equiv \arg\max_{j \in R_{k-1}} I_{\mathbf{u}_{A^*_{k-1}}, u_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) = \arg\max_{j \in R_{k-1}} I_{u_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}), \tag{5}$$

where the second expression results from the additive property of the information function (van der Linden & Pashley, 2000). The MFI is an *ad hoc* procedure that is not based in any formal decision theory. However, there is still an underlying loss function and a risk to be minimized. Appendix B first demonstrates how to frame the MFI in terms of a loss function and defines the risk to be minimized. Then, a method to incorporate respondent burden into the MFI is derived. The approach is similar to the approach used for the MEPV. Specifically, if we again consider the cost of respondent burden to be proportional to the number of items administered, and we assume that the trade-off between respondent burden and the precision of the estimate is a constant value, call it α_F, then we can use the following loss function to generalize the MFI to incorporate respondent burden:

$$l\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right) = -J_{\mathbf{u}_{A_{k-1}}, U_j}(\hat{\theta}_{\mathbf{u}_{A_{k-1}}}) + \alpha_F k, \tag{6}$$

where $-J_{\mathbf{u}_{A_{k-1}}, U_j}(\hat{\theta}_{\mathbf{u}_{A_{k-1}}})$ is the observed Fisher information. Also note that there is no asterisk (*) on the set A_{k-1}, indicating that the set of previously administered items is not required to be optimally selected, although in most cases it will be.

Using this loss function one can minimize the risk, which is the expectation of the loss function. Selecting the item at the current stage that minimizes the risk yields the following selection rule:

$$i^*_k \equiv \arg\max_{j \in R_{k-1}} I_{u_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}). \tag{7}$$

Once the item is selected, stopping is dictated by the following criterion:

$$\begin{aligned}\text{Stop } &\Leftrightarrow \alpha_F \geq \max_{z \in R_{k+1}} I_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) \\ \text{Continue} &\Leftrightarrow \alpha_F < \max_{z \in R_{k+1}} I_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}}).\end{aligned} \tag{8}$$

Note that Equation 7 is identical to the definition of the MFI (Equation 5). In Equation 8 the subscript *F* on α_F differentiates the value from α in the MEPV-*b*. The α_F value serves the same purpose as the αvalue in the MEPV-*b*: When there are no more items in the bank that contribute an information gain greater than α_F , the CAT stops. Because we consider each stage independently, incorporating respondent burden results only in changing the stopping rule. Also, note that the stopping rule amounts to imposing a threshold on the "worth" of administering an item, and the derivation gives mathematical justification for the use of such an approach (See Appendix B).

Selecting α.. Reviewing the loss function given in Equation 6, the value of α_F represents the relative importance (or relevance) of respondent burden. Higher values indicate that respondent burden has more importance, while lower values indicate respondent burden has less importance relative to the precision. The stopping rule implies a practical interpretation for α_F and therefore helps guide the choice of values: α_F can be thought of as the minimum information gain that is worth subjecting an examinee to an additional item. For example $\alpha = 0.2$ means it is only worth asking the next item if the item selected offers at least 0.2 information units for the current estimate of θ.

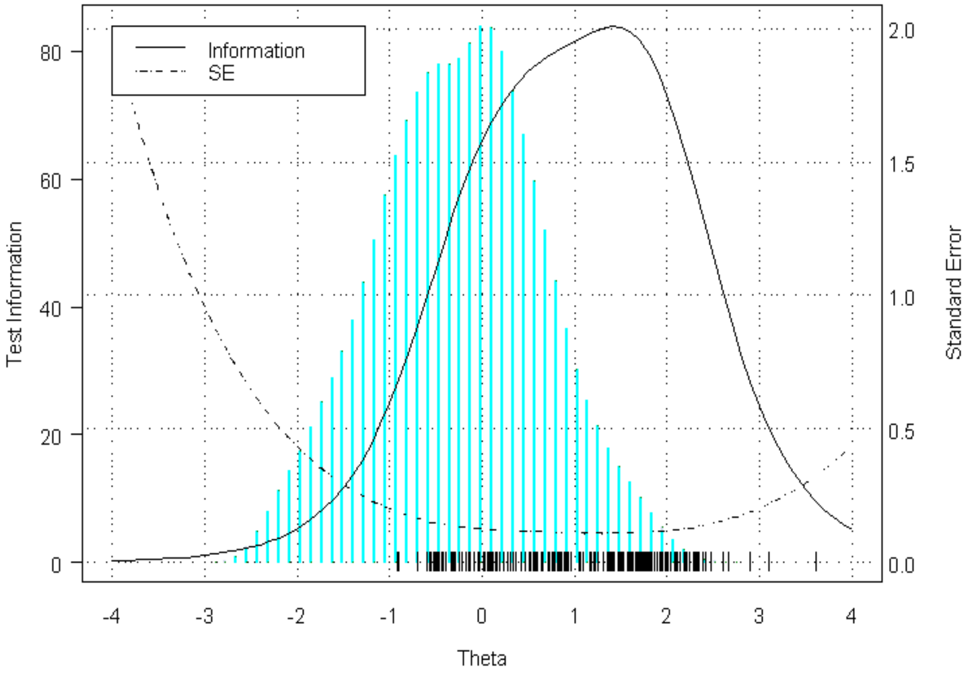
We can use this interpretation for α_F to reduce the set of potential values. Since there are many more applications involving dichotomous items, a rule of thumb has been developed to identify poor items: Items with a discrimination value below 0.8 are typically considered undesirable except in rare special cases. It can also be shown that the maximum information value of any item is equal to one-fourth the value of the square of the discrimination parameter. Therefore, with dichotomous items it seems reasonable to argue that it is not worth administering an item if the information it provides is less than 0.2.

Although it is not clear how this rule of thumb might map to polytomous items, it still indicates that α values near 0.2 might be reasonable. Also Dodd, Koch, and De Ayala (1989) used 0.5 as a minimum threshold for the information contributed by an item, but had little justification for selecting that value. Therefore, we explored the following values for this study: $\alpha_F \in \{0, 0.2, 0.4, 0.6, 0.8, 1, 1.25, 1.5\}$. Simulations will indicate the behavior of each α_F and how this models the trade-off between information and burden for polytomous items.

Study Design

The item bank consisted of 62 four-category response depression items inquiring about depressive symptom experiences. These items are described extensively by Choi and Swartz (2009). We had response data from 730 respondents. The graded response model (Samejima, 1969) was fit to the items using MULTILOG 7.03 (Thissen, Chen, & Bock, 2003). The total information function for the item bank had a mode higher than the standardized mean of 0.0 (see Figure 1).

Figure 1. Total Bank Information and Standard Error Functions and Distribution of Respondent's Estimated θ (N=730)



We performed two studies. One involved post-hoc simulations using the real responses from the 730 respondents. A second study involved simulating responses from known θs. For the simulated responses, we generated a grid of θs evenly distributed from – 3 to 3 at .5 increments. Using the item parameter estimates, response patterns were generated for each θ. This was then replicated 500 times resulting in a total of 6,500 simulees (simulated respondents).

In both studies, we compared the MFI with the MFI-*b* according to the number of items administered, and the standard error of measurement (SE). The CAT using the standard MFI selection criterion continued administering items until a precision less than or equal to .32 (roughly equivalent to a classical reliability of .9) was achieved. As will be discussed further below, since there were many items in the bank that had information values well above α_F, we augmented the MFI-*b* with the standard MFI stopping rule to develop a composite stopping rule. The composite rule stopped administering items when SE ≤ 0.32 was achieved, regardless of the information units any additional items offered (i.e., imposing the SE ≤ 0.32 stopping rule in addition to Equation 8). Also all simulated CATs, regardless of their selection criterion, were limited to a maximum length of 20 items.

Results

Table 1 summarizes the results for the simulated data, while Table 2 summarizes results for the real data. Reliability, average bias, root mean squared error (RMSE), correlation between the θ estimates from the CAT administration and the θ estimates from the full item bank (r), average number of items administered, and average of the squared standard error of the estimates (Mean SE^2) are reported. For the SE < 0.32 stopping rule in these tables, the root mean squared deviation (RMSD) between the θ estimates from the full item bank and the simulated CAT administration is reported instead of RMSE, since the true θ is unknown. With the exception of the standard MFI selection criterion (equivalent to α_F), as α increased, the average number of items administered decreased. Also as the value for α_F increased, the reliability and the correlations decreased, while the bias, mean SE, and the RMSE increased.

Table 1. Results for Simulated Data Using MFI-*b* Without Composite Stopping Rule and With < 0.32 Composite Stopping Rule

Stopping Rule and Selection Criterion	Reliability	Bias	RMSE/ RMSD	Corre- lation	Ave. No. of Items	Mean SE^2
No composite stopping rule						
MFI ($\alpha_F = 0$)	0.9010	0.0277	0.3632	0.9847	11.0	0.0990
MFI-*b*, $\alpha_F = 0.2$	0.9292	0.0444	0.2981	0.9902	19.6	0.0708
MFI-*b*, $\alpha_F = 0.4$	0.9154	0.0717	0.3518	0.9875	17.3	0.0846
MFI-*b*, $\alpha_F = 0.6$	0.8992	0.1018	0.4237	0.9831	15.3	0.1008
MFI-*b*, $\alpha_F = 0.8$	0.8843	0.1085	0.4691	0.9805	13.5	0.1157
MFI-*b*, $\alpha_F = 1.0$	0.8548	0.1558	0.5840	0.9710	11.5	0.1452
MFI-*b*, $\alpha_F = 1.25$	0.8440	0.1382	0.6070	0.9697	10.1	0.1560
MFI-*b*, $\alpha_F = 1.5$	0.8272	0.1162	0.6418	0.9670	7.1	0.1728
SE < 0.32 composite stopping rule						
MFI ($\alpha_F = 0$)	0.9010	0.0277	0.3632	0.9847	11.0	0.0990
MFI-*b*, $\alpha_F = 0.2$	0.8990	0.0305	0.3656	0.9847	10.7	0.1010
MFI-*b*, $\alpha_F = 0.4$	0.8853	0.0578	0.4102	0.9824	8.5	0.1147
MFI-*b*, $\alpha_F = 0.6$	0.8701	0.0868	0.4704	0.9785	6.8	0.1299
MFI-*b*, $\alpha_F = 0.8$	0.8572	0.0935	0.5081	0.9765	6.8	0.1428
MFI-*b*, $\alpha_F = 1.0$	0.8300	0.1413	0.6124	0.9676	4.6	0.1700
MFI-*b*, $\alpha_F = 1.25$	0.8221	0.1262	0.6304	0.9667	4.1	0.1779
MFI-*b*, $\alpha_F = 1.5$	0.8112	0.1063	0.6579	0.9652	3.6	0.1888

Table 1 shows that when the MFI-*b* is used alone (i.e., no composite stopping rule), the standard MFI procedure had the smallest bias, but fell between the conditions where α_F = 0.4 and α_F = 0.6 for reliability, RMSE and r, while it ranked between α_F = 1.0 and α_F = 1.25 in term of number of items used. When a composite rule was used, the standard MFI procedure fell where expected (α_F = 0) in terms of the summary measures (Tables 1 and 2).

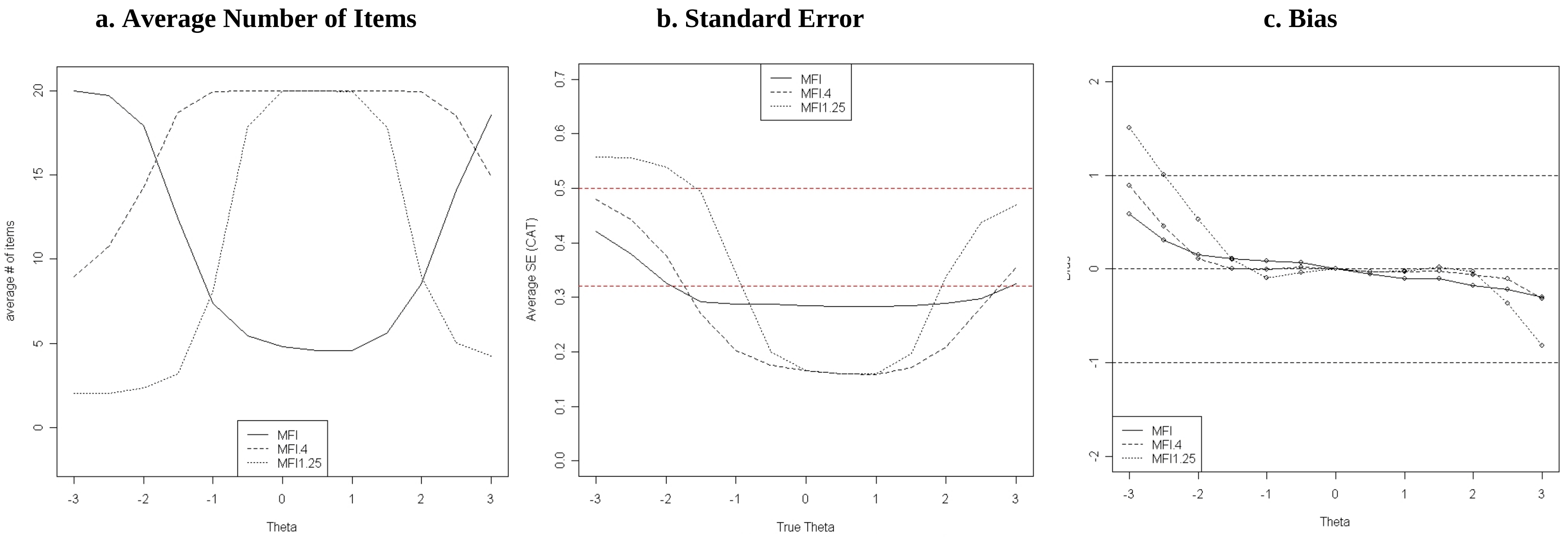
Table 2. Results for Real Data Using MFI-*b* Without Composite Stopping Rule and With < 0.32 Composite Stopping Rule

Stopping Rule and Selection Criterion	Reliability	Bias	RMSE/ RMSD	Corre-lation	Ave. No. of Items	Mean SE^2
No composite stopping rule						
MFI (α_F = 0)	0.9079	0.0466	0.3413	0.9301	9.2	0.0921
MFI-*b*, α_F = 0.2	0.9437	0.0160	0.2386	0.9649	19.8	0.0563
MFI-*b*, α_F = 0.4	0.9348	0.0285	0.2668	0.9558	18.5	0.0652
MFI-*b*, α_F = 0.6	0.9161	0.0410	0.3096	0.9404	16.5	0.0839
MFI-*b*, α_F = 0.8	0.9035	0.0567	0.3347	0.9310	14.8	0.0965
MFI-*b*, α_F = 1.0	0.8591	0.0918	0.4169	0.8937	12.3	0.1409
MFI-*b*, α_F = 1.25	0.8533	0.0896	0.4307	0.8848	11	0.1467
MFI-*b*, α_F = 1.5	0.8384	0.1051	0.4448	0.8785	6.8	0.1616
SE < 0.32 composite stopping rule						
MFI (α_F = 0)	0.9079	0.0466	0.3413	0.9301	9.2	0.0921
MFI-*b*, α_F = 0.2	0.9070	0.0501	0.3428	0.9292	9.0	0.0927
MFI-*b*, α_F = 0.4	0.8983	0.0622	0.3627	0.9196	7.7	0.1017
MFI-*b*, α_F = 0.6	0.8810	0.0723	0.3937	0.9042	6.3	0.1190
MFI-*b*, α_F = 0.8	0.8710	0.0850	0.4103	0.8960	5.5	0.1290
MFI-*b*, α_F = 1.0	0.8301	0.1199	0.4752	0.8604	4.1	0.1699
MFI-*b*, α_F = 1.25	0.8277	0.1192	0.4805	0.8567	4.0	0.1723
MFI-*b*, α_F = 1.5	0.8176	0.1309	0.4970	0.8472	3.3	0.1824

Figure 2 shows the number of items, SE and bias conditional on the true θ for the simulated data conditions using the MFI-*b* alone. Figure 3 shows the same graphs for the simulated data conditions using the MFI-*b* with the composite stopping rule. Figure 4 shows the graphs for the real data conditions using the MFI-*b* alone, and Figure 5 shows the graphs for the real data with MFI-*b* using the composite stopping rule. In all of the figures, the solid line represents the performance of the standard MFI procedure. For simplicity only selected α_F values for the MFI-*b* are shown.

Aside from the shift that occurred because the real respondents did not cover the entire range of θ, the simulated and real data conditions behaved similarly. First we compare the CAT simulations using the MFI-*b* without the composite stopping rule to the CAT simulations using the standard MFI criterion. Simulations using MFI-*b* administered fewer items in the extreme values of θ, and more items where the information function for the bank is near its peak (see Figure 2a) when compared to the standard MFI criterion. As α_F increased, the number of items administered on average decreased. Also the MFI-*b* simulations had lower SEs than the standard MFI simulations near the peak of the item bank information function. This trend reversed for θs at the extremes (–3 or 3). The bias was fairly small and fairly consistent for all MFI-*b* simulations and the standard MFI

Figure 2. Results Conditional on *θ* for MFI-*b* in Simulated Data Conditions

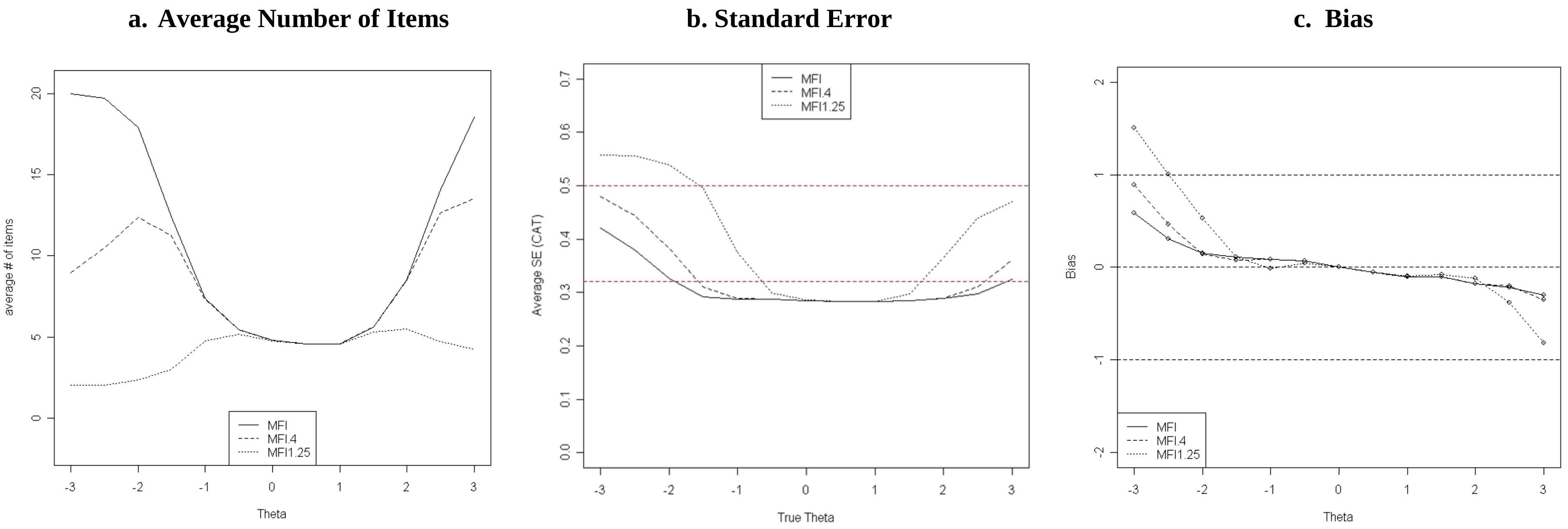


a. Average Number of Items

b. Standard Error

c. Bias

Figure 3. Results Conditional on θ for MFI-*b* With SE < 0.32 Composite Stopping Rule in Simulated Data Conditions

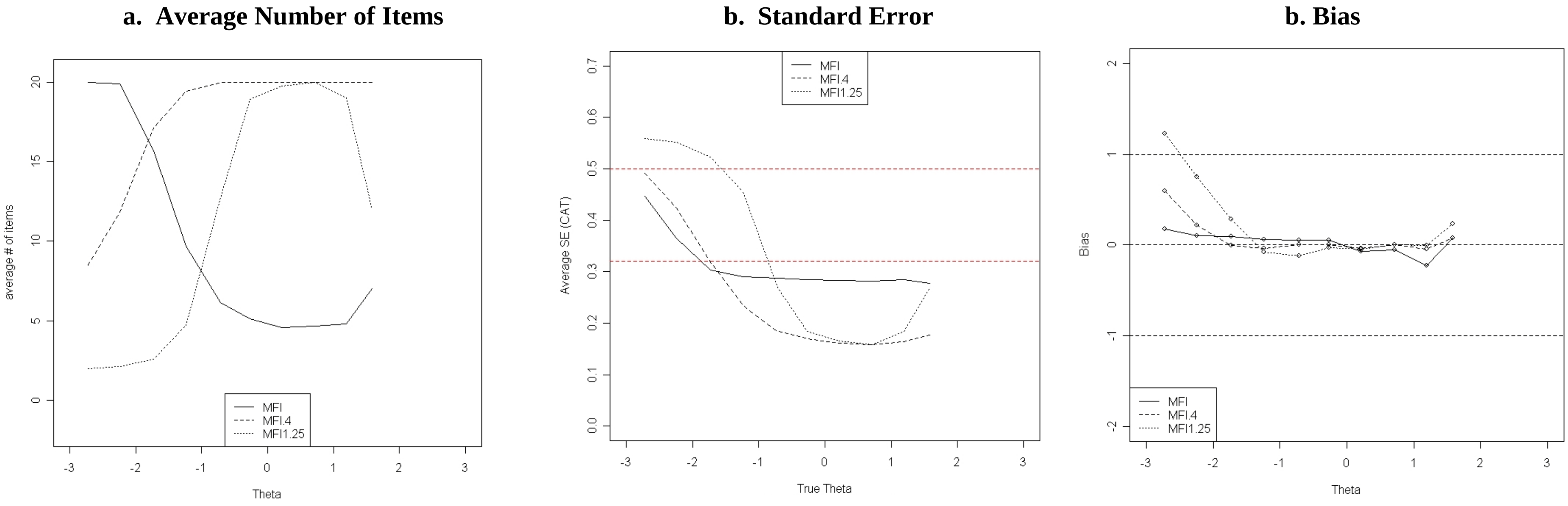


a. Average Number of Items

b. Standard Error

c. Bias

Figure 4. Results Conditional on θ for MFI-*b* in Real Data Conditions

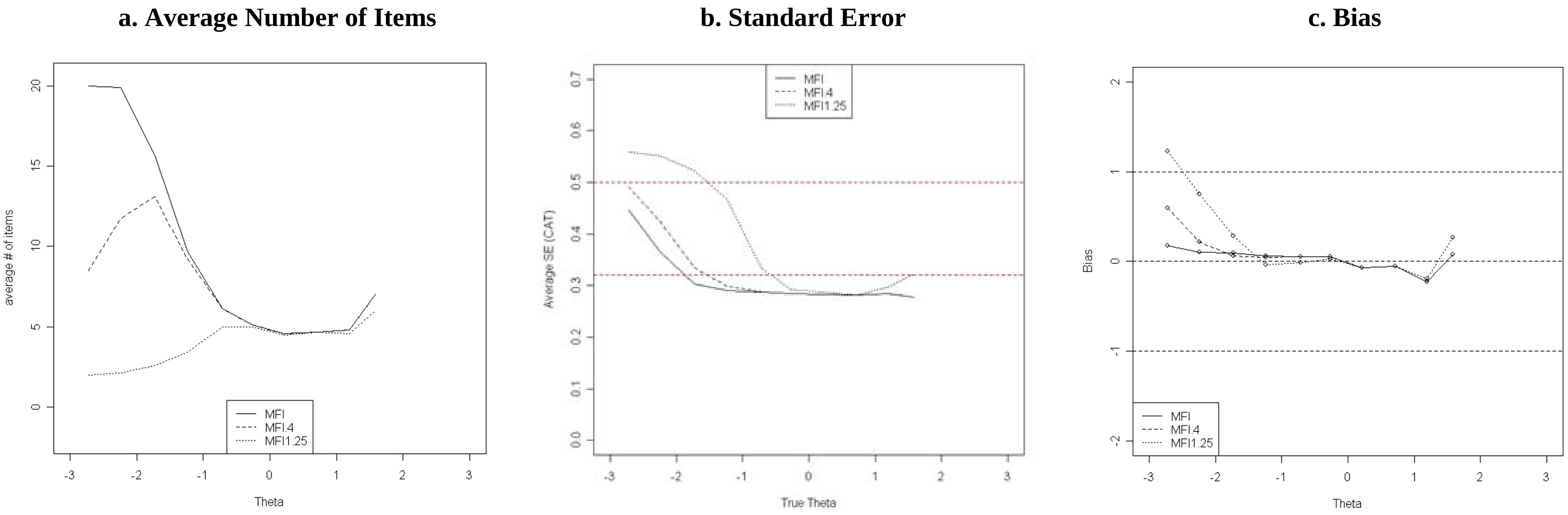


a. Average Number of Items

b. Standard Error

b. Bias

Figure 5. Results Conditional on *θ* for MFI-*b* With
Standard Error < 0.32 Composite Stopping Rule in Real Data Conditions



a. Average Number of Items

b. Standard Error

c. Bias

simulation for θs between –2 and 2. At extreme θs, the magnitude of the bias became larger, and this magnitude increased as α_F increased.

For the CAT simulations using the MFI-*b* composite stopping rule with SE ≤ 0.32 conditions, the central range of the plots (Figures 3 and 5) agreed with the CAT simulations using the standard MFI procedure in terms of the number of items administered and the SEs. The agreement over this range is expected; over this range the MFI-*b* composite CAT algorithms stopped administering items because the SE of the θ estimate fell below 0.32. Also, as expected, the bias was relatively unaffected.

Discussion

Using the MFI-*b*, we expected to see a relatively large decrease in the number of items administered (relatively large decrease in respondent burden) with a minimal increase in both the standard error of measurement and the magnitude of the bias. These were, however, not the results seen. Without the additional imposed rule to stop when the SE ≤ 0.32, the MFI-*b* in fact, administered more items than the standard MFI when the θ estimates fell between –1.5 and 2.0. This is because, in the item bank used, many items were highly informative in this region. Even at the highest value studied, the α_F threshold was still much lower than the information provided by many items located in this range. Therefore, the MFI-*b* continued to administer items until the maximum allowable number of items was attained. To ameliorate this, we introduced the composite rule: stop the MFI-*b* if either the item contributed information less than α_F or the standard error of the θ estimate falls at or below 0.32. This composite rule, although somewhat *ad hoc*, does control respondent burden better than the standard MFI stopping criterion. In the region where many items contributed a high degree of information, the standard error stopping rule keeps the number of items administered (and hence the respondent burden) at a reasonable level. In areas where the information contributed by many items is small, the α_F threshold imposes a cutoff determining when items are not worth administering.

Incorporating respondent burden into the selection algorithm using the composite rule resulted in shorter tests on average than the standard MFI. Based on the current studies, $\alpha_F = 0.4$ appears to be a reasonable value. For this value of α_F, the MFI-*b* composite rule resulted in about a 50% reduction in the number of items administered for θ estimates falling at the lower extreme, while there was less than 20% increase in the standard error of measurement. The bias increased a bit more substantially for θs at the extremes, but this is to be expected because there was not much information with fewer items administered, and the prior then has stronger influence on the estimate.

Also notice that most of the reduction in the respondent burden occurred at at θs below –1.5 or above 2. In patient reported outcome settings, respondents whose estimates are at extremes are also those who are most likely to be affected by respondent burden. Take depressive symptoms as an example. In our sample, a substantial number of people had a θ estimate below –1.5 (see the histogram in Figure 1). These people are at the low end of the depression continuum and are the people who would receive at least five fewer items under the MFI-*b* with $\alpha_F = 0.4$.

If respondent burden were not controlled, these people who are experiencing little to no depressive symptoms might get annoyed that they are repeatedly asked questions about symptoms they do not experience. In addition, administering additional items did not substantially reduce the standard error of measurement or the bias. The MFI-*b* accounts for this balance, and keeps the burden lower across the θ range. In practice, the appropriate balance (determined by the value of α_F.) would be determined empirically based on multiple factors, e.g., testing purposes, item bank characteristics.

Limitations

There are several limitations to this study. First, there is no definitive operating value for α_F. The current values studied show potential and utility of the method, and this was the intent of the current study. However, selecting an α_F value for a CAT to be used in practice would depend on the trait being measured and the importance of respondent burden as determined by experts. For example, an expert would select a small value of α_F if it is more important to have less biased estimates than to reduce the respondent burden. This is a preference-based approach to determining α_F. An area of future research is to explore standardized methods for determining preference-based α_F values.

A second limitation of this study is that we modeled the loss associated with respondent burden to be the same for all items at all values of θ. For an initial exploration into the problem, this was insightful and showcases the benefit of the procedure. This simple loss structure might not closely reflect reality for all traits or outcomes one would want to measure with CAT. For example, someone who is experiencing severe depression or fatigue might find items in general more burdensome than someone experiencing very little or no depression or fatigue. Also, certain items may be more burdensome in terms of the reading comprehension level they require, or the subject matter they include. For example, it is possible that an item about drug abuse or suicide may be more burdensome than an item about positive affects. Although not considered in this study, the loss function could be generalized to include such considerations. Potential future research would explore more accurate models of burden for the loss function.

A third limitation is that this study only considered the MEPV briefly as a basis for the framework and examined the MFI in more detail. However, it is likely that any selection method can be modeled using a loss function (or at least closely approximated by a loss function), and then the loss function could be developed to include respondent burden much as we have done in this study. In fact, the simplified loss function used for the MFI in this study could be re-expressed as a multi-objective loss function as described by van der Linden (2005). In this way, this study also outlined a general approach to adapting other selection methods to incorporate burden. We speculate that similar benefits will be seen with other selection methods as they are adapted to incorporate respondent burden.

Despite the simplified form of the loss function used to model respondent burden, this study showed that incorporating respondent burden further reduced that burden than a selection method that does not, especially at θ values where the item bank has poor coverage. This is especially evident in the real data CAT simulations. Figure 5 shows that most of the respondents had θ estimates on the low end of the scale where the total information of the bank was low (Figure 1 shows total information). Figures 3 and 5 show that the MFI-*b* composite stopping rule with $\alpha_F = 0.4$ yielded the best benefit in terms of burden for these respondents. Those with θ estimates at the lower extreme received about 50% fewer items but had relatively minimal reduction in the precision of their θ estimates. As mentioned previously, these respondents typically are the ones for who burden is more of a problem.

Conclusions

Typical CAT selection methods that do not consider respondent burden will administer a large number of items to respondents whose θs fall in areas where the bank has poor coverage because the selection criteria disregard respondent burden and the distribution of informative items in the bank. This might overburden those for whom burden is a problem. Therefore, reducing respondent burden for people whose θ estimates fall in a region where information is scarce is not only efficient but also beneficial. In summary, we have developed a straightforward way to incorporate respondent burden into a familiar item selection criterion that is easy to interpret, computationally easy to implement, and reduces the respondent burden better than the traditional CAT methods when the item bank consists of polytomous items.

APPENDIX A

The MEPV is a special case of the BASIS method when the BASIS method is restricted to a single-stage decision, and uses squared error loss with the EAP estimator for the theta estimate. The BASIS method is outlined in Swartz et al.(2009). Let $l\left(\theta,\mathbf{u}_{A_{k-1}},u_{i_k}\right)$ represent a loss function, that is, a function that models penalties that drive the problem. Specifically, in the CAT setting, our loss function (Equation 1) will model penalties associated with the precision of the estimate and respondent burden. If we restrict the decision to first selecting a single item at the current stage, and then deciding to either stop the test or continue the test and administer an additional item, the BASIS method defines the following selection and stopping rules (Swartz, et al., 2009):

$$\begin{aligned}
i_k^* &\equiv \arg\min_{j\in R_{k-1}} E_{U_j|\mathbf{u}_{A_{k-1}^*}}\left[E_{\theta|\mathbf{u}_{A_{k-1}^*},U_j}\left[l\left(\theta,\mathbf{u}_{A_{k-1}^*},U_j\right)|\,\mathbf{u}_{A_{k-1}^*},U_j=u_j\right]\right]\\
\text{Stop} \;\Leftrightarrow\; & E_{\theta|\mathbf{u}_{A_{k-1}^*},u_{i_k^*}}\left[l\left(\theta,\mathbf{u}_{A_{k-1}^*},u_{i_k^*}\right)|\,\mathbf{u}_{A_{k-1}^*},u_{i_k^*}\right]\le \min_{z\in R_k} E_{U_z|u_{A_k^*}}\left[E_{\theta|\mathbf{u}_{A_k^*},u_z}\left[l\left(\theta,\mathbf{u}_{A_k^*},U_z\right)|\,\mathbf{u}_{A_k^*},U_z=u_z\right]\right],\\
\text{Continue} \;\Leftrightarrow\; & E_{\theta|\mathbf{u}_{A_{k-1}^*},u_{i_k^*}}\left[l\left(\theta,\mathbf{u}_{A_{k-1}^*},u_{i_k^*}\right)|\,\mathbf{u}_{A_{k-1}^*},u_{i_k^*}\right]> \min_{z\in R_k} E_{U_z|u_{A_k^*}}\left[E_{\theta|\mathbf{u}_{A_k^*},u_z}\left[l\left(\theta,\mathbf{u}_{A_k^*},U_z\right)|\,\mathbf{u}_{A_k^*},U_z=u_z\right]\right].
\end{aligned} \tag{9}$$

In the above equations, i_k^* is the item to be administered at stage k, and $E_{\bullet|\bullet}\left[\bullet\right]$ is the conditional expectation operator. Starting with the general loss function described in Equation 1, then defining $G\left(\theta,\mathbf{u}_{A_{k-1}},u_{i_k}\right)$as outlined in Equation 3, using $\alpha(\theta)=\alpha_F$, and $C\left(\theta,\mathbf{u}_{A_{k-1}},u_{i_k}\right)=k$ as outlined in the text, results in the following loss function:

$$l\left(\theta,\mathbf{u}_{A_{k-1}},u_{i_k}\right)=\left(\theta-E(\theta\,|\,\mathbf{u}_{A_{k-1}},u_{i_k})\right)^2+\alpha k. \tag{10}$$

Using this loss function, we can further simplify the BASIS method as follows. Starting with i_k^*, we have

$$\begin{aligned}
i_k^* &\equiv \arg\min_{j\in R_{k-1}} E_{U_j|\mathbf{u}_{A_{k-1}^*}}\left\{E_{\theta|\mathbf{u}_{A_{k-1}^*},U_j}\left[l\left(\theta,\mathbf{u}_{A_{k-1}^*},u_j\right)|\,\mathbf{u}_{A_{k-1}^*},U_j=u_j\right]\right\}\\
&=\arg\min_{j\in R_{k-1}}\sum_{u=1}^{m_j} p_j(u\,|\,\mathbf{u}_{A_{k-1}^*})\left\{\int\left[\left(\theta-E(\theta\,|\,\mathbf{u}_{A_{k-1}^*},U_j=u)\right)^2+\alpha k\right]g(\theta\,|\,\mathbf{u}_{A_{k-1}^*},U_j=u)d\theta\right\}\\
&=\arg\min_{j\in R_{k-1}}\sum_{u=1}^{m_j} p_j(u\,|\,\mathbf{u}_{A_{k-1}^*})\left\{\alpha k+\int\left[\left(\theta-E(\theta\,|\,\mathbf{u}_{A_{k-1}^*},U_j=u)\right)^2\right]g(\theta\,|\,\mathbf{u}_{A_{k-1}^*},U_j=u)d\theta\right\}
\end{aligned} \tag{11}$$

Where the αk in the last line is not part of the integral because it is constant with respect to θ, and the integral of the posterior is 1. Note that the integral of the squared error loss multiplied by the posterior density is the posterior variance. Continuing from Equation 11 above:

$$\begin{aligned}
&\arg\min_{j\in R_{k-1}} \sum_{u=1}^{m_j} p_j(u\mid \mathbf{u}_{A^*_{k-1}})\left\{\alpha k + \int\left[\left(\theta - E(\theta\mid \mathbf{u}_{A^*_{k-1}}, U_j = u)\right)^2\right] g(\theta\mid \mathbf{u}_{A^*_{k-1}}, U_j = u)d\theta\right\} \\
&\quad = \arg\min_{j\in R_{k-1}} \sum_{u=1}^{m_j} p_j(u\mid \mathbf{u}_{A^*_{k-1}})\left\{\alpha k + \mathrm{var}(\theta\mid \mathbf{u}_{A^*_{k-1}}, U_j = u)\right\} \qquad (12) \\
&\quad = \arg\min_{j\in R_{k-1}} \sum_{u=1}^{m_j} p_j(u\mid \mathbf{u}_{A^*_{k-1}})\,\mathrm{var}(\theta\mid \mathbf{u}_{A^*_{k-1}}, U_j = u)
\end{aligned}$$

The second line above follows because the term αk is constant with respect to the summation and is also constant with respect to the minimization; therefore it does not affect the argument to be minimized. Note that the last expression is identical to the definition of the MEPV in Equation 2.

To define the stopping rule, we will first consider the left-hand side (LHS) and right-hand side (RHS) of the inequality separately.

$$\begin{aligned}
\mathrm{LHS} &= E_{\theta\mid \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}}\left[l\left(\theta, \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}\right)\mid \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}\right] \\
&= \int\left[\left(\theta - E(\theta\mid \mathbf{u}_{A^*_{k-1}}, u_{i^*_k})\right)^2 + \alpha k\right] g(\theta\mid \mathbf{u}_{A^*_{k-1}}, u_{i^*_k})d\theta \qquad (13) \\
&= \mathrm{var}(\theta\mid \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}) + \alpha k,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{RHS} &= \min_{z\in R_k} E_{U_z\mid \mathrm{u}_{A^*_k}}\left[E_{\theta\mid \mathbf{u}_{A^*_k}, U_z}\left[l\left(\theta, \mathbf{u}_{A^*_k}, U_z\right)\mid \mathbf{u}_{A^*_k}, U_z = u_z\right]\right] \qquad (14) \\
&= E_{U_{i^*_{k+1}}\mid \mathrm{u}_{A^*_k}}\left[E_{\theta\mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}}\left[l\left(\theta, \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}\right)\mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}\right]\right]
\end{aligned}$$

where i^*_{k+1} replaces the minimization over $z \in R_k$ in the last line because i^*_{k+1} is defined as the item that achieves that minimum. Because the second expectation within the square brackets is simply the LHS expression except evaluated at k+1, we can continue from equation 14 above and replace the quantity as follows:

$$\begin{aligned}
&E_{U_{i^*_{k+1}}\mid \mathrm{u}_{A^*_k}}\left[E_{\theta\mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}}\left[l\left(\theta, \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}\right)\mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}\right]\right] \\
&\quad = E_{U_{i^*_{k+1}}\mid \mathrm{u}_{A^*_k}}\left[\mathrm{var}(\theta\mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}) + \alpha(k+1)\right] \qquad (15) \\
&\quad = E_{U_{i^*_{k+1}}\mid \mathrm{u}_{A^*_k}}\left[\mathrm{var}(\theta\mid \mathbf{u}_{A^*_k}, U_{i^*_{k+1}})\right] + \alpha k + \alpha.
\end{aligned}$$

The last step follows from properties of the expectation operation.

Substituting the simplified versions of LHS and RHS into the expression for the stopping condition in Equation 9 yields:

$$\begin{aligned} LHS \le RHS &\Rightarrow \mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}) + \alpha k \le E_{U_{i^*_{k+1}} | \mathrm{u}_{A^*_k}} \left[\mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}) \right] + \alpha k + \alpha \\ &\Rightarrow \mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}) - E_{U_{i^*_{k+1}} | \mathrm{u}_{A^*_k}} \left[\mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}) \right] \le \alpha \end{aligned} \tag{16}$$

So we find the optimal decision according to the BASIS method as follows

$$i^*_k \equiv \arg\min_{j \in R_k} \sum_{u=1}^{m_j} p(U_j = u \,|\, \mathbf{u}_{A^*_{k-1}}) \,\mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_{k-1}}, U_j = u)$$

$$\text{Stop } \Leftrightarrow \mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}) - E_{U_{i^*_{k+1}} | \mathrm{u}_{A^*_k}} \left[\mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}) \right] \le \alpha,$$

$$\text{Continue} \Leftrightarrow \mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_{k-1}}, u_{i^*_k}) - E_{U_{i^*_{k+1}} | \mathrm{u}_{A^*_k}} \left[\mathrm{var}(\theta \,|\, \mathbf{u}_{A^*_k}, U_{i^*_{k+1}}) \right] > \alpha$$

Note that not accounting for burden is equivalent to setting $\alpha = 0$. Then the loss function (Equation 10) becomes simply squared error loss (Equation 3) and the BASIS method will always continue to administer items until the bank is exhausted, unless an additional stopping criterion is imposed. This is equivalent to the MEPV selection method, as presented in Equation 2.

APPENDIX B

The MFI is an *ad hoc* procedure. To derive an item selection procedure based on Fisher's information and respondent burden requires that we recall the definition of Fisher's information:

$$I_{U_{A_k}}(\theta) \equiv E\left[-\frac{\partial}{\partial \theta^2} \ln\left(L(\theta \,|\, U_{A_k}) \right) \right], \tag{16}$$

where the expectation is taken over the responses $U_{A_k} = (U_{i_1}, U_{i_2}, U_{i_3}, \ldots, U_{i_k})$, A_k is any set of administered items, not necessarily optimally selected, and $L(\theta \,|\, U_{A_k})$ is the likelihood function of θ given the responses. Also, recall the definition of observed information:

$$J_{\mathbf{u}_{A_k}}(\theta) \equiv -\frac{\partial}{\partial \theta^2} \ln\left(L(\theta \,|\, \mathbf{u}_{A_k}) \right), \tag{17}$$

where $\mathbf{u}_{A_k}$ is the vector of observed responses (as discussed above.)

Note that Fisher's information is simply the expectation of the observed information function with respect to the item responses. Both functions are additive (van der Linden & Pashley, 2000): $J_{\mathbf{u}_{A_k}}(\theta) \equiv -\frac{\partial}{\partial \theta^2} \ln\left(L(\theta \,|\, \mathbf{u}_{A_k}) \right) = \sum_{s=1}^{k} -\frac{\partial}{\partial \theta^2} \ln\left(L(\theta \,|\, u_{i_s}) \right)$. Next, if we view the negative of the observed information as a loss function, and consider minimizing the risk, defined as the expected loss (Ferguson, 1967) we have the following:

$$\begin{aligned} i_k &\equiv \arg\min_{j\in R_{k-1}} E\left[-J_{\mathbf{u}_{A^*_{k-1}},U_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) \right] \\ &= \arg\max_{j\in R_{k-1}} E\left[J_{\mathbf{u}_{A^*_{k-1}}}(\hat{\theta}_{\mathbf{u}_{A_{k-1}}}) + J_{U_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) \right] \\ &= \arg\max_{j\in R_{k-1}} \left\{ J_{\mathbf{u}_{A^*_{k-1}}}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) + E\left[J_{U_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) \right] \right\} \\ &= \arg\max_{j\in R_{k-1}} \left\{ E\left[J_{U_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) \right] \right\} \\ &= \arg\max_{j\in R_{k-1}} I_{U_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) \equiv MFI. \end{aligned} \tag{18}$$

Again the * on the set A^*_{k-1} indicates that the previously administered items were optimally selected using the MFI.

Therefore, the MFI is a method that selects the item that minimizes the risk (defined as the expected loss) at a given value of θ (in this case the estimate of θ based on the previous responses). Minimizing the risk at a given value of θ, although not guaranteed to give an optimal decision (DeGroot, 1969; Ferguson, 1967), is a reasonable approach to making a decision, and is the approach employed with the current MFI selection procedure. Therefore, we can extend the loss function to include patient burden:

$$l\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right) = -J_{\mathbf{u}_{A_{k-1}},U_j}(\hat{\theta}_{\mathbf{u}_{A_{k-1}}}) + \alpha(\theta) C\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right). \tag{19}$$

Making the same simplifying assumptions as we did for the MEPV and assigning $\alpha(\theta) = \alpha_F$ and $C\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right) = k$ yields the following loss function:

$$l\left(\theta, \mathbf{u}_{A_{k-1}}, u_{i_k}\right) = -J_{\mathbf{u}_{A_{k-1}},U_j}(\hat{\theta}_{\mathbf{u}_{A_{k-1}}}) + \alpha_F k. \tag{20}$$

The subscript *F* on α highlights the fact that the value associated with Fisher's information method may be different from the MEPV-*b*. As with the MEPV-*b*, the item selection does not change because $\arg\max_{j\in R_{k-1}} E\left[J_{\mathbf{u}_{A^*_{k-1}},U_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) - \alpha_F k \right] = \arg\max_{j\in R_{k-1}} E\left[J_{\mathbf{u}_{A^*_{k-1}},U_j}(\hat{\theta}_{\mathbf{u}_{A^*_{k-1}}}) \right]$ since the second term is constant with respect to the expectation and maximization. Meanwhile, as with MEPV, the new loss function does change the stopping rule:

$$\begin{aligned} \text{Stop } &\Leftrightarrow J_{\mathbf{u}_{A^*_{k-1}},u^*_k}(\hat{\theta}_{\mathbf{u}_{A^*_k}}) - \alpha_F k \geq \max_{z\in R_{k+1}} E\left[J_{\mathbf{u}_{A^*_k},U_z}(\hat{\theta}_{\mathbf{u}_{A^*_k}}) - \alpha_F(k+1) \right], \\ \text{Continue } &\Leftrightarrow J_{\mathbf{u}_{A^*_{k-1}},u^*_k}(\hat{\theta}_{\mathbf{u}_{A^*_k}}) - \alpha_F k < \max_{z\in R_{k+1}} E\left[J_{\mathbf{u}_{A^*_k},U_z}(\hat{\theta}_{\mathbf{u}_{A^*_k}}) - \alpha_F(k+1) \right]. \end{aligned} \tag{22}$$

Note that since there is now a penalty term for asking additional items, the stopping rule compares the loss at the current stage *k* after administering the *k* items, with the risk (expected loss) that would be incurred after administering the next item. Simplifying the stopping condition further yields:

$$\begin{aligned}
\text{Stop } &\Leftrightarrow J_{\mathbf{u}_{A_k^*}}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) - \alpha_F k \geq \max_{z \in R_{k+1}} E\left[J_{\mathbf{u}_{A_k^*},U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) - \alpha_F (k+1) \right] \\
&\Rightarrow J_{\mathbf{u}_{A_k^*}}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) - \alpha_F k \geq \max_{z \in R_{k+1}} E\left[J_{\mathbf{u}_{A_k^*}}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) + J_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) - \alpha_F k - \alpha_F \right] \\
&\Rightarrow J_{\mathbf{u}_{A_k^*}}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) - \alpha_F k - J_{\mathbf{u}_{A_k^*}}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) + \alpha_F k + \alpha_F \geq \max_{z \in R_{k+1}} E\left[J_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) \right] \\
&\Rightarrow \alpha_F \geq \max_{z \in R_{k+1}} E\left[J_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) \right] \\
&\Rightarrow \alpha_F \geq \max_{z \in R_{k+1}} I_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}})
\end{aligned} \tag{23}$$

And similarly for the continue condition, so that the MFI-*b* becomes:

$$\begin{gathered}
i_k^* \equiv \arg\max_{j \in R_{k-1}} I_{\mathbf{u}_{A_{k-1}^*},U_j}(\hat{\theta}_{\mathbf{u}_{A_{k-1}^*}}) = \arg\max_{j \in R_{k-1}} I_{U_j}(\hat{\theta}_{\mathbf{u}_{A_{k-1}^*}}) \\
\text{Stop } \Leftrightarrow \alpha_F \geq \max_{z \in R_{k+1}} I_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}}) \\
\text{Continue} \Leftrightarrow \alpha_F < \max_{z \in R_{k+1}} I_{U_z}(\hat{\theta}_{\mathbf{u}_{A_k^*}})
\end{gathered} \tag{21}$$

References

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer-Verlag.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16 Suppl 1*, 95-108.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, Fla.: St. Lucie Press.

Cella, D., & Chang, C. H. (2000). A discussion of item response theory and its applications in health status assessment. *Medical Care, 38*(9 Suppl), II66-72.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33,* 419-440.

DeGroot, M. H. (1969). *Optimal statistical decisions*. New York: McGraw-Hill.

Dodd, B. G., Koch, W. R., & Deayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129-143.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. New York: Academic Press.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*(10), 2277-2291.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.).

Boca Raton FL: Chapman & Hall/CRC.

Hart, D. L., Wang, Y. C., Stratford, P. W., & Mioduski, J. E. (2008a). Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function. *Quality of Life Research, 17*(8), 1081-1091.

Hart, D. L., Wang, Y. C., Stratford, P. W., & Mioduski, J. E. (2008b). A computerized adaptive test for patients with hip impairments produced valid and responsive measures of function. *Archives of Physical Medicine and Rehabilitation, 89*(11), 2129-2139.

Jenkinson, C., Fitzpatrick, R., Garratt, A., Peto, V., & Stewart-Brown, S. (2001). Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). *Journal of Neurology Neurosurgery and Psychiatry, 71*(2), 220-224.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing : Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision-theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386.

McHorney, C. A. (2003). Ten recommendations for advancing patient-centered outcomes measurement for older persons. *Annals of Internal Medicine, 139*(5 Pt 2), 403-409.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Owen, R. J. (1975). Bayesian sequential procedure for quantal response in context of adaptive mental testing. *Journal of the American Statistical Association, 70*(350), 351-356.

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*, 1-20.

Reeve, B. B. (2006). Special issues for building computerized-adaptive tests for measuring patient-reported outcomes: The National Institute of Health's investment in new technology. *Medical Care, 44*(11 Suppl 3), S198-S204.

Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Quality of Life Research, 6*(6), 595-600.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs no. 17, 34*(4 pt 2).

Science Advisory Committee of the Medical Outcomes Trust (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*(3), 193-205.

Swartz, R. J., & Choi, S. W. (2008, June). *Computerized adaptive testing item selection procedures for dichotomous items - What do we lose by being greedy?* . Paper presented at the 73rd Annual Meeting of the Psychometric Society, Durham, NH.

Swartz, R. J., Choi, S. W., & Herrick, R. C. (2009, Bayesian adaptive sequential item selection (BASIS): An optimal approach to item selection. . *UT MD Anderson Cancer Center Department of Biostatistics Working Paper Series, 53*, http://www.bepress.com/mdandersonbiostat/paper53

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory (Version 7.02327.3) [Computer software]. Lincolnwood, IL: Scientific Software International.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: a primer* (2 ed., pp. 101-134). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. J. (2005). *Linear models of optimal test design*. New York, NY: Springer.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Norwell, MA: Kluwer Academic.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-26). Norwell, MA: Kluwer Academic.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics, 24*(3), 271-292.

Vos, H. J. (2000). A Bayesian procedure in the context of sequential mastery testing. *Psicologica, 21*, 191-211.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for 'anxiety' (Anxiety-CAT). *Quality of Life Research, 16 Suppl 1*, 143-155.