Statistica Sinica **27** (2017), 1987-2010
doi:https://doi.org/10.5705/ss.202015.0304

COMPUTERIZED ADAPTIVE TESTING THAT ALLOWS FOR RESPONSE REVISION: DESIGN AND ASYMPTOTIC THEORY

Shiyu Wang[1], Georgios Fellouris[2] and Hua-Hua Chang[2]

[1] *University of Georgia and* [2] *University of Illinois at Urbana Champaign*

Abstract: In Computerized Adaptive Testing (CAT), items are selected in real time and are adjusted to the test-taker's ability. While CAT has become popular for many measurement tasks, such as educational testing and patient reported outcomes, it has been criticized for not allowing examinees to review and revise their answers. In this work, we propose a novel CAT design that preserves the efficiency of a conventional CAT, but allows test-takers to revise their previous answers at any time during the test. The proposed method relies on a polytomous Item Response model that describes the first response to each item, as well as any subsequent responses to it. Each item is selected in order to maximize the Fisher information of the model at the current ability estimate, which is given by the maximizer of a partial likelihood function. We establish the strong consistency and asymptotic normality of the final ability estimator under minimal conditions on the test-taker's revision behavior. We present the findings of two simulation studies that illustrate our theoretical results, as well as the behavior of the proposed design in a realistic item pool.

Key words and phrases: Asymptotic normality, computerized adaptive testing, consistency, item response theory, martingale limit theory, nominal response model, sequential design.

1. Introduction

A main goal in educational assessment is the accurate estimation of the test-taker's ability. In a conventional paper-pencil test, this estimation is based on the examinee's responses to a preassembled set of items. However, in Computerized Adaptive Testing (CAT), as it was originally conceived by Lord (1971), items are selected in real time and are tailored to the examinee's ability, which is learned as the test progresses. This feature is especially important for examinees at the two extreme ends of the ability distribution, who may otherwise receive items that are either too difficult or too easy.

The design of CAT is based on Item Response Theory (IRT) models, which describe the probability of a correct answer given the examinee's ability and the item itself. The simplest IRT model is the Rasch model (Rasch (1993)), in which the probability of a correct answer is equal to $H(\theta - b)$, where H is the cdf of the logistic distribution, θ is an unknown, scalar parameter that represents the ability of the examinee, and b a known, scalar parameter that reflects the difficulty of the item. A generalization of the Rasch model is the three-parameter logistic model (3PL) model, in which the probability of a correct answer is $c+(1-c)H(a(\theta-b))$, where $c \in (0,1)$ represents the probability of guessing the right answer, and $a > 0$ is the discrimination parameter of the item. An intermediate model, the two-parameter logistic model (2PL), arises when we set $c = 0$.

A standard approach for item selection in CAT, proposed by Lord (1980), is to select the item with the maximum Fisher information at each step. In the case of the Rasch model, this means that item i should be selected such that its difficulty parameter b_i equals to θ. Since θ is unknown, this suggests setting b_i equal to an estimate of θ based on the first $i-1$ observations. For the adaptive estimation of θ, Lord (1971) proposed the Stochastic Approximation algorithm of Robbins and Monro (1951). However, this non-parametric approach can be very inefficient with binary data, as was shown by Lai and Robbins (1979). For this reason, Wu (1985, 1986) suggested using the Maximum Likelihood Estimator (MLE) of θ based on the first $i-1$ observations. Following this approach, coupled with the information maximizing item selection strategy, Ying and Wu (1997) established the strong consistency and asymptotic normality of the resulting estimator under the Rasch model, whereas Chang and Ying (2009) extended these results to the case of the 2PL and 3PL model. Alternative item selection algorithms have been proposed in the literature, such as the approximate Bayes procedures (Owen (1969, 1975)), the maximum global-information criterion (Chang and Ying (1996)), and various criteria in Veerkamp and Berger (1997). Moreover, many modifications on the original item selection algorithms have been suggested in order to incorporate non-statistical constraints, such as content coverage, answer key distribution, and exposure control (Wang et al. (2016a); Chang and Ying (1999); Luecht (1998); Swanson and Stocking (1993)).

Thanks to these statistical advances, as well as the rapid development of modern technology, CAT has become popular for many kinds of measurement tasks, such as educational testing, patient reported outcomes, and quality of life measurement. Examples of large-scale CATs include the Graduate Management Admission Test (GMAT), the National Council Licensure Examination (NCLEX) for nurses, and the Armed Services Vocational Aptitude Battery (ASVAB) (Chang and Ying (2007)). Beyond the problem of ability estimation, CAT has been applied to mastery testing (Sie et al. (2015); Bartroff, Finkelman and Lai (2008)) and cognitive diagnosis (liu2013rate). However, currently operational CAT programs typically do not allow examinees to revise their responses to previously administered items during the test (Vispoel et al. (1999)), and this is one of the reasons that some testing programs have decided to switch to other modes of testing , such as Multistage Adaptive Testing (Luecht and Nungester (1998)).

A main argument against response revision among practitioners and researchers who oppose this feature is that it violates the adaptive nature of CAT. Specifically, it has been argued that allowing for response revision decreases estimation efficiency and increases bias (Stocking (1997); Vispoel et al. (1999)), as well as that it gives the opportunity to disingenuous examinees to artificially inflate their test scores by adopting deceptive test-taking strategies (Wainer (1993); Kingsbury (1996); Wise et al. (1999)). On the other hand, it has been argued that response revision in CAT can minimize measurement error, leading to more accurate inference, and that it can lower the anxiety levels of the examinees, leading to a friendlier testing environment (Wise (1996); Vispoel, Hendrickson and Bleiler (2000)). It has been reported that examinees would favor the response revision feature in CAT (Vispoel and Coffman (1992); Han (2015)), whereas the desire for review opportunities has also been verified in other studies of computerized tests, e.g. (Schmidt, Urry and Gugel (1978)).

Overall, the absence of the opportunity to revise in CAT has been a main concern for both examinees and testing companies, and a number of modified CAT designs have been proposed in order to incorporate this feature (Stocking (1997); Vispoel, Hendrickson and Bleiler (2000); Han (2013)). In order to prevent the potential dangers of revision, these designs have postulated quite limited revision rules, such as an upper bound on the number of items that can be revised.

Under such rules and restrictions, it has been reported that response revision does not impact significantly the estimation accuracy and efficiency of CAT. However, these conclusions were based only on simulation experiments and not supported theoretically.

In this work, we propose and analyze a novel CAT design whose goal is to preserve the advantages of conventional CAT with respect to estimation efficiency, but at the same time to allow examinees to revise their answers at any time during the test.

In order to achieve this, we use a different modeling framework than that of a typical CAT design. Indeed, although most operational CAT programs employ multiple-choice items, they model them in a dichotomous way, specifying the probability of each response being either right or wrong. We use a polytomous IRT model, the nominal response model proposed by Bock (1972), and specify the probability that the examinee selects each category of a given item. Based on this model, we postulate a joint probability model for the first answer to each item and any subsequent revisions to it, and we update the ability parameter after each response with the maximizer of the likelihood of all responses, first answers, *and* revisions. We do not make any assumptions regarding the decision of the examinee to revise or not at each step and, whenever the examinee asks for a new item, we select the one with the maximum Fisher information at the current estimate of the ability level.

We provide an asymptotic study of the proposed method as the number of administered items goes to infinity, apparently the first rigorous analysis of a CAT design that allows for response revision. Our main result is that the proposed estimator is asymptotically normal under a stability assumption on the cumulative Fisher information. That is satisfied, for example, when the number of revisions is small relative to the number of items.

We consider separately the case of a CAT that is based on the nominal response model, but does not allow for response revision. Again, there has apparently not been any theoretical analysis of a conventional CAT based on a polytomous IRT model, so the corresponding asymptotic results are of independent interest. They help us illustrate the conceptual and technical differences between the traditional CAT setup, where the number of observed responses coincides with the number administered items at any time during the test, and the proposed setup in which the number of responses and items are , in general, different.

The rest of the paper is organized as follows. In Section 2, we introduce the nominal response model and present its main properties. In Section 3 we consider the design and analysis of a CAT that is based on the nominal response model, but in which response revision is not allowed. In Section 4, we present and analyze the proposed CAT design that allows for response revision. In Section 5, we present the findings of two simulation studies that illustrate our results and evaluate our proposed design in a realistic setup. We conclude in Section 6. Throughout the paper, we focus on a single examinee whose ability is quantified by an unknown, scalar parameter $\theta \in \mathbb{R}$, and we denote by $\mathsf{P}_\theta/\mathsf{E}_\theta/\mathsf{Var}_\theta$ the corresponding probability measure/expectation/variance.

2. Nominal Response Model

Let X be the response to a multiple-choice item with $m \geq 2$ categories. We write $X = k$ when the examinee chooses category $k \in [m] := \{1, \ldots, m\}$, and we assume that

$$\mathsf{P}_\theta(X = k) = \frac{\exp(a_k\theta + c_k)}{\sum_{h=1}^{m} \exp(a_h\theta + c_h)}, \quad k \in [m], \tag{2.1}$$

where $\{a_k, c_k\}_{1\leq k\leq m}$ are known, item-specific real numbers such that

$$\sum_{k=1}^{m} |a_k| \neq 0, \quad \sum_{k=1}^{m} |c_k| \neq 0, \quad \text{and} \quad \sum_{k=1}^{m} a_k = \sum_{k=1}^{m} c_k = 0. \tag{2.2}$$

Thus, the distribution of X is specified by the ability of the examinee, θ, and the item-specific vector $\mathbf{b} = (a_2, \ldots, a_m, c_2, \ldots, c_m)$. To lighten the notation, we write

$$p_k(\theta; \mathbf{b}) := \mathsf{P}_\theta(X = k), \quad k \in [m], \tag{2.3}$$

and we denote by $\mathbb{B}$ the subset of $\mathbb{R}^{2m-2}$ in which $\mathbf{b}$ takes values. For the log-likelihood function, the score function, and the Fisher information based on a single observation we write

$$\begin{aligned} &\ell(\theta; \mathbf{b}, k) := \log\big(p_k(\theta; \mathbf{b})\big), \ s(\theta; \mathbf{b}, k) := \frac{\partial}{\partial\theta}\ell(\theta; \mathbf{b}, k), \quad k \in [m], \\ &J(\theta; \mathbf{b}) := \mathsf{Var}_\theta[s(\theta; \mathbf{b}, X)]. \end{aligned} \tag{2.4}$$

With a direct computation it follows that

$$s(\theta; \mathbf{b}, k) = a_k - \bar{a}(\theta; \mathbf{b}), \quad k \in [m], \tag{2.5}$$

$$J(\theta; \mathbf{b}) = \sum_{k=1}^{m} \Big(a_k - \bar{a}(\theta; \mathbf{b})\Big)^2 p_k(\theta; \mathbf{b}), \tag{2.6}$$

where $\bar{a}(\theta; \mathbf{b})$ is a weighted average of the a_k's,

$$\bar{a}(\theta; \mathbf{b}) := \sum_{h=1}^{m} a_h\, p_h(\theta; \mathbf{b}). \tag{2.7}$$

The derivative of $s(\theta; \mathbf{b}, k)$ with respect to θ does not depend on k. And for every $k \in [m]$, we have

$$s'(\tilde{\theta}; \mathbf{b}) := \frac{\partial}{\partial\theta} s(\theta; \mathbf{b}, k)\Big|_{\theta=\tilde{\theta}} = -J(\tilde{\theta}; \mathbf{b}). \tag{2.8}$$

In the special case of binary data ($m = 2$), $a_1 = -a_2$, $c_1 = -c_2$,

$$p_2(\theta;\mathbf{b}) = 1 - p_1(\theta;\mathbf{b}) = \frac{\exp(2a_2\theta + 2c_2)}{1+\exp(2a_2\theta+2c_2)}, \tag{2.9}$$

i.e., we recover the 2PL model with discrimination parameter $2|a_1|$ and difficulty parameter $-c_2/a_2$.

For a given item parameter vector $\mathbf{b}$, take

$$\begin{aligned} a^*(\mathbf{b}) &:= \max_{k\in[m]} a_k, \quad k^*(\mathbf{b}) := \{k\in[m] : a_k = a^*(\mathbf{b})\}, \\ a_*(\mathbf{b}) &:= \min_{k\in[m]} a_k, \quad k_*(\mathbf{b}) := \{k\in[m] : a_k = a_*(\mathbf{b})\}. \end{aligned} \tag{2.10}$$

Here $k^*(\mathbf{b})$ and $k_*(\mathbf{b})$ are singletons when $m=3$, but this is not necessarily the case when $m>3$. For any given item parameter vector $\mathbf{b}\in\mathbb{B}$, it is easy to see that

$$\lim_{\theta\to-\infty} \bar{a}(\theta;\mathbf{b}) = a_*(\mathbf{b}), \quad \lim_{\theta\to\infty} \bar{a}(\theta;\mathbf{b}) = a^*(\mathbf{b}), \quad \lim_{|\theta|\to\infty} J(\theta;\mathbf{b}) = 0.$$

We denote by $J^*(\theta)$ and $J_*(\theta)$ the maximal and minimal, respectively, Fisher information in the item pool for an examinee with ability level θ,

$$J_*(\theta) := \inf_{\mathbf{b}\in\mathbb{B}} J(\theta;\mathbf{b}) \quad \text{and} \quad J^*(\theta) := \sup_{\mathbf{b}\in\mathbb{B}} J(\theta;\mathbf{b}). \tag{2.11}$$

The proofs of the following two lemmas are presented in S1 in the supplementary material.

Lemma 1. *Let $g : \mathbb{R}\times\mathbb{B}\to\mathbb{R}$ be a jointly continuous function and set*

$$g^*(\cdot) := \sup_{\mathbf{b}\in\mathbb{B}} g(\cdot,\mathbf{b}), \quad g_*(\cdot) := \inf_{\mathbf{b}\in\mathbb{B}} g(\cdot,\mathbf{b}).$$

If $\mathbb{B}$ is compact, then g^* and g_* are continuous functions, and if $x_n\to x_0$, then

$$\sup_{\mathbf{b}\in\mathbb{B}} |g(x_n,\mathbf{b}) - g(x_0,\mathbf{b})| \to 0.$$

Lemma 2. *If $\mathbb{B}$ is compact, then*

$$|s(\theta;b,\cdot)| \le K, \quad 0 < J_*(\theta) \le J^*(\theta) \le K,$$

where K is a constant that does not depend on θ or $\mathbf{b}$.

In what follows, the subset $\mathbb{B}\subset\mathbb{R}^{2m-2}$, that represents the underlying item bank/pool, is assumed to be *compact.*

3. Standard CAT with Nominal Response Model

In this section we consider the design of a CAT that is based on the nominal response model, but is conventional in that it does not allow for response revision.

3.1. Problem formulation

Let n be the total number of items that will be administered to the examinee and let X_i denote the response to item i, where $i \in [n] := \{1, \ldots, n\}$. To lighten the notation, we assume that each item has the same number of categories $m \geq 2$, and we write $X_i = k$ if the examinee chooses category k in item i, where $k \in [m]$ and $i \in [n]$. The responses are assumed to be governed by the nominal response model (2.1)–(2.3), thus

$$\mathsf{P}_\theta(X_i = k) := p_k(\theta; \mathbf{b}_i), \quad k \in [m],\ i \in [n], \tag{3.1}$$

where θ is the unknown ability parameter, and the item parameter vector $\mathbf{b}_i := (a_{i2}, \ldots, a_{im}, c_{i2}, \ldots, c_{im})$ takes values in a compact set $\mathbb{B} \subset \mathbb{R}^{2m-2}$ with components satisfying (2.2). In practice, there is only a finite number of items in a given item bank and there are further restrictions on the exposure rate of the items (Chang and Ying (1999)), so each $\mathbf{b}_i$ cannot actually take any value in $\mathbb{B}$. Nevertheless, this assumption will allow us to obtain a benchmark for the achievable large-sample performance in CAT.

We assume that the responses are conditionally independent given the selected items, in the sense that

$$\mathsf{P}_\theta(X_{1:i} | \, \mathbf{b}_{1:i}) = \prod_{j=1}^{i} \mathsf{P}_\theta(X_j | \mathbf{b}_j), \quad i \in [n], \tag{3.2}$$

where $X_{1:i} \equiv (X_1, \ldots, X_i)$ and $\mathbf{b}_{1:i} \equiv (\mathbf{b}_1, \ldots, \mathbf{b}_i)$. In a paper-pencil test where the selected items are fixed in advance, $\mathbf{b}_{1:n}$ is a deterministic vector. This is not the case in CAT, where items are determined in real time based on the already observed responses. Specifically, if we denote by $\mathcal{F}_i^X := \sigma(X_1, \ldots, X_i)$ the information contained in the first i responses, then $\mathbf{b}_i$ must be a $\mathcal{F}_{i-1}^X$-measurable, $\mathbb{B}$-valued random vector for every $2 \leq i \leq n$, whereas $\mathbf{b}_1$ is arbitrary.

The specification of the *item selection strategy*, $(\mathbf{b}_i)_{1 \leq i \leq n}$, is a major component in CAT design. Here we adopt the standard approach of selecting each item in order to maximize the Fisher information at the current estimate of the ability level, i.e., item $i+1$ is selected such that

$$J(\widehat{\mathbf{b}}_i) = \max_{\mathbf{b} \in \mathbb{B}} J(\hat{\theta}_{i-1}; \mathbf{b}), \quad 2 \leq i \leq n, \tag{3.3}$$

where J is the Fisher information function of the nominal response model given by (2.6), and $\hat{\theta}_i$ is an estimate of θ based on the first i responses. Due to assumptions (3.1)–(3.2), the conditional log-likelihood and score functions of the first i responses given arbitrary selected items $\mathbf{b}_{1:i}$ take the form

$$L_i(\theta) := \log \mathsf{P}_\theta(X_{1:i} \mid \mathbf{b}_{1:i}) = \sum_{j=1}^{i} \ell(\theta; \mathbf{b}_j, X_j),$$

$$S_i(\theta) := \frac{d}{d\theta} L_i(\theta) = \sum_{j=1}^{i} s(\theta; \mathbf{b}_j, X_j), \tag{3.4}$$

where $\ell(\theta; \mathbf{b}_j, X_j)$ and $s(\theta; \mathbf{b}_j, X_j)$ are the log-likelihood function and score function, respectively, of the j^{th} response, defined in (2.4). Our estimate of θ based on the first i observations is given by the root of $S_i(\theta)$, which exists and is unique for every $i > n_0$ such that

$$n_0 := \max\Big\{ i \in \{1, \ldots, n\} : X_j \in k^*(\mathbf{b}_j)\ \forall\, j \leq i \quad \text{or} \quad X_j \in k_*(\mathbf{b}_j)\ \forall\, j \leq i \Big\},$$

where $k^*(\mathbf{b})$ and $k_*(\mathbf{b})$ are defined in (2.10). Thus the root of $S_i(\theta)$ exists and is unique as long as either at least one of the first i responses does not correspond to a category with the largest a-value, or to a category with the smallest a-value. For $i \leq n_0$, we need an alternative ability estimator, such as the Bayesian estimator in Bock and Aitkin (1981). Alternatively, for $i \leq n_0$ we can set $\hat{\theta}_0 = 0$ and $\hat{\theta}_i = \hat{\theta}_{i-1} + d$ (resp. $\hat{\theta}_i = \hat{\theta}_{i-1} - d$) if the initial responses have the largest (resp. smallest) a-value, where d is some predetermined constant.

Lemma 3. $\mathsf{P}_\theta(S_n(\hat{\theta}_n) = 0$ *for all large* $n) = 1$.

The proof of Lemma 3 is presented in S2 of the supplementary material.

3.2. Asymptotic analysis

In this section, we establish the asymptotic properties of the proposed ability estimator, $\hat{\theta}_n$, assuming that (3.1)–(3.2) hold. Specifically, we establish strong consistency for an arbitrary item selection strategy and its asymptotic normality when the information-maximizing item selection strategy (3.3) is adopted.

We first show that for an arbitrary item selection strategy, $(\mathbf{b}_i)_{1 \leq i \leq n}$, the corresponding score function $\{S_n(\theta)\}_{n \in \mathbb{N}}$ is a martingale with mean 0 and predictable variation equal to the conditional Fisher information

$$I_n(\theta) := \sum_{i=1}^{n} J(\theta; \mathbf{b}_i), \quad n \in \mathbb{N}, \tag{3.5}$$

where $J(\theta; \mathbf{b}_i)$ is the Fisher information of the i^{th} item given by (2.6). From (2.8) it follows that

$$S_n^{'}(\tilde{\theta}) := \frac{d}{d\theta} S_n(\theta)\Big|_{\theta = \tilde{\theta}} = \sum_{i=1}^{n} -J(\tilde{\theta}; \mathbf{b}_i) = -I_n(\tilde{\theta}). \tag{3.6}$$

Lemma 4. *For any item selection strategy, the score function* $\{S_n(\theta)\}_{n \in \mathbb{N}}$ *is a* $\{\mathcal{F}_n\}$*-martingale under* P_θ*, with bounded increments, mean* 0 *and predictable variation*

$$\langle S(\theta)\rangle_n := \sum_{i=1}^{n} \mathsf{E}_\theta \left[\left(S_i(\theta) - S_{i-1}(\theta)\right)^2 \,|\, \mathcal{F}_{i-1} \right] = I_n(\theta).$$

The proof is presented in S2 of the supplementary materials.

Theorem 1. For any item selection strategy, $\hat{\theta}_n \to \theta$ P_θ-a.s. and

$$\frac{I_n(\hat{\theta}_n)}{I_n(\theta)} \to 1 \quad \mathsf{P}_\theta - \text{a.s.}. \tag{3.7}$$

The proof is presented in Appendix S2 of the supplementary materials. It is interesting to note that (3.7) remains valid for any strongly consistent estimator of θ and that the strong consistency of $\hat{\theta}_n$ is established for any item selection strategy. This is due to the compactness of the item parameter space, $\mathbb{B}$. Were this not the case, the resulting estimator could fail to be consistent (see Chang and Ying (2009) for a counterexample).

Theorem 2. If $I_n(\theta)/n$ converges in probability to some positive constant under P_θ, then as $n \to \infty$ we have

$$\sqrt{I_n(\hat{\theta}_n)}\,(\hat{\theta}_n - \theta) \longrightarrow \mathcal{N}(0,1). \tag{3.8}$$

When the information-maximizing item selection strategy (3.3) is adopted, in which case

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}\left(0, [J^*(\theta)]^{-1}\right). \tag{3.9}$$

Proof. From Lemma 4 we know that $\{S_n(\theta)\}_{n\in\mathbb{N}}$ is a martingale with bounded increments, mean 0, and predictable variation $\{I_n(\theta)\}_{n\in\mathbb{N}}$. Then, if $I_n(\theta)/n$ converges in probability to some positive constant under P_θ, we can apply the Martingale Central Limit Theorem (see, e.g., Billingsley (2008), p.481) and obtain

$$\frac{S_n(\theta)}{\sqrt{I_n(\theta)}} \longrightarrow \mathcal{N}(0,1).$$

From Lemma 3 and a Taylor expansion of $S_n(\theta)$ around $\hat{\theta}_n$ it follows that there exists some $\tilde{\theta}_n$ that lies between $\hat{\theta}_n$ and θ such that

$$\begin{aligned} 0 = S_n(\hat{\theta}_n) &= S_n(\theta) + S_n^{'}(\tilde{\theta}_n)(\hat{\theta}_n - \theta) \\ &= S_n(\theta) - I_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta) \quad \mathsf{P}_\theta - \text{a.s.}, \end{aligned} \tag{3.10}$$

where the second equality follows from (3.6). Consequently,

$$\frac{I_n(\tilde{\theta}_n)}{I_n(\theta)} \sqrt{I_n(\theta)}\,(\hat{\theta}_n - \theta) \to \mathcal{N}(0,1).$$

Since $\tilde{\theta}_n$ lies between $\hat{\theta}_n$ and θ, similarly to (3.7), we can show that

$$\frac{I_n(\tilde{\theta}_n)}{I_n(\theta)} \to 1 \quad \mathsf{P}_\theta - \text{a.s.}.$$

Thus, from an application of Slutsky's theorem we obtain

$$\sqrt{I_n(\theta)}\,(\hat{\theta}_n - \theta) \longrightarrow \mathcal{N}(0,1). \tag{3.11}$$

From (3.7) and another application of Slutsky's theorem we obtain (3.8).

In order to prove the second part of this theorem, it suffices to show that

$$\frac{1}{n} I_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} J(\theta; \widehat{\mathbf{b}}_i) \to J^*(\theta) \quad \mathsf{P}_\theta - \text{a.s}, \tag{3.12}$$

where $(\widehat{\mathbf{b}}_i)_{1\le i\le n}$ are the item parameters selected according to (3.3). To prove (3.12) it suffices to show that $J(\theta; \widehat{\mathbf{b}}_n) \to J^*(\theta)$ P_θ–a.s. Since $J(\theta; \mathbf{b})$ is jointly continuous and $\hat{\theta}_n$ strongly consistent, from Lemma 1 it follows that

$$\sup_{\mathbf{b}\in\mathbb{B}} |J(\hat{\theta}_n; \mathbf{b}) - J(\theta; \mathbf{b})| \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.13}$$

and, consequently,

$$|J(\hat{\theta}_n; \widehat{\mathbf{b}}_n) - J(\theta; \widehat{\mathbf{b}}_n)| \to 0 \quad \mathsf{P}_\theta - a.s..$$

Therefore, it suffices to show that $J(\hat{\theta}_n; \widehat{\mathbf{b}}_n) \to J^*(\theta)$ $\mathsf{P}_\theta -$a.s. From the definition of $(\widehat{\mathbf{b}}_n)$ in (3.3) we have $J(\hat{\theta}_{n-1}; \widehat{\mathbf{b}}_n) = J^*(\hat{\theta}_{n-1})$, and from the triangle inequality we obtain

$$\begin{aligned} |J(\hat{\theta}_n; \widehat{\mathbf{b}}_n) - J^*(\theta)| &\le |J(\hat{\theta}_n; \widehat{\mathbf{b}}_n) - J(\hat{\theta}_{n-1}; \widehat{\mathbf{b}}_n)| + |J^*(\hat{\theta}_{n-1}) - J^*(\theta)| \\ &\le \sup_{\mathbf{b}\in\mathbb{B}} |J(\hat{\theta}_n; \mathbf{b}) - J(\hat{\theta}_{n-1}; \mathbf{b})| + |J^*(\hat{\theta}_{n-1}) - J^*(\theta)|. \end{aligned}$$

From (3.13), the first term in the upper bound goes to 0 $\mathsf{P}_\theta -$ a.s.. From the continuity of J^* (recall Lemma 1) and the strong consistency of $\hat{\theta}_n$, the second term in the upper bound goes to 0, which completes the proof.

Remark: The resulting estimator is asymptotically efficient in the sense that if we could employ an oracle item selection method and select each item i such that $J(\theta; \mathbf{b}_i) = J^*(\theta)$, where $J^*(\theta)$ is the maximum Fisher information an item can achieve at the true ability level θ, the asymptotic distribution of the resulting estimator is that of (3.9).

3.3. Discussion of the design

The proposed CAT design based on the nominal response model is similar, but not identical to the CAT design of Chang and Ying (2009) that is based on dichotomous logistic models. We point out the difference in the dichotomous

2PL model in which each item is characterized by the difficulty parameter and the discrimination parameter.

Chang and Ying (2009) select only the difficulty parameter in order to maximize the Fisher information while assuming that the discrimination parameter is bounded. This is not a very realistic assumption, and their asymptotic analysis relies on a closed-form expression for the difficulty parameter.

We assume that all components of the item parameter vector are bounded and establish the consistency of the resulting estimator for an arbitrary item selection strategy. We select all components of the item parameter vector to maximize the Fisher information function, and our analysis does not require a closed-form expression for the item parameters defined by (3.3).

4. CAT with Response Revision

4.1. A novel CAT

In this section we propose and analyze a CAT design in which examinees are allowed to revise their previous answers. We consider multiple-choice items with m categories and assume that the total number of items that will be administered, n, is fixed. Here, after each response, the examinee decides whether to revise the answer to a previous item or to proceed to a new item. Examinees are not allowed to switch back to previously selected answers, and each item can be revised at most $m-2$ times during the test. We need restrict ourselves to items with $m \geq 3$ categories.

Consider the time during the test at which the examinee has completed t responses and let f_t be the number of distinct items that have been administered until this time, the number of revisions until this time being $t - f_t$. For each item $i \in [f_t]$,we denote by g_t^i the number of responses on item i up to this time, so $g_t^i \leq m-1$. If C_t is the set of items that can still be revised at this time, then $C_t = \{i \in [f_t] : g_t^i < m-1\}$, and the decision of the examinee is described as

$$d_t := \begin{cases} 0, & \text{the } t+1^{th} \text{ response corresponds to a new item,} \\ i, & \text{the } t+1^{th} \text{ response is a revision of item } i \in C_t. \end{cases}$$

For each item $i \in [f_t]$ we denote by X_j^i the selected answer at the j^{th} attempt on this item and by $X_{1:j}^i := (X_1^i, \ldots, X_j^i)$ the set of all selected answers in the first j attempts on this item, where $j \in [g_t^i]$. Thus, we write $X_1^i = k$ if category $k \in [m]$ is the first answer to item i, and $X_j^i = k$ if category $k \notin X_{1:j-1}^i$ is selected on the j^{th} attempt on item i, where $2 \leq j \leq g_t^i$ whenever $g_t^i \geq 2$.

Information is generated from the content of the responses, from the decisions of the examinee to revise or not, and the identity of the items that are chosen for revision. Specifically, if $\mathcal{G}_t := \sigma(d_s, s \in [t])$ is the σ-algebra of the first t decisions of the examinee regarding revision, and $\mathcal{F}_t^X := \sigma(X^i_{1:g^i_t}, i \in [f_t])$ the σ-algebra of the first t responses, then $\mathcal{F}_t := \mathcal{G}_t \vee \mathcal{F}_t^X$is the σ-algebra that contains all available information after t responses. The number of items that have been administered until this time, f_t, is $\mathcal{G}_{t-1}$-measurable, since it can be fully recovered by $d_1, \ldots, d_{t-1}$.

For each $i \in [n-1]$, item $i+1$ needs to be selected at the $\{\mathcal{G}_t\}$-stopping time

$$\tau_i := \min\{t \geq 1 : \; f_t = i \quad \text{and} \quad d_t = 0\},$$

i.e., the first time the examinee has answered i distinct items and does not want or is not allowed to revise any more items. Since the total number of items to be administered is n, the test stops at the random time τ_n, which is determined by the test-taker's *revision strategy*, $(d_t)_{1 \leq t \leq \tau_n}$. Our goal is to propose a design that will guarantee the reliable estimation of the test-taker's ability *for any revision strategy*, that is no matter when and which items the test-taker chooses to revise. We postulate only a statistical model for the responses.

4.2. The proposed design

We assume that the first response to each item is governed by the nominal response model. That is, for every item $i \in [n]$,

$$\mathsf{P}_\theta(X^i_1 = k \,|\, \mathbf{b}_i) = p_k(\theta; \mathbf{b}_i), \quad k \in [m], \tag{4.1}$$

where $p_k(\theta; \mathbf{b})$ is the pmf of the nominal response model defined in (2.1)-(2.3), θ is an unknown, scalar parameter that represents the ability of the test-taker and $\mathbf{b}_i := (a_{ik}, c_{ik})_{2 \leq k \leq m}$ is a $\mathbb{B}$-valued vector that characterizes item i. The item parameter $\mathbf{b}_{i+1}$ needs to be selected at time τ_i based on all the available information until this time, and we say that $(\mathbf{b}_i)_{2 \leq i \leq n}$ is an *item selection strategy* if $\mathbf{b}_{i+1}$ is a $\mathbb{B}$-valued, $\mathcal{F}_{\tau_i}$-measurable random vector for every $i \in [n-1]$. The ultimate goal is to obtain an $\mathcal{F}_{\tau_n}$-measurable statistic $\hat{\theta}_n$ that is close to the true ability θ under minimal assumptions on how the examinee chooses to revise.

As before, we suggest that item $i+1$ should be selected such that

$$J(\widehat{\mathbf{b}}_{i+1}) = \max_{\mathbf{b} \in \mathbb{B}} J(\hat{\theta}_{\tau_i}; \mathbf{b}), \tag{4.2}$$

where J is the Fisher information function of the nominal response model given by (2.6), and $\hat{\theta}_{\tau_i}$ is an $\mathcal{F}_{\tau_i}$-measurable statistic. Therefore, the item selection method (4.2) requires an estimate of θ at all times at which items are selected,

$(\tau_i)_{1\leq i\leq n}$.

For the adaptive estimation of θ we use the maximizer of the *partial* likelihood of all observed responses, conditioned on the selected items and *the revision decisions of the examinee.* We describe the proposed estimator for an arbitrary item selection strategy, not necessarily (4.2), and at every time t, not only at $(\tau_i)_{1\leq i\leq n}$. Thus, for any revision strategy $(d_t)_{1\leq t\leq \tau_n}$ and any item selection strategy $(\mathbf{b}_i)_{1\leq i\leq n}$, we suggest updating the ability of the examinee after t responses with the maximizer of

$$L_t(\theta) := \log \mathsf{P}_\theta(X^i_{1:g^i_t}\ ,\ 1\leq i\leq f_t\ \Big|\ \mathcal{G}_t, \mathbf{b}_{1:f_t}). \tag{4.3}$$

We assume that responses coming from different items are conditionally independent, so

$$\mathsf{P}_\theta(X^i_{1:g^i_t}\ ,\ 1\leq i\leq f_t\ \Big|\ \mathcal{G}_t, \mathbf{b}_{1:f_t}) = \prod_{i=1}^{f_t} \mathsf{P}_\theta(X^i_{1:g^i_t}\ |\ \mathcal{G}_t, \mathbf{b}_i). \tag{4.4}$$

We assume that the response on a given item is independent of the revision strategy of the examinee, in the sense that for every item $i\in\{1,\ldots,f_t\}$ we have

$$\begin{aligned}\mathsf{P}_\theta(X^i_{1:g^i_t}\,|\,\mathcal{G}_t, \mathbf{b}_i) &= \mathsf{P}_\theta(X^i_{1:g^i_t}\,|\,g^i_t, \mathbf{b}_i) \qquad (4.5)\\ &= \mathsf{P}_\theta(X^i_1\,|\,\mathbf{b}_i)\cdot \prod_{j=2}^{g^i_t} \mathsf{P}_\theta\left(X^i_j\,|\,X^i_{1:j-1}, \mathbf{b}_i\right).\end{aligned}$$

The second equality follows from the definition of conditional probability and it is understood that the second factor in the right-hand side is equal to 1 whenever $g^i_t = 1$. Each probability $\mathsf{P}_\theta(X^i_1\,|\,\mathbf{b}_i)$ is determined by (4.1), according to which the first answer to each item is governed by the nominal response model. Thus, it remains to specify the contribution of revisions. We assume that the nominal response model also determines revisions, in the sense that

$$\mathsf{P}_\theta\left(X^i_j = k\,|\,X^i_{1:j-1}, \mathbf{b}_i\right) = \frac{p_k(\theta;\mathbf{b}_i)}{\sum_{h\notin X^i_{1:j-1}} p_h(\theta;\mathbf{b}_i)}, \quad k\notin X^i_{1:j-1}. \tag{4.6}$$

Assumptions (4.1), (4.4), (4.5), and (4.6) imply that the conditional log-likelihood function takes the form

$$L_t(\theta) = \sum_{i=1}^{f_t}\Big[\ell\left(\theta;\mathbf{b}_i, X^i_1\right) + 1_{\{g^i_t\geq 2\}} \sum_{j=2}^{g^i_t} \ell(\theta;\mathbf{b}_i, X^i_j|X^i_{1:j-1})\Big], \tag{4.7}$$

where $\ell(\theta;\mathbf{b}_i, X^i_1)$ is defined according to (2.4), and for every $2\leq j\leq g^i_t$ we set

$$\ell(\theta;\mathbf{b}_i, X^i_j\,|\,X^i_{1:j-1}) := \log \mathsf{P}_\theta\left(X^i_j\,|\,X^i_{1:j-1}, \mathbf{b}_i\right).$$

Then, the corresponding score function takes the form

$$S_t(\theta) := \frac{d}{d\theta} L_t(\theta) = \sum_{i=1}^{f_t} \Big[s\left(\theta; \mathbf{b}_i, X_1^i\right) + 1_{\{g_t^i \geq 2\}} \sum_{j=2}^{g_t^i} s\left(\theta; \mathbf{b}_i, X_j^i \,|\, X_{1:j-1}^i\right) \Big], \quad (4.8)$$

where $s(\theta; \mathbf{b}_i, X_1^i)$ is defined according to (2.5), and for every $2 \leq j \leq g_t^i$ we have

$$s(\theta; \mathbf{b}_i, X_j^i \,|\, X_{1:j-1}^i) := \frac{d}{d\theta} \ell\left(\theta; \mathbf{b}_i, X_j^i \mid X_{1:j-1}^i\right). \quad (4.9)$$

The proposed estimate for θ after t responses, $\hat{\theta}_t$, is the root of the score function $S_t(\theta)$. The root exists and is unique for every t that is larger than some random time and a preliminary estimation procedure is needed until this time. We can show that $\hat{\theta}_{\tau_n}$ is the root of $S_{\tau_n}(\theta)$ for all large n with probability 1.

4.3. Discussion of the proposed design

Assumptions (4.1)-(4.4) are analogous to (3.1)-(3.2) in the context of a conventional CAT. The additional modeling assumptions that we impose are (4.5) and (4.6).

The proposed design does not introduce any additional item parameters to the ones used in the conventional CAT of the previous section, and contrary to previous CAT designs in the literature that allow for response revision, the proposed method takes into account all responses of the examinee on a given item during the test, not only the last one. Examinees benefit by revising wrong answers, but have to be cautious with revisions, since every recorded answer during the test contributes to item selection and interim ability estimation.

We do not make any assumptions regarding when and what items the test-taker chooses to revise. Incorporating such information could lead to alternative estimators and item selection methods. But would make the design more vulnerable to model mispecification.

4.4. Asymptotic properties

We assume that (4.1), (4.4), (4.5), and (4.6) hold, and study the asymptotic behavior of the proposed final ability estimator. We establish its strong consistency for any item selection strategy and revision behavior, and its asymptotic normality when the items are selected according to (4.2) and the total number of revisions is small relative to the total number of distinct items.

We show that the conditional score function, $S_t(\theta)$, is a martingale with predictable variation equal to the conditional Fisher information

$$I_t(\theta) := \sum_{i=1}^{f_t} \left[J(\theta; \mathbf{b}_i) + 1_{\{g_t^i \geq 2\}} \sum_{j=2}^{g_t^i} J\left(\theta; \mathbf{b}_i \,|\, X_{1:j-1}^i\right) \right] = \frac{d}{d\theta} S_t(\theta), \tag{4.10}$$

where $J(\theta; \mathbf{b}_i)$ is defined as in (3.5), and for every $2 \leq j \leq g_t^i$ we set

$$J\left(\theta; \mathbf{b}_i \,|\, X_{1:j-1}^i\right) := \mathsf{Var}_\theta \left[s(\theta; \mathbf{b}_i, X_j^i \,|\, X_{1:j-1}^i) \right], \tag{4.11}$$

where $s(\theta; \mathbf{b}_i, X_j^i \,|\, X_{1:j-1}^i)$ is defined in (4.9).

Lemma 5. *For any item selection strategy and any revision strategy,*

(i) $\{S_t(\theta)\}_{t \in \mathbb{N}}$ *is a* $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$*-martingale under* P_θ *with bounded increments, mean zero and predictable variation equal to the conditional Fisher information* (4.10)*, i.e.,*

$$\langle S(\theta) \rangle_t := \sum_{u=1}^{t} \mathsf{E}_\theta \left[(S_u(\theta) - S_{u-1}(\theta))^2 \,|\mathcal{F}_{u-1} \right] = I_t(\theta).$$

(ii) $\{S_{\tau_n}(\theta)\}_{n \in \mathbb{N}}$ *is a* $\{\mathcal{F}_{\tau_n}\}_{n \in \mathbb{N}}$*-martingale with mean 0 and predictable variation* $\{I_{\tau_n}(\theta)\}_{n \in \mathbb{N}}$.

The proof of Lemma 5 is presented in S3 of the supplementary materials.

Lemma 6. *Fix* $\mathbf{b}_i \in \mathbb{B}$, $j \in \{2, \ldots, m-1\}$*, and* $X_{1:j-1}^i$ *and let*

$$p_k(\theta; \mathbf{b}_i | X_{1:j-1}^i) := \mathsf{P}_\theta \left(X_j^i = k \,|\, X_{1:j-1}^i, \mathbf{b}_i \right),$$
$$\bar{a}(\theta; \mathbf{b}_i | X_{1:j-1}^i) := \sum_{k \notin X_{1:j-1}^i} a_{ki} \, p_k(\theta; \mathbf{b}_i | X_{1:j-1}^i).$$

(i) *The conditional score in* (4.9) *and the conditional Fisher information* (4.11) *admit satisfy*

$$s(\theta; \mathbf{b}_i, X_j^i = k \,|\, X_{1:j-1}^i) = a_{ki} - \bar{a}(\theta; \mathbf{b}_i | X_{1:j-1}^i), \quad k \notin X_{1:j-1}^i,$$
$$J\left(\theta; \mathbf{b}_i \,|\, X_{1:j-1}^i\right) = \sum_{k \notin X_{1:j-1}^i} \left(a_{ki} - \bar{a}(\theta; \mathbf{b}_i \,| X_{1:j-1}^i) \right)^2 p_k(\theta; \mathbf{b}_i | X_{1:j-1}^i),$$

and are bounded by a constant that does not depend on θ *or* $\mathbf{b}_i$.

(ii) $\bar{a}(\theta; \mathbf{b}_i | X_{1:j-1}^i) \to a_*(\mathbf{b}_i)$ *as* $\theta \to -\infty$ *and* $\bar{a}(\theta; \mathbf{b}_i | X_{1:j-1}^i) \to a^*(\mathbf{b}_i)$ *as* $\theta \to +\infty$.

The proof of Lemma 6 follows by direct computation.

Theorem 3. For any item selection method and any revision strategy, as $n \to \infty$ we have

$$\hat{\theta}_{\tau_n} \to \theta \quad \text{and} \quad \frac{I_{\tau_n}(\hat{\theta}_{\tau_n})}{I_{\tau_n}(\theta)} \to 1 \quad \mathsf{P}_\theta\text{-a.s.}. \tag{4.12}$$

The proof of Theorem 4.1 is presented in S3 of the supplementary materials.

Theorem 4. If $I_{\tau_n}(\theta)/n$ converges in probability to a positive number, then

$$\sqrt{I_{\tau_n}(\hat{\theta}_{\tau_n})}\,(\hat{\theta}_{\tau_n}-\theta) \to \mathcal{N}(0,1). \tag{4.13}$$

This holds when items are selected according to (4.2), and the number of revisions is much smaller than the number of items in the sense that $\tau_n - n = o_p(n)$, in which case

$$\sqrt{n}(\hat{\theta}_{\tau_n}-\theta) \to \mathcal{N}\left(0,[J^*(\theta)]^{-1}\right). \tag{4.14}$$

Proof. Assume for the moment that as $n \to \infty$

$$\frac{S_{\tau_n}(\theta)}{\sqrt{I_{\tau_n}(\theta)}} \to \mathcal{N}(0,1). \tag{4.15}$$

Since $S_{\tau_n}(\hat{\theta}_{\tau_n}) = 0$ for large enough n with probability 1, with a Taylor expansion around θ we have

$$\begin{aligned} 0 = S_{\tau_n}(\hat{\theta}_{\tau_n}) &= S_{\tau_n}(\theta) + S^{'}_{\tau_n}(\tilde{\theta}_{\tau_n})(\hat{\theta}_{\tau_n}-\theta) \\ &= S_{\tau_n}(\theta) - I_{\tau_n}(\tilde{\theta}_{\tau_n})(\hat{\theta}_{\tau_n}-\theta) \quad \mathsf{P}_\theta - \text{a.s.}, \end{aligned} \tag{4.16}$$

where $\tilde{\theta}_{\tau_n}$ lies between $\hat{\theta}_{\tau_n}$ and θ. From (4.15) and (4.16) we obtain

$$\frac{I_{\tau_n}(\tilde{\theta}_{\tau_n})}{I_{\tau_n}(\theta)}\sqrt{I_{\tau_n}(\theta)}\,(\hat{\theta}_{\tau_n}-\theta) \to \mathcal{N}(0,1).$$

Thus, from (4.12) it follows that the ratio on the left-hand side goes to 1 almost surely, and from Slutsky's theorem we obtain (4.13). Therefore, in order to prove the first part of the theorem, it suffices to show that if $I_{\tau_n}(\theta)/n$ converges in probability to some positive number, then (4.15) holds. We define the martingale-difference array

$$Y_{nt} := \frac{S_t(\theta)-S_{t-1}(\theta)}{\sqrt{n}} 1_{\{t\leq\tau_n\}}, \quad t\in\mathbb{N}, \quad n\in\mathbb{N}.$$

Since $\{S_t(\theta)\}$ is an $\{\mathcal{F}_t\}$-martingale and τ_n an $\{\mathcal{F}_t\}$-stopping time, then $\{t \leq \tau_n\} = \{\tau_n \leq t-1\}^c \in \mathcal{F}_{t-1}$ and we have

$$\mathsf{E}_\theta\left[Y_{nt}|\mathcal{F}_{t-1}\right] = \frac{1_{\{t\leq\tau_n\}}}{\sqrt{n}}\,\mathsf{E}_\theta[S_t(\theta)-S_{t-1}(\theta)\,|\,\mathcal{F}_{t-1}] = 0.$$

The increments of $\{S_t(\theta)\}_{t\in\mathbb{N}}$ are uniformly bounded, which implies that for every $\epsilon > 0$ we have, as $n \to \infty$,

$$\sum_{t=1}^{\infty}\mathsf{E}_\theta\left[Y_{nt}^2\,\mathbb{1}_{\{|Y_{nt}|>\epsilon\}}\right] \to 0. \tag{4.17}$$

Therefore, from the Martingale Central Limit Theorem (see, e.g. Ex. 35.12 in

Billingsley (2008)) and Slutsky's theorem it follows that if $\sum_{t=1}^{\infty} \mathsf{E}_\theta[Y_{nt}^2 \,|\, \mathcal{F}_{t-1}]$ converges in probability to a positive number, then

$$\sqrt{\frac{n}{I_{\tau_n}(\theta)}} \sum_{t=1}^{\infty} Y_{nt} \longrightarrow \mathcal{N}(0,1).$$

But

$$\sum_{t=1}^{\infty} \mathsf{E}_\theta[Y_{nt}^2 \,|\, \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{t=1}^{\tau_n} \mathsf{E}_\theta \left[(S_t(\theta) - S_{t-1}(\theta))^2 \,|\, \mathcal{F}_{t-1} \right] = \frac{I_{\tau_n}(\theta)}{n},$$

$$\sqrt{\frac{n}{I_{\tau_n}(\theta)}} \sum_{t=1}^{\infty} Y_{nt} = \frac{1}{\sqrt{I_{\tau_n}(\theta)}} \sum_{t=1}^{\tau_n} [S_t(\theta) - S_{t-1}(\theta)] = \frac{S_{\tau_n}(\theta)}{\sqrt{I_{\tau_n}(\theta)}},$$

which completes the proof of the first part of the theorem. In order to prove the second part, it suffices to show that

$$\frac{I_{\tau_n}(\theta)}{n} \to J^*(\theta)$$

in probability as $n \to \infty$. From (4.10) it follows that the Fisher information function can be decomposed as

$$I_{\tau_n}(\theta) = \sum_{i=1}^{n} J(\theta; \mathbf{b}_i) + I_{\tau_n}^R(\theta),$$

where $I_{\tau_n}^R(\theta)$ is the part of the information coming from revisions, i.e.,

$$I_{\tau_n}^R(\theta) := \sum_{i=1}^{n} 1_{\{g_{\tau_n}^i \geq 2\}} \sum_{j=2}^{g_{\tau_n}^i} J\left(\theta; \mathbf{b}_i \,|\, X_{1:j-1}^i\right). \tag{4.18}$$

Let $(\widehat{\mathbf{b}}_i)_{2 \leq i \leq n}$ be the information maximizing item selection strategy defined in (4.2). Then, from (3.12) we have

$$\frac{1}{n} \sum_{i=1}^{n} J(\theta, \widehat{\mathbf{b}}_i) \to J^*(\theta) \quad \mathsf{P}_\theta - \text{a.s.},$$

whereas from Lemma 6(ii), for any revision strategy we have

$$\frac{1}{n} I_{\tau_n}^R(\theta) \leq K \frac{\tau_n - n}{n},$$

where K is some constant that does not depend on θ. The upper bound goes to 0 in probability when $\tau_n - n = o_p(n)$, which completes the proof.

5. Simulation Study

In this section we present the results of two simulation studies in which we compared the proposed CAT design that allows for response revision, to which we refer as RCAT, with a conventional CAT that does not allow for response revision, when both are based on the nominal response model (2.1). In the first study we illustrate our asymptotic results, whereas in the second study we compare the two designs in a realistic item pool. Specifically, in both studies items were selected according to the information-maximizing item selection strategies (3.3) and (4.2) for CAT and RCAT, respectively, however in the second study items are selected from a discrete item pool without replacement.

For both studies, when revision is allowed we assumed that at most n_1 items could be revised during the test and that the examinee decided to revise a previous answer after the t^{th} response with probability p_t that satisfies the recursion

$$p_{t+1} = p_t - \frac{0.5}{n_1}, \quad p_1 = 0.5.$$

For n_1, we considered the cases $n_1/n = 0.1, 0.5, 1$. At any given time the examinee decided to revise, we assumed that all previous items that could still be revised were equally likely to be selected. The revised responses were simulated according to the conditional probability model (4.6).

We replicated the two studies for ability levels in the set $\{-3, -2, -1, 0, 1, 2, 3\}$. For each scenario, we computed the root mean square error (RMSE) of the final ability estimator, $\sqrt{\mathsf{E}_\theta[(\hat{\theta}_n - \theta))^2]}$ and $\sqrt{\mathsf{E}_\theta[(\hat{\theta}_{\tau_n} - \theta))^2]}$, for CAT and RCAT, respectively, on the basis of $1,000$ simulation runs (examinees).

5.1. An idealized item pool

In the first study we considered an idealized item pool of items that was simulated based on Passos, Berger and Tan (2007). Each item had $m = 3$ categories, which means that each item could be revised at most once whenever revision was allowed. The parameters of the nominal response model were restricted to $a_2 \in [-0.18, 4.15]$, $a_3 \in [0.17, 3.93]$, $c_2 \in [-8.27, 6.38]$ and $c_3 \in [-7.00, 8.24]$, whereas $a_1 = c_1 = 0$. The test length was $n = 50$ items.

The results are summarized in Table 1. We observe that the RMSE in RCAT is, typically, slightly smaller than that in CAT and slightly larger than the quantity that is suggested by our asymptotic analysis, $(\sqrt{nJ^*(\theta)})^{-1}$. The RMSE in RCAT seems to slightly outperform this benchmark in the case $\theta = -2$ when the number of revisions is large. For an examinee with this ability, we plot in Figure 1 the evolution of the normalized total information $I_t(\theta)/f_t$, as well as the corresponding information from first responses, $\sum_{i=1}^{f_t} J(\theta; \mathbf{b}_i)/f_t$, and revisions, $I_t^R(\theta)/f_t$, where $1 \leq t \leq \tau_n$.

In Figure 2 we compare the approximate 95% confidence intervals that are

Table 1. RMSE in CAT and RCAT.

θ	$(\sqrt{nJ^*(\theta)})^{-1}$	CAT	RCAT		
			Expected Number of Revision		
			4	18	26
-3	0.097	0.104	0.105	0.107	0.100
-2	0.071	0.075	0.073	0.070	0.070
-1	0.068	0.072	0.072	0.072	0.071
0	0.068	0.074	0.072	0.072	0.072
1	0.068	0.077	0.072	0.069	0.070
2	0.068	0.075	0.072	0.070	0.070
3	0.071	0.079	0.076	0.073	0.072

obtained after i distinct items have been answered with a CAT and RCAT, respectively,

$$\hat{\theta}_i \pm 1.96 \cdot (I_i(\hat{\theta}_i))^{-1/2} \quad \text{and} \quad \hat{\theta}_{\tau_i} \pm 1.96 \cdot (I_{\tau_i}(\hat{\theta}_{\tau_i}))^{-1/2}, \quad 1 \leq i \leq n,$$

for an examinee with ability parameter $\theta = -3$. Our asymptotic results guarantee the validity of the final confidence interval ($i = n$) when n is large. The graph indicates that revision improves the estimation of θ.

5.2. A discrete item pool

In the second study, an item pool was constructed based on a random sample of 135 multiple choice items from the Chinese Proficiency Tests (HSK), a large scale international standardized exam for non-native Chinese speakers with more 500,000 test takers annually (Wang et al. (2016b)). The item parameters were calibrated based on the responses of 10,000 examinees. The MULTILOG (Thissen (1991)) was used to calibrate the item parameters of the nominal response model that are described by figure S4 of the supplementary material. Each item had $m = 4$ categories, which means that each item could be revised at most twice when revision was allowed. We considered 3 levels for the test length(n), and 3 levels for the maximum number of items that could be revised (n_1), to which we refer as “small”, “medium” and “large”. Specifically, we considered $n = 20(n_1 = 5, 10, 20)$, $30(n_1 = 5, 15, 30)$ and $40(n_1 = 5, 20, 40)$. The results are documented in Table 2 and show that the positive effect of revisions in the efficiency of the proposed estimator is much more intense than in the case of the idealized item pool, especially when the number of revisions is large. However, due to the discreteness of item pool, the RMSEs were much larger than $(\sqrt{nJ^*(\theta)})^{-1}$.

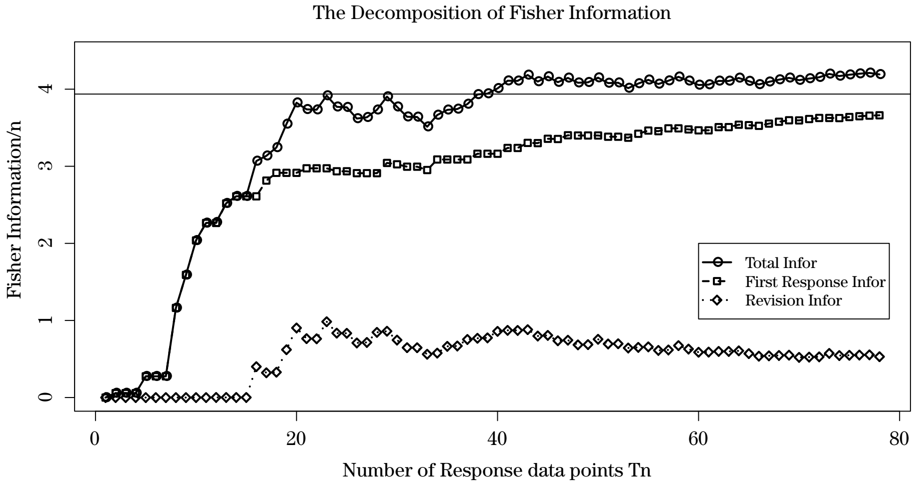

Figure 1. Decomposition of the Fisher information. The solid line represents the evolution of the normalized accumulated Fisher information, $\{I_t(\hat{\theta}_t)/f_t, 1 \leq t \leq \tau_n\}$, in a CAT with response revision. The dashed line with squares (diamonds) represents the corresponding information from first responses (revisions). The horizontal line represents the maximal Fisher information, $J^*(\theta)$. The true ability value is $\theta = -2$.

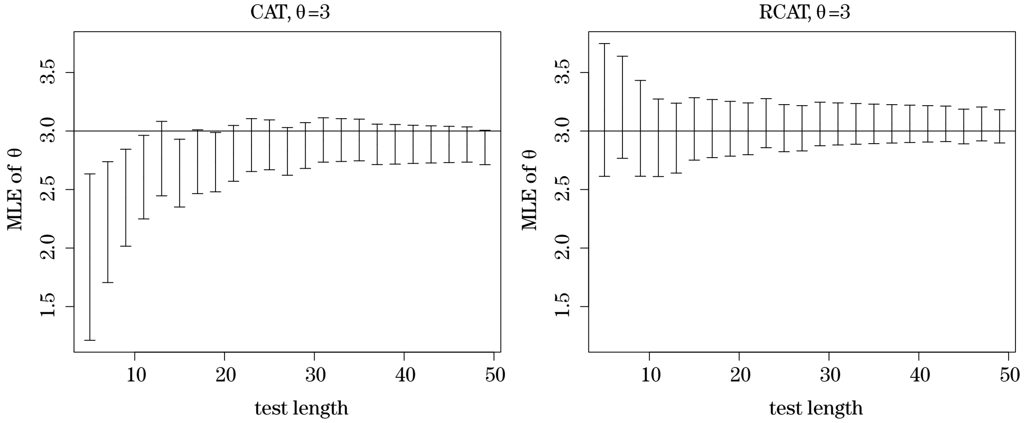

Figure 2. 95% Confidence Intervals. The left-hand side presents 95% confidence intervals, $\hat{\theta}_i \pm 1.96 \cdot (I_i(\hat{\theta}_i))^{-1/2}$, $1 \leq i \leq n$, in a standard CAT. The right-hand side presents the corresponding intervals $\hat{\theta}_{\tau_i} \pm 1.96 \cdot (I_{\tau_i}(\hat{\theta}_{\tau_i}))^{-1/2}$, $1 \leq i \leq n$ in the proposed RCAT design that allows for response revision. In both cases, the true value of θ is -3.

6. Conclusion

An attractive feature of our approach from a practical point of view is that it does not require any additional calibration effort to the one needed by the corresponding conventional CAT that is based on the nominal response model.

Table 2. RMSE of CAT and RCAT in a realistic item pool.

θ			-3	-2	-1	0	1	2	3
$n = 20$	Design	Condition							
		$(\sqrt{nJ^*(\theta)})^{-1}$	0.084	0.059	0.059	0.080	0.081	0.107	0.165
	CAT		0.283	0.230	0.301	0.346	0.338	0.300	0.333
		small	0.264	0.211	0.258	0.342	0.324	0.276	0.315
	RCAT	medium	0.267	0.194	0.278	0.342	0.328	0.276	0.292
		large	0.248	0.181	0.259	0.333	0.320	0.250	0.289
$n = 30$	Design	Condition							
		$(\sqrt{nJ^*(\theta)})^{-1}$	0.068	0.048	0.048	0.065	0.066	0.087	0.135
	CAT		0.256	0.200	0.236	0.303	0.291	0.265	0.314
		small	0.246	0.178	0.227	0.296	0.287	0.253	0.300
	RCAT	medium	0.232	0.181	0.221	0.276	0.278	0.224	0.274
		large	0.222	0.159	0.205	0.283	0.275	0.217	0.267
$n = 40$	Design	Condition							
		$(\sqrt{nJ^*(\theta)})^{-1}$	0.059	0.042	0.042	0.057	0.058	0.075	0.117
	CAT		0.243	0.178	0.213	0.279	0.289	0.260	0.309
		small	0.247	0.173	0.208	0.269	0.277	0.257	0.300
	RCAT	medium	0.209	0.149	0.190	0.260	0.253	0.215	0.271
		large	0.206	0.145	0.186	0.257	0.251	0.202	0.247

Thus, a traditional CAT system based on the nominal response model may easily be modified to allow for response revision. At the same time, we should underline that examinees do not recover the flexibility they enjoy in a paper-pencil test, as all responses on an item during the test, and not only the last one, contribute to the estimation of the ability parameter. We believe that this feature helps protect the resulting ability estimator against certain deceptive test-taking strategies by the examinees, which is an issue we explore in our current research.

Our work opens a number of research directions. First of all, since items are drawn without replacement, this may call for modifications of the item selection strategy, in the spirit of Chang and Ying (1999). More empirical and theoretical work is required in order to understand the effect of different revision behaviors on the proposed methodology. It remains an open problem to develop reliable models for the revision behavior of examinees, which could potentially be incorporated in the ability estimation and item selection algorithms.

Supplementary Materials

The supplementary materials contain the proofs of Lemmas and Theorems in Sections 2, 3, 4, as well as the distribution of the item parameters in the discrete item pool.

Acknowledgment

The authors thank the Chinese Testing International Company for providing the HSK item pool.

References

Bartroff, J., Finkelman, M. and Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika* **73**, 473–486.

Billingsley, P. (2008). *Probability and Measure*. John Wiley & Sons.

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

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

Chang, H.-H. and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement* **20**, 213–229.

Chang, H.-H. and Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement* **23**, 211–222.

Chang, H.-H. and Ying, Z. (2007). Computerized adaptive testing. *The Sage Encyclopedia of Measurement and Statistics* **1**, 170–173.

Chang, H.-H. and Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics* **37**, 1466–1488.

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement* **37**, 259–275.

Han, K. T. (2015). Happy cat: Options to allow test takers to review and change responses in cat. In *In the International Association of Computerized Adaptive Testing.*

Kingsbury, G. (1996). Item review and adaptive testing. In *Annual Meeting of the National Council on Measurement in Education, New York, NY.*

Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *The Annals of Statistics* **7**, 1196–1221.

Lord, F. M. (1971). Robbins-monro procedures for tailored testing. *Educational and Psychological Measurement* **31**, 3–31.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Routledge.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement* **22**, 224–236.

Luecht, R. M. and Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement* **35**, 229–249.

Owen, R. J. (1969). A bayesian approach to tailored testing. *ETS Research Bulletin Series* **1969**, i–24.

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

Passos, V. L., Berger, M. P. and Tan, F. E. (2007). Test design optimization in cat early stage with the nominal response model. *Applied Psychological Measurement* **31**, 213–232.

Rasch, G. (1993). *Probabilistic Models for Some Intelligence and Attainment Tests.* ERIC.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics* **22**, 400–407.

Schmidt, F. L., Urry, V. W. and Gugel, J. F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement* **38**, 265–273.

Sie, H., Finkelman, M. D., Bartroff, J. and Thompson, N. A. (2015). Stochastic curtailment in adaptive mastery testing improving the efficiency of confidence interval–based stopping rules. *Applied Psychological Measurement* **39**, 278–292.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement* **21**, 129–142.

Swanson, L. and Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement* **17**, 151–166.

Thissen, D. (1991). *Multilog User's Guide: Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory.* Scientific Software International.

Veerkamp, W. J. and Berger, M. P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics* **22**, 203–226.

Vispoel, W. P. and Coffman, D. D. (1992). Computerized adaptive testing of music-related skills. *Bulletin of the Council for Research in Music Education* , 29–49.

Vispoel, W. P., Hendrickson, A. B. and Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement* **37**, 21–38.

Vispoel, W. P., Rocklin, T. R., Wang, T. and Bleiler, T. (1999). Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test? *Journal of Educational Measurement* **36**, 141–157.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice* **12**, 15–20.

Wang, S., Lin, H., Chang, H.-H. and Douglas, J. (2016a). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement* **53**, 45–62.

Wang, S., Zheng, Y., Zheng, C., Su, Y. and Li, P. (2016b). An automated test assembly design for a large-scale chinese proficiency test. *Applied Psychological Measurement* **40**, 233–237.

Wise, S. L. (1996). A critical analysis of the arguments for and against item review in computerized adaptive testing. In *Annual Meeting of the National Council on Measurement in Education (NCME)*, vol. 1996.

Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A. and Severance, D. D. (1999). Examinee judgments of changes in item difficulty: Implications for item review in computerized adaptive testing. *Applied Measurement in Education* **12**, 185–198.

Wu, C. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association* **80**, 974–984.

Wu, C. J. (1986). Maximum likelihood recursion and stochastic approximation in sequential designs. *Lecture Notes-Monograph Series* **8**, 298–313.

Ying, Z. and Wu, C. J. (1997). An asymptotic theory of sequential designs based on maximum likelihood recursions. *Statistica Sinica* **7**, 75–91.

Department of Educational Psychology, University of Georgia, Athens, GA 30602, USA.

E-mail: swang44@uga.edu

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: fellouri@illinois.edu

Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: hhchang@illinois.edu

(Received September 2015; accepted November 2016)