Statistics and Its Interface Volume 13 (2020) 317–333

Computerized adaptive test using raw responses for item selection: theoretical results and applications for the up-and-down method

Cheng-Der Fuh, Edward Haksing Ip*, and Shyh-Huei Chen

Modern computerized adaptive testing (CAT) is finding applications that contain more intensive assessments, collected over nontraditional devices such as tablets and smartphones. In this paper, we introduce an CAT algorithm that uses raw responses to adaptively select items and does not require updating the ability estimate at every administration of an item. The proposed algorithm is especially useful in adaptive assessment situations in which updating ability estimate at each administration is either not feasible or too costly to implement. Specifically, an a-stratified multistage up-and-down method is proposed as an approximation to the commonly used recursive maximum likelihood estimate (R-MLE). Using Markov chain tools, we derive theoretical results for the statistical properties of the up-and-down method. We also report empirical studies for the performance of the proposed method. Both simulation experiments and real data analysis are included. Limitations of the method such as reduced statistical efficiency are also discussed. Overall, despite the limitations, our results show that the up-and-down method is a promising alternative to the classical R-MLE and well-suited for some CAT applications such as ecological momentary assessments.

Keywords and phrases: Up-and-down method, Computerized adaptive tests, Markov chains, Recursive maximum likelihood estimate, ACT, PROMIS.

1. INTRODUCTION

Two growing technology-related trends are emerging in the field of educational, psychological, and health assessment. Consider the assessment of depression and anxiety in the study of psychopathology. Currently, most such assessments are administered at a few fixed time points, often times with lengthy instruments. This means that the assessments (1) may not be able to capture the true nature of the disorders, which might be fluid and dynamic, and (2) may introduce potential bias into the assessment as a result of recall bias, respondent burden and fatigue. As an example for (1), a positive response to a self-reported depression item like "I feel blue" (today), could be more related to daily fluctuation in mood rather than attributable to a psychopathologic state of depression. To address these limitations of traditional assessment, there is a growing trend to employ tools such as ecological momentary assessment (EMA; [7, 26]) to either collect data at regular timepoints but with much higher frequencies (e.g., several times a day). Alternatively, adaptive momentary assessment as well as just-in-time adaptive intervention (JITAI) based on cues and other information (such as habitual time schedule) are also becoming commonplace (e.g., [19]). Instead of paper-and-pencil administration, EMA and JITAI often rely on modern information and communication technology (ICT) such as smartphones and tablets to collect realtime or close-to-realtime information.

The second growing trend is the use of computerized adaptive tests (CAT; [15, 16, 22, 29, 25]). In a CAT administration, items are adaptively presented to the respondent. In other words, for each respondent, a customized set of items is excerpted from a larger pool of available items. The items are administered sequentially, and the presented item is typically selected based on a current estimate of the targeted latent trait. As an example of CAT in the study of depression, Gibbons et al. [11] demonstrated that CAT can be successfully applied to a depression inventory with increased precision of depression measurement and reduced respondent burden.

It is natural to envision applications that combine the two technologies - EMA/JITAI and CAT (e.g., see [27]). Because of the high frequency of assessing respondents, EMA/JITAI augmented with CAT could be a powerful tool for delivering assessment via smartphone and at the same time minimizing respondent burden. Continuing to use depression as an example, recent uses of smartphone-enabled JITAI for depression include the EU-sponsored ICT4Depression project [28]. Examples of mobile platforms for CAT are provided in [10, 12]. As the field of educational and psychological assessment is gravitating toward using mobile devices as data collection and assessment tools, there is clear and pressing need for enhancing the methodology that underlies CAT for more efficient and sustainable deployment on such devices.

In this paper, we propose a CAT algorithm that is based on a simple decision rule for selecting items. The decision is based upon the raw response and does not necessitate

[*] Corresponding author.

the computationally demanding procedure of computing the ability estimate. The proposed CAT algorithm is especially suitable for CAT designs such as EMA/JITAI deployed on a mobile device. The algorithm is based on the so-called up-and-down method, originally described in [8], and recently applied in settings such as dose-design [21]. In psychometric assessment, the idea of not using intermediate ability estimation to select items can be traced back to intelligence testing (Binet and Simon [1]), and is related to the multi-stage test approach proposed by Lord [17]. Dodd et al. [9] suggested a CAT algorithm that does not update ability at each iteration when responses are all correct or all incorrect. Briefly, the idea underlying the up-and-down method for CAT is to use the raw response to a current item to decide whether an easier or a more difficult item (assuming the items are all first linearly ordered) should be presented next. In other words, if the current response is positive (e.g., correct in a cognitive test) then a more difficult item will be presented, and vice versa. The up-and-down algorithm is easy to understand, simple to explain, and convenient to implement. Surprisingly, it can be proved that this seemingly unsophisticated algorithm (at first-glance) possesses many desirable statistical properties. In this paper we discuss the behavior of the CAT estimate derived from the up-and-down method. Additionally we present empirical evidence that the proposed method performs reasonably well when compared to traditional and more computationally intensive CAT algorithms.

This article makes several contributions to the psychometric literature. First, the proposed method opens new avenue for the study of a family of CAT algorithms that is based on raw responses and does not require realtime updating of trait estimates at each and every administration of an item. We call such algorithms raw response driven CAT (RRD-CAT). This type of RRD-CAT algorithms has the potential for widespread applications of CAT on mobile devices or client-server implementations in which technical, logistic, or intellectual property considerations present challenges for realtime updates or require elaborate communication between the mobile device and the remote server. The saving from not being required to compute an ability estimate has important implications especially for emerging adaptive designs. We will elaborate on this point and discuss a few examples of potential applications in the Discussion section. The paper also contributes to the theory of CAT. Specifically, we study the behavior of the up-and-down estimates from a new perspective of Markov chain. Because of the sequential design of CAT, responses are not conditionally independent given the trait estimate. Asymptotic analysis could therefore be challenging and indeed theoretical studies of the statistical behavior of CAT algorithms are few and far between. Chang and Ying [4] provides asymptotic analysis of the standard procedure of selecting items based on maximizing the Fisher information [18, pg. 151–153]. As far as we know, our paper represents a first attempt to understand the behavior of a raw-response driven CAT algorithm from a Markov chain perspective. The Markov chain related tools developed in this paper have the potential to be translated and applied to other similar CAT designs. Again, we discuss such opportunities in the Discussion section.

The paper is organized as follows: First we provide background on both standard CAT algorithm based on recursive maximum likelihood and the up-and-down method. We provide the theoretical results regarding the proposed up-and-down CAT and report results from two simulation experiments. We then describe two real data analyses to investigate the real world behaviors of the up-and-down CAT implementation as compared to the standard approach. The first data analysis uses items from the ACT math test, and the second data analysis uses depression items from the Patient Reported Outcome Measurement Information System (PROMIS). Finally we provide a discussion and a brief conclusion.

2. BACKGROUND

2.1 Fisher information based CAT and R-MLE

Consider the two-parameter logistic (2PL) item response model, which has the form:

$$p(\theta) = P(Y = 1|\theta) = \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}, \tag{1}$$

where $Y = 0, 1$ is the binary response, θ is the latent trait, and a and b are the discrimination and difficulty item parameters, respectively. The 1PL IRT model can be obtained by setting $a = 1$ for all items. The 2PL model appears to be more commonly used in the CAT literature because of its flexibility and in many cases, better fit to response data. In this paper, we report both 2PL and 1PL results.

At a given iteration in CAT, the standard approach is to select the item with the maximum Fisher item information as the next presented item [18], computed at the examinee's current estimated ability level under the 2-PL model. This method is also known as recursive maximum likelihood estimation (R-MLE). R-MLE selects an item to maximize precision in estimating an examinee's θ. Thus R-MLE requires updating θ after every administration of an item and the computation of the maximum Fisher information. When b is close to θ, items with high a values have high information. Consequently, items with high a values tend to be more frequently exposed than items with lower a values. Overexposure of some items is a potential problem in R-MLE for high-stake testing such as licensure exams [3, 14].

2.2 Up-and-down method: one-parameter logistic IRT

It is easier to illustrate the key features of the up-and-down method using the one-parameter logistic IRT (i.e., $a = 1$ in Equation (1) for all items):

0. Order the items by their difficulty parameter b from smallest to largest. Without loss of generality, we assume that b on consecutive items in the ordered list differ by one unit Δ.
1. Start the first item that has the median value of b.
2. The difficulty level of the next test item, b, is increased by one unit when the response is 1. Otherwise, it is decreased by one unit.
3. Repeat Step 2 until M items have been administered.

The length of the test M is determined a priori. When the test is finished, the method of maximum likelihood is applied to the collection of M responses to estimate the examinee's ability parameter θ.

Consider an examinee with ability level θ, the data set generated by the up-and-down procedure produces the sequence $(\mathbf{x}, \mathbf{y}) = \{(x_0, y_0), \ldots, (x_n, y_n)\}$, where x_t represents the difficulty level for the tth selected item and y_t is the corresponding response value. The sequence $\{X_t, t = 0, 1, \cdots, n\}$ forms a Markov chain (MC) on a state space $\{b_j = x_0 + j\Delta, j \in Z\}$, where Z is the set of integers. A key element in understanding the behavior of the up-and-down method is the Markov kernel, in this case the transition probability matrix - i.e., how the state x_t transitions into the next state x_{t+1}.

Observe that

$$\begin{aligned} & P(X_{i+1} = X_i + \Delta | (X_k, Y_k), 0 \leq k \leq i) \\ = \ & P(Y_i = 1 | X_i) = e^{(\theta - X_i)} / [1 + e^{(\theta - X_i)}], \\ & P(X_{i+1} = X_i - \Delta | (X_k, Y_k), 0 \leq k \leq i) \\ = \ & P(Y_i = 0 | X_i) = 1/[1 + e^{(\theta - X_i)}]. \end{aligned}$$

It follows that the transition probability matrix $P = (p_{x_1, x_2})$ is such that $p_{x_1, x_1+\Delta} + p_{x_1, x_1-\Delta} = 1$. Specifically,

$$\begin{aligned} p_{x_1, x_2} &= P\{X_2 = x_2 | X_1 = x_1\} \\ &= \begin{cases} e^{(\theta - x_1)}/[1 + e^{(\theta - x_1)}], & x_2 = x_1 + \Delta, \\ 1/[1 + e^{(\theta - x_1)}], & x_2 = x_1 - \Delta. \end{cases} \end{aligned} \tag{2}$$

For simplicity in our illustration, we assume that $X_0 = 0$, $\Delta = 1$, and $b_i = i$. Figure 1 graphically depicts the dynamic of the Markov chain using several values of X. An arrow represents the direction of a transition. Table 1 shows subsets of the the transition probability matrices for two different levels of $\theta = 0$ (less able student) and $\theta = 2$ (able student). Only $P(X_2 = 1 | X_1)$ is shown as $P(X_2 = 0 | X_1)$ is simply the one-complement of the first quantity. Here X_2 and X_1 can be considered as the difficulty level presented at two consecutive administrations.

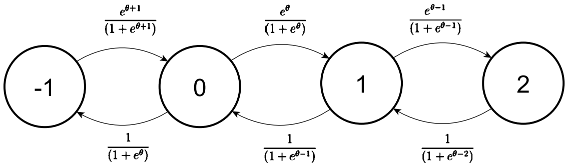
$\frac{e^{\theta+1}}{(1+e^{\theta+1})}$
$\frac{e^{\theta}}{(1+e^{\theta})}$
$\frac{e^{\theta-1}}{(1+e^{\theta-1})}$
-1
0
1
2
$\frac{1}{(1+e^{\theta})}$
$\frac{1}{(1+e^{\theta-1})}$
$\frac{1}{(1+e^{\theta-2})}$

Figure 1. An illustration of the Markov chain for a subset of item difficulty level X $(-1, 0, 1, 2)$ in up-and-down CAT.

Table 1. Illustration of transition matrix for sequence of presented item difficulty for two values of θ

	$\theta = 0$	$\theta = 2$
X_1	$P(X_2 = X_1 + 1 \| X_1)$	$P(X_2 = X_1 + 1 \| X_1)$
0	0.50	0.88
1	0.27	0.73
2	0.12	0.50
3	0.05	0.27

Table 1 shows that for a more able student ($\theta = 2$) the likelihood of getting a more difficult question (by correctly responding to the current question) when the current question is an easy one ($X_1 = 0$) is high (0.88). However when the current question is already difficult ($X_1 = 3$) then it is unlikely he will get a more difficult question ($p = 0.27$). The chance is even lower ($p = 0.05$) for the less able student. The transition probability matrix encapsulates the dynamic of which the level of item difficulty is driven by the true ability, and plays a key role in the proof for convergence properties of the up-and-down method. More background for MC is provided later in a separate subsection.

2.3 Up-and-down method: two-parameter logistic IRT

We follow the a-stratified multistage computerized adaptive testing [3], which was designed for the R-MLE. The idea is to partition the item bank into different groups that have different levels of a parameter and then apply the one-parameter logistic IRT CAT within each partition by treating the a parameter as fixed within the specific partition. The a-stratified multistage CAT for R-MLE can be represented as follows:

1. Partition the item bank into K levels according to the a-parameter values of items. The first item stratum contains items with smallest a's, the next stratum contains items with second smallest a's, etc.
2. Accordingly, partition the test into K stages.
3. Start with stage $k = 1$, select n_k items based on the matching of item difficulty (b-matching) parameter b with the updated estimator $\hat{\theta}_n$, stop after n_k items have been administered. $\hat{\theta}_n$ is the MLE of θ [3]. (p. 4 Section 2.4)
4. Repeat Step 3 for $k = 2, \ldots, K$.

Note that in step 3, depending on the selection method, the selection of the n_k in R-MLE items may take different forms. For example, one can update θ within the stratum for each item and match b that is closest to the updated

θ using the maximum information criterion. Alternatively, one can use a simpler scheme such as minimizing the cumulative difference between the observed response and the expected probability of response with respect to θ [29, 13]. For the up-and-down method, we replace the third step with the procedure described in 1PL IRT model by treating the a parameter as fixed. Hereafter we also assume that the number of items selected from each partition n_k are all equal to M. The total number of administered items is therefore MK. The up-and-down method also requires an additional step 5 that computes the MLE estimate after all MK responses have been collected. Unlike standard CAT using R-MLE, the up-and-down method does not update $\hat{\theta}$ after every response; the maximum likelihood estimate is calculated once after all responses have been collected.

The a-stratified up-and-down procedure inherits the advantages of stratification scheme in a-stratified multistage CAT in terms of item exposure. Both decrease exposure rates of high a items and increase exposure rates of low a items [3]. In other words, the up-and-down scheme for the 2PL model is also able to balance the exposure of items across different values of the a parameter.

2.4 Markov chain (MC)

Recall that a MC is a sequence of random variables $\{X_0, X_1, \cdots, X_n\}$ characterized by (1) the state space S, which is a finite or countable set (here we assume S has N finite elements) (2) an initial distribution π_0 of the states, and (3) the transition probability matrix from a state at time $(t-1)$ to a state at time t: $p_{x_t,x_{t-1}} = P(X_t = x_t | X_{t-1} = x_{t-1})$, where x_t, x_{t-1} represent states in S. To describe the Markov property with a pithy phrase: an MC satisfies the assumption that "conditional on the present, the future does not depend on the past". Hence the dynamic of a MC is determined by the initial distribution and the transition probability matrix. The MC is called homogeneous if the transition probability matrix does not change over time. Hence for a homogeneous MC one can use the first two time points to label the transition probability matrix as p_{x_1,x_2}. As we have derived above, in the up-and-down MC (of difficulty level X), the transition matrix p_{x_1,x_2} is homogeneous over time (sequence of items) for each individual; however, the matrix is a function of θ, which varies across respondents.

The behavior of a MC is often described in terms of several attributes. A MC is irreducible if any given state is accessible to any other state in S; is aperiodic if the period is 1 for all states; and is recurrent if for any state, the probability of returning to the same state (recurrence) is 1. Positive recurrent refers to a recurrent chain that has finite expected recurrrence time. Note that aperiodicity and positive recurrency can be defined at the state level. When a state in a MC possesses the properties of being irreducible, aperiodic and positive recurrent, it is said to be ergodic. Ergodicity of all states of a MC implies that the MC is ergodic, or in a sense "well behaved." For example, an ergodic MC always has a unique stationary distribution, which means that the distribution of the states remains unchanged with transition. Mathematically, it means that the stationary distribution π is invariant: $\pi = \pi\mathbf{P}$ where P is the transition probability matrix. The tool that we use to prove convergence and explore asymptotic behavior of the up-and-down method relies on showing that the MC produced by the sequence of the difficulty level of the selected items is ergodic for any θ, implying a unique stationary distribution exists.

Because observations from the MC do not form independent and identically distributed (i.i.d.) samples, traditional methods for proving large-sample properties do not immediately apply for studying the behavior of the observations generated by the up-and-down method. In this paper an embedded renewal process within the MC, called a regenerative process, was used to explore the asymptotic behavior of the up-and-down estimate. The key idea of the regenerative process is that for a recurrent MC, there exists a recurrent state Δ_r which is visited infinitely often (i.o.). For a fixed state Ω, the cycles $\{X_j; j = T_\Omega^{(n)}, \cdots, T_\Omega^{(n+1)} - 1\ \}$ are i.i.d. for $n = 1, 2, \cdots$, where $T_\Omega^{(n)}$ is the time of the n-th return to Ω. For example, if time for the n-th return is $T_\Omega^{(n)} = 100$, and time for the $(n+1)$-th return is $T_\Omega^{(n+1)} = 121$, then $\{X_j; j = 100, \cdots, 120\}$ form a cycle in the i.i.d. sample. The method using the regenerative process of i.i.d. samples embedded within the MC greatly facilitates the proof of the asymptotic behavior of the MC.

3. LIKELIHOOD FUNCTION FOR UP-AND-DOWN

Let X_i be the test level and Y_i be the corresponding response. The contribution of likelihood function for the 2-PL IRT during the kth stage is given by

$$\text{(3)} \qquad \begin{aligned} & L_{n_k}(\theta) \\ = \;& C\prod_{i=0}^{n_k} f(Y_i|X_i) = f(X_0)\prod_{i=0}^{n_k}[p_i(\theta)]^{Y_i}[1-p_i(\theta)]^{1-Y_i}, \end{aligned}$$

where C is determined by the design for choosing the difficulty level of the first test item, and

$$p_i(\theta) = P(Y_i = 1|\theta) = \frac{e^{a_k(\theta-b_i)}}{1+e^{a_k(\theta-b_i)}},$$

where a_k represents the (constant) value of a designated to the kth stratum. Here $X_i = b_i$ and C is independent of θ. Because for each stratum in the kth stage of a-stratified multistage computerized adaptive testing the a values of the items are considered uniform, in the following discussion we treat a as a fixed value $(= 1)$ in each stage. For simplicity, the difficulty parameter b of the first test items is set to be 0 hereafter. In this case, $C = 1$.

In contrast to R-MLE, for the up-and-down method, estimates of θ is not needed in the selection process of test items. We only need to calculate an estimate of θ at the end of the test. Here we denote the maximum likelihood estimate derived from the up-and-down procedure based on the observations $x_0, y_0, \cdots, x_n, y_n$ by $\hat{\theta}_n$ (i.e., $\hat{\theta}_n$ is the maximizer of the likelihood function $L_n(\theta) = \prod_{k=1}^{K} L_{n_k}(\theta)$ in Eq. (3).

There is no closed-form solution for $\hat{\theta}_n$ in general; therefore, numerical algorithms, such as Newton-like method specifically BFGS [20], are needed to solve for $\hat{\theta}_n$. Since $L_n(\theta)$ is a concave function of θ, the initial value does not impact convergence of BFGS.

Although the likelihood function given in (3) is identical to the likelihood function of a non-adaptive test (in which $X_0, X_1, \ldots, X_n$ are determined before the administration of the test), Y_i in (3) now depends on $Y_{i-1}, \ldots, Y_1$ in the up-and-down method. As a result, asymptotic analysis using standard likelihood methods does not apply to the maximizer of $\prod_{k=1}^{K} L_{n_k}(\theta)$. We need to study the Markovian structure imposed by the up-and-down method to understand the asymptotic behaviour of $\hat{\theta}_n$.

Continuing from Eq. (2), denote the transition probability matrix by $f(X_1, X_2; \theta)$, and define $g(X_1, X_2; \theta) = \log f(X_1, X_2; \theta)$. Since for given observations x_k, $k = 1, \cdots, n$,

$$\frac{1}{n}\sum_{k=1}^{n} \frac{\partial^2}{\partial\theta^2} g(X_{k-1}, X_k; \theta) = \frac{1}{n}\sum_{k=1}^{n} \frac{-e^{(\theta - x_k)}}{[1 + e^{(\theta - x_k)}]^2} < 0,$$

for any $\theta, L_n(\theta)$ is a concave function, the maximum likelihood estimate is the unique root of the score equation

$$\frac{1}{n}\frac{\partial}{\partial\theta} L_n(\theta) = \frac{1}{n}\sum_{k=1}^{n} \frac{\partial}{\partial\theta} g(x_{k-1}, x_k; \theta) = 0. \tag{4}$$

4. MAIN RESULT REGARDING BEHAVIOR OF UP-AND-DOWN ESTIMATE

The main result regarding the asymptotic behavior of the up-and-down estimate follows the following step-by-step argument.

- The score equation in Eq. (4) has a root near θ_0. This can be established by showing that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \frac{\partial}{\partial\theta} g(X_{t-1}, X_t; \theta)\Big|_{\theta=\theta_0} \to 0 \quad \text{in probability.} \tag{5}$$

- The slope of the score equation is negative. We will show that

$$\lim_{t\to\infty} \frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2}{\partial\theta^2} g(X_{t-1}, X_t; \theta)\Big|_{\theta=\theta_0} \to -I \quad \text{in probability,} \tag{6}$$

where the information matrix I is given by

$$\begin{aligned} I &= E_{\theta_0}\left[\frac{\partial}{\partial\theta} g(X_1, X_2; \theta)\Big|_{\theta=\theta_0}\right]^2 \\ &= -E_{\theta_0}\left\{\frac{\partial^2}{\partial\theta^2} g(X_1, X_2; \theta_0)\right\} < \infty. \end{aligned} \tag{7}$$

- The quantity $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean 0 and asymptotic variance I^{-1}.

The idea of proof for these results requires first approximating the score function by an additive function of the MC $\{X_n, n \geq 0\}$. To prove that (5) and (6) hold, we need the law of large numbers. To prove asymptotic normality, we need the Central Limit Theorem for an additive function of the MC. In these proofs, we use the regenerative process to represent the additive function of the MC as sum of i.i.d. random variables. Wald's equations for MC will be applied to deduce the moment conditions for each time epoch t, $t = 1, \cdots, n$. We are now ready to present the main result for the up-and-down estimate.

To avoid singularity of the Fisher information in the two-parameter logistic model, we assume that there exist $0 < C_1 < C_2 < \infty$ such that $a \in [C_1, C_2]$.

Theorem 4.1. *Consider the two-parameter logistic model (1). Let $\{X_n, n \geq 0\}$ be a Markov chain on a countable state space $S = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$, with transition probability p_{ij} defined as (2) for $i, j \in S$. Then*

1) *$\hat{\theta}_n = \hat{\theta}(x_0, \cdots x_n)$ converges in probability to the true ability θ_0,*
2) *$\sqrt{n}(\hat{\theta}_n - \theta_0) \longrightarrow N(0, I^{-1})$ in distribution, where $I = E_{\theta_0}\left[\frac{\partial}{\partial\theta} \log f(X_1, X_2; \theta_0)\right]^2$ is the Fisher information, E_{θ_0} denotes the expectation E_π under the true parameter θ_0, and $f(X_1, X_2; \theta)$ denotes the transition probability under parameter θ of the Markov chain (2).*

The proof of Theorem 4.1 is given in the appendix.

5. EMPIRICAL STUDIES

5.1 Simulation experiment 1

To evaluate the performance of the up-and-down method, we compare it to R-MLE within the context of the 2PL model, i.e., using the a-stratified multistage computerized adaptive testing. Two versions of R-MLE were implemented in simulation experiment 1. The first version, which we denote by R-MLE(A), selected $n_k = M = L/K$ items, where L is the total length of the test, from each of the K strata and updated the θ estimate within stratum to match difficulty of the next item. The second version, R-MLE(B), which was based on selecting n_k items (L/K) for each stratum, started with an initial θ, matched items within each stratum for this value of θ and only updated θ after the round of K items have been administered. As described above, in a-stratified CAT, one often encountered problem is the overexposure of some items especially those with high a parameters- i.e., these items are too frequently used when the algorithm starts to hone in the true ability value. Especially when a and b are correlated, for a fixed length test, it means that there may not be sufficient items within some strata to choose from. For example, when ability is high, when the algorithm gets to high a stratum, items with high b will quickly be exhausted. The R-MLE (B) version was designed to allow for some flexibility in item exposure and improve the more even use of items across all a-stratified groups.

For up-and-down, because no updating on ability was used during item administration, the value of θ is restarted at default (value of 0) every time when the item selection process moved to a new stratum.

The metric for evaluation included both bias and root mean square error (RMSE), which are defined as follows:

$$\text{Bias} = \frac{1}{Q}\sum_{i=1}^{Q}(\hat{\theta}_{ni} - \theta_0), \quad \text{RMSE} = \left(\frac{1}{Q}\sum_{i=1}^{Q}(\hat{\theta}_{ni} - \theta_0)^2\right)^{\frac{1}{2}},$$

where $\hat{\theta}_{ni}$, θ_0 are respectively the estimated value and either the true or reference value, and Q is the total number of responses used for the calculation. The simulation experiments were implemented through R for IRT/CAT data generation, and for parameter and ability estimation on an Intel i7-6700 (16 GB) PC. Particularly, the R package ltm [24] was adopted for IRT estimation.

Data were generated from 3,000 individuals by first sampling 300 θ values from $N(0, 1)$. Then, conditional on θ, response data from a 2PL model were independently generated with each θ value replicated 10 times. Two levels were specified for the number of partitions ($K = 5, 10$), and 3 levels of test length ($n = 20, 30, 40$) were specified. There is a tradeoff between making K too large (implying fewer items per stratum) and K too small (implying fewer points of selection). We followed the literature in the choice of K. The determination of K for data analysis can be found on page 214, Chang and Ying [4].

For the 2PL model, we used two conditions for the correlation between the a and b parameters. Lord and Wingersky (1984) reported that a and b parameter estimates often are positively correlated. This phenomenon is being confirmed from a retired item bank of a GRE quantitative test of 360 items in [5, 4]. For the first set of conditions, the correlation between a and b was -0.56, which was the observed correlation for the real educational test data set used (described later). For the second set, we randomly selected a parameters from the same educational test item parameter set, and then given a, generated $d \sim U(-1, 1)$ and assigned b the value d/a. Correlation was -0.01 for this data set. Eventually, we created two levels for item parameters - correlated a and b and uncorrelated a and b, for the simulation study.

The number of items in the item bank was specified at $L = 150$ for both data sets. In summary, this simulation experiment contains a total of 2 (partition) $\times 3$ (test length) $\times 2$ (parameter setting) $\times 3$ (method)= 36 conditions.

Figure 2 shows the trajectories of the ability estimates for 5 individuals with $\theta = -2, -1, 0, 1, 2$, averaged over 10 replication, with $K = 5$ and test length $n = 40$ for the three methods. They all started at an initial value of $\theta = 0$. This small sample is representative for the entire sample in the sense that all three methods tend to converge to the approximately same value. The values from the three methods begin to become quite close after 20 items.

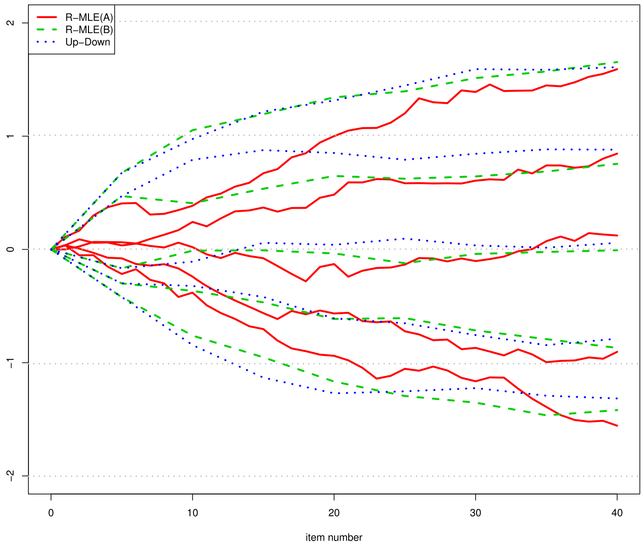

Figure 2. Paths of ability estimates from 5 individuals with true $\theta = -2, -1, 0, 1, 2$, averaged over 10 replications, with $K = 5$, and test length $n = 40$.

The simulation results are shown in the first three rows in both Tables 2 and 3. Two sets of biases and RMSEs are reported. The first set, denoted by (θ), used the true generative θ as reference, whereas the second set, denoted by ($\bar{\theta}_{150}$), used the estimate from all 150 items as the "true" reference value. The R-MLE(A) and (B) methods, especially (A) which had more frequent updating of θ, generally perform better than the up-and-down method. The difference, however, is quite small for $n = 20$ (e.g., when $K = 5$ for the correlated item parameters a and b, the biases across the three methods of R-MLE(A), R-MLE(B), and up-and-down are 0.027, -0.01, and -0.026). The differences are deemed minimal for $n = 30, 40$.

Computationally, R-MLE (A) is the most intensive scheme because it requires updating θ after each item administration. To illustrate computational overhead, we recorded CPU times for the three methods with the following setting: item bank $L = 150$, strata $K = 5$; test length $n = 30$ and number of individuals $N = 300$. The CPU times for R-MLE (A)(30 scorings), and R-MLE (B) (6 scorings),

Table 2. Biases and RMSEs of simulation experiment 1 with $K = 10$

		correlated a and b			uncorrelated a and b		
		$n = 20$	$n = 30$	$n = 40$	$n = 20$	$n = 30$	$n = 40$
R-MLE(A)	Bias (θ)	−0.004	−0.005	−0.012	−0.017	−0.007	−0.011
	RMSE (θ)	0.580	0.486	0.429	0.574	0.493	0.441
	Bias ($\bar{\theta}_{150}$)	−0.002	−0.003	−0.010	−0.011	−0.001	−0.005
	RMSE ($\bar{\theta}_{150}$)	0.516	0.414	0.347	0.502	0.406	0.348
R-MLE(B)	Bias (θ)	−0.017	−0.014	−0.003	−0.024	−0.014	−0.013
	RMSE (θ)	0.571	0.484	0.435	0.568	0.492	0.440
	Bias ($\bar{\theta}_{150}$)	−0.015	−0.011	−0.001	−0.018	−0.008	−0.007
	RMSE ($\bar{\theta}_{150}$)	0.505	0.413	0.350	0.500	0.409	0.347
up-and-down	Bias (θ)	−0.017	−0.013	−0.006	−0.011	−0.014	−0.011
	RMSE (θ)	0.576	0.508	0.462	0.585	0.506	0.450
	Bias($\bar{\theta}_{150}$)	−0.014	−0.011	−0.003	−0.005	−0.008	−0.005
	RMSE($\bar{\theta}_{150}$)	0.511	0.430	0.378	0.518	0.429	0.363
up-and-down (median)	Bias (θ)	−0.021	−0.020	−0.008			
	RMSE (θ)	0.592	0.511	0.455			
	Bias($\bar{\theta}_{150}$)	−0.019	−0.018	−0.006			
	RMSE($\bar{\theta}_{150}$)	0.524	0.434	0.372			

Table 3. Biases and RMSEs of simulation experiment 1 with $K = 5$

		correlated a and b			uncorrelated a and b		
		$n = 20$	$n = 30$	$n = 40$	$n = 20$	$n = 30$	$n = 40$
R-MLE(A)	Bias (θ)	−0.027	−0.012	−0.011	−0.014	−0.015	−0.015
	RMSE (θ)	0.577	0.484	0.441	0.572	0.492	0.438
	Bias ($\bar{\theta}_{150}$)	−0.024	−0.009	−0.008	−0.008	−0.009	−0.009
	RMSE ($\bar{\theta}_{150}$)	0.511	0.411	0.355	0.507	0.412	0.348
R-MLE(B)	Bias (θ)	−0.010	−0.002	0.000	−0.017	−0.009	−0.009
	RMSE (θ)	0.572	0.500	0.438	0.573	0.495	0.439
	Bias ($\bar{\theta}_{150}$)	−0.007	0.000	0.003	−0.011	−0.003	−0.003
	RMSE ($\bar{\theta}_{150}$)	0.507	0.424	0.354	0.509	0.420	0.352
up-and-down	Bias (θ)	−0.026	−0.014	−0.007	−0.019	−0.009	−0.006
	RMSE (θ)	0.578	0.502	0.452	0.583	0.494	0.444
	Bias ($\bar{\theta}_{150}$)	−0.024	−0.012	−0.004	−0.013	−0.003	0.000
	RMSE ($\bar{\theta}_{150}$)	0.504	0.421	0.363	0.512	0.415	0.356
up-and-down (median)	Bias (θ)	−0.021	−0.011	−0.008			
	RMSE (θ)	0.565	0.496	0.444			
	Bias($\bar{\theta}_{150}$)	−0.018	−0.008	−0.005			
	RMSE($\bar{\theta}_{150}$)	0.497	0.419	0.357			

and up-and-down (1 scoring) were respectively 481.5s, 15.7s, and 2.7s.

5.2 Simulation experiment 2

The purpose of this simulation experiment is to examine the efficiency of the up-and-down method. Two sets of experiments were respectively designed to evaluate (1) the performance of the variance estimate for ability, and (2) coverage probability and length of confidence interval of the true ability value. For the investigation of these second-order properties associated with the ability estimate, a large number of replications were needed for each individual. Accordingly the simulation setting was simplified. Because the discrimination parameter a is fixed within a stratum in the a-stratified scheme, we used $a = 1$ for data generation, and step size for b was set at $\Delta = .1$. The up-and-down method always started at $\theta = 0$ at each stratum. Thus it was expected that the efficiency loss would be larger for individuals with θ values that were farther away from 0. Due to symmetry, we only studied the following three variations of non-negative values of ability: $\theta = 0, 1, 2$. For (1), we used 1,000 replications for each individual and varied test length from at the levels of $10, 20, \cdots, 50, 100, 150, 200$, and reported both empirical variances (across the 1,000 replications) and asymptotic variances. For (2), we used $n = 50, 100, 150$ and a larger number of replications of 10,000 for each individual.

Table 4. Empirical variance (VAR) and asymptotic variance (ASY VAR) of R-MLE and up-and-down methods for $\theta = 0, 1, 2$

n	Methods	VAR ($\theta = 0$)	VAR ($\theta = 1$)	VAR ($\theta = 2$)	ASY VAR
10	R-MLE	0.473	0.464	0.523	
	up-and-down	0.550	1.335	11.82	0.410
20	R-MLE	0.212	0.224	0.242	
	up-and-down	0.206	0.241	0.689	0.205
30	R-MLE	0.146	0.145	0.164	
	up-and-down	0.139	0.150	0.195	0.137
40	R-MLE	0.113	0.104	0.107	
	up-and-down	0.115	0.110	0.128	0.103
50	R-MLE	0.085	0.084	0.089	
	up-and-down	0.079	0.089	0.099	0.082
100	R-MLE	0.042	0.044	0.040	
	up-and-down	0.041	0.040	0.045	0.041
150	R-MLE	0.027	0.027	0.028	
	up-and-down	0.026	0.027	0.029	0.027
200	R-MLE	0.020	0.021	0.020	
	up-and-down	0.022	0.022	0.022	0.021

The up-and-down method was directly compared to R-MLE (A) (called R-MLE for this simulation experiment).

By Theorem 4.1, $\sqrt{n}(\hat{\theta}_n - \theta_0) \longrightarrow N(0, I^{-1})$, where I is the Fisher information determined by the Markov chain with transition probability (2), with $\hat{\theta}_n$ being the up-and-down ability estimate. The asymptotic variance of the R-MLE and the up-and-down estimate can be derived from the asymptotic information matrix I_{RMLE}. Let $\tilde{\theta}_n$ denote the R-MLE based on the observations. It is known (Chang and Ying, 1999) that

$$\sqrt{I_{RMLE}(\tilde{\theta}_n)}(\tilde{\theta}_n - \theta_0) \longrightarrow N(0,1) \quad \text{in distribution,} \tag{8}$$

where

$$I_{RMLE}(\tilde{\theta}_n) = \sum_{i=1}^{n} \frac{a_i^2 e^{a_i(\tilde{\theta}_n - b_i)}}{1 + e^{a_i(\tilde{\theta}_n - b_i)}}.$$

The results for the first set of experiment are summarized in Table 4. Except when the test is short (e.g., $n = 10$) and the true ability is far away from the initial estimate of 0.0, the variances of the R-MLE and up-and-down methods (VAR) are quite similar, and the asymptotic variance (ASY VAR) also appears to be a good approximation of the estimated variance.

The second set of experiments focuses on efficiencies across the two methods - R-MLE and up-and-down - in terms of coverage probability (CP) and averaged length of confidence interval (AL). Table 5 summarizes the comparison of efficiencies. With the same coverage probability, a small average length indicates a better interval estimation. Both R-MLE and up-and-down perform well in terms of coverage probability at the level of $\alpha = 0.05$ Type I error. Apparently, the up-and-down method is not as efficient as the R-MLE and requires slightly wider 95% confidence intervals to achieve the same nominal Type I error rate. However, the differences are small. Generally the loss in efficiency is less than 10%. Using $n = 150$ as an example, even at true ability not close to 0 ($\theta = 2$), the loss in efficiency in terms of AL is 7.0%.

Table 5. CP at $\alpha = 0.05$ and AL of the 95% confidence intervals for $\theta = 0, 1, 2$

Method	n	$\theta = 0$		$\theta = 1$		$\theta = 2$	
		CP	AL	CP	AL	CP	AL
R-MLE	50	0.950	1.132	0.952	1.136	0.949	1.146
up-and-down	50	0.946	1.117	0.951	1.145	0.950	1.228
R-MLE	100	0.950	0.795	0.950	0.796	0.948	0.800
up-and-down	100	0.947	0.790	0.948	0.800	0.949	0.827
R-MLE	150	0.955	0.647	0.950	0.647	0.949	0.649
up-and-down	150	0.953	0.646	0.950	0.651	0.951	0.666

The CPU times for the R-MLE and up-and-down for test length $n = 100$ with 1,000 replications were 935.5s and 9.2s, respectively.

6. REAL DATA ANALYSIS

We include two real data examples - one from education and the other from patient-reported outcomes in health science - to illustrate the up-and-down method and compare it against the two R-MLE methods - R-MLE (A) and R-MLE (B). We used a sample of responses from an ACT test, which contained a total of 150 math and science items. We used item parameters from all items in the simulation experiments, but here we only used the math items. A total of $L = 60$ dichotomously scored math items were available for this analysis and the items were calibrated using a 2PL model using responses from a sample of $N = 23{,}096$ students. The item parameters of the 60 items are provided in supplementary materials http://www.intlpress.com/site/pub/files/_supp/sii/2020/0013/0003/sii-2020-0013-0003-s001.pdf. A total of $K = 5$ partitions were used for the a-stratified item selection. Test lengths were set at $n = 20, 25$ and $n = 30$. Bias and RMSE were calculated using the entire sample of students. Because there is no "true" ability value, we used the estimate from all 60 items as a proxy for the true value of θ. The actual raw responses from the students were used in the analysis for selecting items in CAT. The computational times of R-MLE (A), R-MLE (B), and up-and-down methods for test length $n = 30$ were respectively, 100.6, 24.2, and 4.0 minutes.

Table 6 shows the result of the analysis. The up-and-down method generally has slightly higher RMSE. Biases across the three methods are all small and do not differ across methods. The performance of the up-and-down method is quite comparable to the other two methods even when the length of test is short at $n = 20$. In this analysis, we also noticed that not all up-and-down method administrated the designated number of items. This is due to ceiling/flooring effect within stratum. In other words, when the up-and-down method administers an item that has maximum (minimum) difficulty within a stratum, and receives a correct response, the up-and-down method cannot proceed even when not all required number of within-stratum items have been administered. The method would simply move on to the next stratum. In this ACT data analysis, when $n = 30$ items were used, respectively 10, 651, and 5713 examinees were administered fewer than 20, 25, and 30 items.

Table 6. Bias and RMSE: ACT Math

Method	Metric	$n = 20$	$n = 25$	$n = 30$
R-MLE (A)	Bias	−0.019	−0.018	−0.015
	RMSE	0.288	0.242	0.203
R-MLE (B)	Bias	−0.023	−0.022	−0.020
	RMSE	0.289	0.244	0.206
up-and-down	Bias	−0.009	0.006	0.003
	RMSE	0.317	0.263	0.222

Because fewer items imply lower level of accuracy, we studied the effect of getting fewer items on standard error of the ability estimate from the up-and-down method. We plotted the standard errors of the ability estimates from the up-and-down method for examinees with fewer than 20, 25, and 30 items as a function of the estimated ability. Figure 3 (a)–(c) show the different scenarios. We plotted a random sample of 100 to alleviate overlapping of points. For comparison, the corresponding estimates for the R-MLE of the same individuals are also plotted. Additionally, the graph also shows those that were given the full set of 60 items (black dot) for the condition $n = 30$ (Fig. 3(d)). It can be seen that the two R-MLE methods provide approximately the same SE, while the up-and-down method has slightly higher SE. However, the differences begin to disappear at approximately 25 items. Note that the R-MLE methods do not necessarily have the same number of items administered as the up-and-down method. It is interesting to note that those that were administered fewer than 20, 25, and 30 items tended to have ability estimates between 0 and 1. We further investigated the phenomenon and found that it was an artifact of the high correlation between the a and b parameters as well as the asymmetric distribution of the difficulty parameters, which tended to be left-skewed. Figure 4 shows the distribution of the items on the b parameter on the 5 strata (labeled K1 through K5). When the up-and-down algorithm started at the initial value of $\theta = 0$ for the first stratum (low a), because of the correlation between a and b, which was approximately −0.6 for the ACT data, there were often not sufficient easy items in the same stratum for lower ability students. We will discuss this further in the last section on remarks and further research.

In the second real data example, we used 28 items that measure depression and used a subset of available data collected from the Patient Reported Outcome Measurement Information System (PROMIS [23]). Data from a total of $N = 768$ individuals were made available. Examples of items on the PROMIS depression scale are "I felt hopeless", "I felt that I had nothing to look forward to", and "I withdrew from other people." These items were adopted on a 7-day timeframe. In the original data, the response format was on a 5-point ordered scale: never(1), rarely(2), sometimes(3), often(4), and always(5). We used two ways to dichotomize the responses: (I) $\{1\}$ versus $\{2, 3, 4, 5\}$, and (II) $\{1, 2\}$ versus $\{3, 4, 5\}$. The correlations between the a and b parameters for the dichotomizing schemes (I) and (II) were respectively 0.469 and 0.030. Because each participant only responded to a small subset of questions, we could not use raw responses for implementing CAT. To circumvent this problem, we first calibrated the items based on the two dichotomization schemes using a 2-PL model, and then generated a full set of responses for $n = 768$ individuals.

Table 7 summarizes the result for the PROMIS data analysis. Consistent with the ACT data analysis, the RMSEs for up-and-down are generally larger than the other two R-MLEs. The performance of the up-and-down method for set (II) is worse than for set (I). Biases are generally low across the three methods. The higher values in RMSE for short tests ($n = 12, 16$) in PROMIS reflect the larger variance error in the up-and-down estimate. For realistic correlation values between a and b, the RMSE for the up-and-down method becomes smaller when the correlations are smaller. The computational times of R-MLE (A), R-MLE (B), and up-and-down methods for test length $n = 12$ were respectively, 283.2, 64.7, and 9.9 seconds.

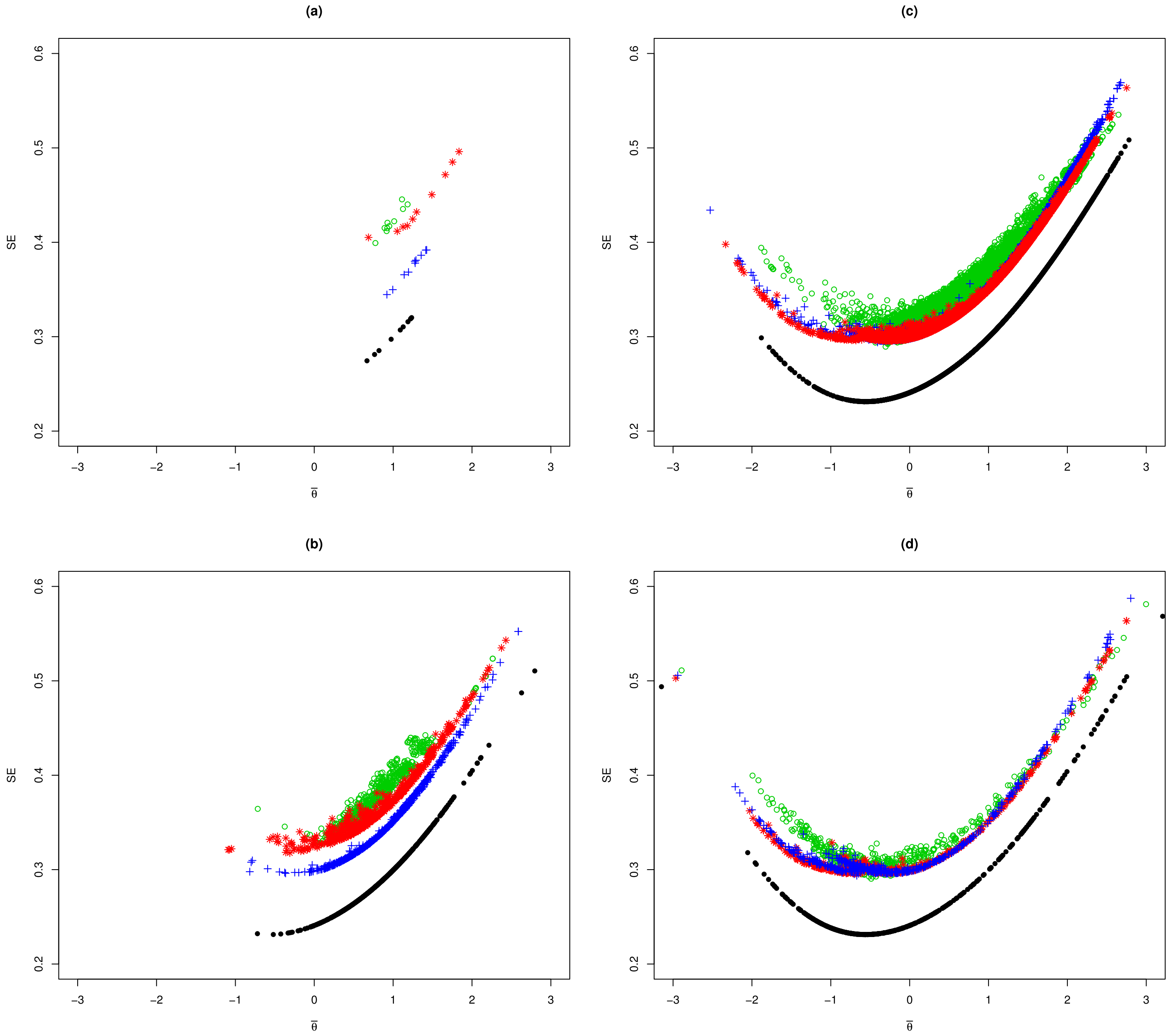

Figure 3. Plot of standard error (SE) against ability with sample from up-and-down method stopped before administrating 20 (panel (a)), 25 (panel (b)), and 30 (panel (c)) items. Panel (d) shows SE for up-and-down CAT method stopped after administrating 30 items (500 randomly selected individuals shown). Methods: red $\star$ = R-MLE (A); blue $+$= R-MLE (B); green $\circ$ = up-and-down; black $\bullet$= R-MLE from all 60 items.

7. CONCLUDING REMARKS, DISCUSSION, AND FURTHER RESEARCH

In this paper we propose an up-and-down method in the context of CAT. The up-and-down method is not new and has been applied in areas such as toxicity assessment (e.g., Bruce [2]). Applying it to CAT is novel and the method has several appeals from a psychometric perspective. First, it is intuitive and easy to understand. Second, the method is easy to implement in CAT, both in terms of software implementation as well as deployment on different platforms (e.g., smartphone and tablet). Third, the saving in computation could be substantial. The up-and-down method does not require computationally intensive procedure for updating ability estimate at each item selection decision point. From our simulation study, the ratio between the CPU time for up-and-down to CPU time for R-MLE was approximately 1 : 200. If the computation requires communication overhead between the data device and the server where computation of R-MLE occurs, this ratio could even be higher.

Table 7. Bias and Root Mean Square Error (RMSE): PROMIS data

		(I)		(II)	
Method	Metric	$n = 12$	$n = 16$	$n = 12$	$n = 16$
R-MLE (A)	Bias	−0.028	−0.022	−0.051	−0.041
	RMSE	0.166	0.115	0.127	0.088
R-MLE (B)	Bias	−0.031	−0.024	−0.035	−0.042
	RMSE	0.183	0.130	0.152	0.103
up-and-down	Bias	−0.052	−0.052	−0.043	−0.039
	RMSE	0.234	0.187	0.287	0.218

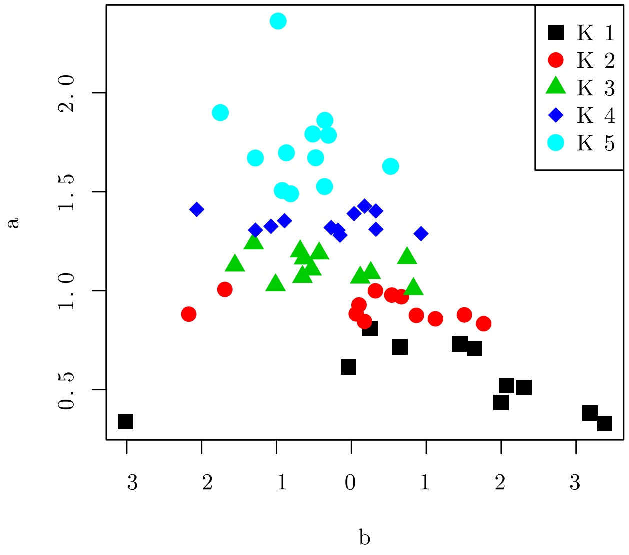

Figure 4. Scatterplot of item parameters by stratum for the ACT study.

Indeed, in many applications, the overhead input/output times across client and server are much higher than the actual CPU times. Additionally, using raw responses to select item can be easily implemented on the client side. This could potentially circumvent logistic issues such as software licensing. Thus, using raw responses in the up-and-down method for CAT is highly suitable for applications that require data management on distributed client-server systems.

The current article potentially opens a new avenue of research into adaptive algorithms, which we call RRD-CAT - that are based on raw responses for item selection. A potential powerful application include JITAI such as adaptation to mobile phone-delivered interventions to the dynamic of an individual's psychological, social, and contextual state. Another possibility is the application of the algorithm in the context of learning and cognitive diagnostics in education. With the omnipresence of tablet in classroom and the development of new software-based learning tools, the up-and-down algorithm could provide an efficient method for implementing adaptive assessment and learning opportunities in such situations.

The current article provides both theoretical basis and empirical evaluation of the up-and-down CAT method. Theoretical results concerning the asymptotic behavior of the up-and-down estimate using Markov chain-based method are derived. Additionally, the tools that were developed (e.g., Theorem A.1) for this purpose could serve as vehicles for generalizing the up-and-down or Markov chain-based method in other RRD-CAT settings. In this paper we provide equations for computing the asymptotic standard error of up-and-down estimates. Empirical studies including simulation experiments and real data analysis provide evidence that the up-and-down method perform reasonably well when compared to the computationally intensive R-MLE method in terms of accuracies, coverage probability of confidence intervals, and efficiency of the estimates. Understandably, by design the up-and-down is not as efficient as R-MLE. However, our data analysis shows that the loss in efficiency is generally quite small especially when the test is not too short.

The ACT real data analysis reveals a potential improvement of the up-and-down method. Figure 4 suggests that using a common starting value at 0.0 each strata may not be highly efficient because for some strata the distributions of the b parameter are not be centered around 0.0. Therefore we investigated the performance of a more flexible scheme - using the median of the b parameter within each stratum as starting value. We used the same setup as in Simulation Study I (setting I). The performance of the more flexible scheme is reported in the last rows of Tables 2 and 3, and labeled up-and-down (median). The result shows that there is no noticeable difference between the two up-and-down methods. For example, the median method appears to perform better for $n = 30$ and $K = 5$ but for $K = 10$ the result is not better. Further work will be required to explore more efficient solution.

We identify several limitations of the current study and some other future research directions as well. First, the asymptotic variance estimate may not work well in short test. Currently a bootstrap procedure to correct bias in asymptotic variance is under investigation. Second, our preliminary study suggested that the up-and-down procedure could be rather robust to model misspecification. Because misspecification in the parametric form of the item response function or in the dimensionality of the item response model could have an effect on CAT item selection and ability estimate, it is important to study the robustness of RRD-CAT methods. We speculate that up-and-down could proved to be robust because of the direct use of raw response for selecting a subsequent item. Clearly the robustness of the up-and-down algorithm to model misspecification will require further research. Finally, we have not investigated the performance of the up-and-down for classification purpose (e.g., pass versus no pass) such as for licensure exams. The issue is of practical importance and a comparison between the R-MLE and the up-and-down methods will be of value to practitioners of such exams.

APPENDIX A. TECHNICAL RESULTS FOR ASYMPTOTIC ANALYSIS

We introduce several prerequisite results that would be used for proving the main theorem (Theorem 4.1). Denote P_π as the response probability, whose initial distribution is the stationary distribution π, and let E_π be the expectation under P_π. In order to employ the technique of regenerative process to establish the consistency and asymptotic normality, we need to show that the Markov chain (2) is both irreducible and positive recurrent.

Consider a Markov chain $\{X_n, n \geq 0\}$ on a countably state space $S = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$, with transition probability p_{ij} for $i, j \in S$. Denote $\sum_{n=1}^{\infty} P\{X_\nu(\omega) \neq i, 0 < \nu < n; X_n(\omega) = i | X_0(\omega) = i\}$ by $f^*_{i,i}$. A state $i \in S$ is called recurrent if $f^*_{i,i} = 1$. Assume that state i is recurrent, let T_i be the first regeneration time of X_n to state i, that is,

$$T_i = \begin{cases} \inf\{n \geq 1, X_n = i\}; \\ \infty, \quad \text{if no such } n \text{ exist.} \end{cases}$$

A recurrent state i is called positive if and only if $E(T_i) < \infty$. The irreducibility of the Markov chain implies that if a state i is positive recurrent, then all states are positive recurrent.

Now we state the key theorem for proving Theorem 4.1.

Theorem A.1. *Let $\{X_n, n \geq 0\}$ be an ergodic (irreducible, aperiodic and positive recurrent) Markov chain on a countably state space $S = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$, with stationary distribution π. Let h be a real-valued function on the state space S. Suppose $E_\pi(|h|) < \infty$. The following holds.*

(a) $\sum_{t=1}^{N} h(X_t)/N$ converges to $E_\pi\{h(X_1)\}$ in probability.
(b) If $E_\pi(|h|^2) < \infty$ and $\sigma^2 := Var(\sum_{t=1}^{T_{x_0}} h(X_t)) < \infty$, then

$$(9) \qquad \frac{\sqrt{N}}{\sigma\sqrt{\pi(x_0)}} \left\{ \frac{\sum_{t=1}^{N} h(X_t)}{N} - \frac{E_\pi\{h(X_1)\}}{\pi(x_0)} \right\} \longrightarrow N(0,1) \quad \textit{in distribution.}$$

The proof of Theorem A.1 is given after the proof of Theorem 4.1. The following theorems lay the ground for the conditions of ergodicity in Theorem A.1.

By the definition of (2), it is easy to see that the properties of irreducibility and aperiodicity hold for the Markov chain produced by the up-and-down method. To prove the property of positive recurrent, we need following Theorems.

Theorem A.2. *Let $\{X_n, n \geq 0\}$ be a Markov chain on a countably state space $S = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$, with transition probability p_{ij} for $i, j \in S$. Let $0 < \alpha_i < 1$ and $\beta_i = 1 - \alpha_i$ be given numbers such that*

$$\begin{aligned} &p_{i,i+1} = \alpha_i, \quad p_{i,i-1} = \beta_i, \quad \textit{for} \quad i \geq 0; \\ &p_{i,i-1} = \alpha_i, \quad p_{i,i+1} = \beta_i, \quad \textit{for} \quad i < 0. \end{aligned}$$

(a) The state 0 is recurrent, i.e., $f^*_{0,0} = 1$ if and only if

$$\sum_{r\geq 1} \frac{\beta_1 \times \cdots \times \beta_r}{\alpha_1 \times \cdots \times \alpha_r} = \infty, \quad \sum_{r\geq 1} \frac{\beta_{-1} \times \cdots \times \beta_{-r}}{\alpha_{-1} \times \cdots \times \alpha_{-r}} = \infty.$$

(b) The recurrent state 0 is positive if and only if

$$\sum_{r\geq 1} \frac{\alpha_1 \cdots \alpha_{r-1}}{\beta_1 \cdots \beta_{r-1}\beta_r} < \infty, \quad \sum_{r\geq 1} \frac{\alpha_{-1} \cdots \alpha_{-(r-1)}}{\beta_{-1} \cdots \beta_{-(r-1)}\beta_{-r}} < \infty.$$

Remark. Results of Theorem A.2 can be found in [6], in which only the one-sided Markov chain is studied. The argument there can be generalized easily to the above results for two-sided Markov chain.

Using Theorem A.2, we now show that the Markov chain generated by the up-and-down method is positive recurrent.

Theorem A.3. *The Markov chain with transition probability defined in (2) is positive recurrent.*

Proof. Let n_0 be the integer satisfying $n_0 - 1 \leq \theta < n_0$. Denote

$$\begin{aligned} &p_{n_0+i,n_0+i+1} = \alpha_i, \quad p_{n_0+i,n_0+i-1} = \beta_i, \qquad \text{for} \quad i \geq 0, \\ &p_{n_0+i,n_0+i-1} = \alpha_i, \quad p_{n_0+i,n_0+i+1} = \beta_i, \qquad \text{for} \quad i < 0. \end{aligned}$$

Note that the logistic curve is monotone increasing. Hence $\beta_i/\alpha_i > 1$ for $n \geq n_0$. We conclude

$$\sum_{r\geq 1} \frac{\beta_1 \times \cdots \times \beta_r}{\alpha_1 \times \cdots \times \alpha_r} = \infty, \quad \sum_{r\geq 1} \frac{\beta_{-1} \times \cdots \times \beta_{-r}}{\alpha_{-1} \times \cdots \times \alpha_{-r}} = \infty.$$

It follows from Theorem A.2 (a) that we have $f_{n_0,n_0} = 1$ and n_0 is a recurrent state.

Next we show that $E(T_{x_0}) < \infty$, for all x_0. Recall that $\beta_i > 1/2$, $\alpha_i/\beta_i < 1$ and α_i is monotone decreasing for $n \geq n_0$. We have

$$\sum_{r\geq 1} \frac{\alpha_1 \times \cdots \times \alpha_{r-1}}{\beta_1 \times \cdots \times \beta_r} < \infty.$$

By a similar argument, we have

$$\sum_{r\geq 1}\frac{\alpha_{-1}\times\cdots\times\alpha_{-(r-1)}}{\beta_{-1}\times\cdots\times\beta_{-r}}<\infty.$$

Theorem A.2 (b) hence leads that x_0 is positive recurrent. Since the Markov chain (2) is irreducible and this implies that whole states in S are positive recurrent. That is, the Markov chain is an irreducible, aperiodic and positive recurrent Markov chain. Then, the vector $\varphi = (\cdots, 1/E(T_{-1}), 1/E(T_0), 1/E(T_1), \cdots)$ is a stationary probability distribution for (2). □

Remark. Note that the up-and-down method is a nonparametric method of selecting test items. Hence, as long as the item response function is continuous, strictly monotone increasing, and ranges over $(0,1)$, Theorem A.3 remains to hold.

APPENDIX B. ASYMPTOTIC BEHAVIOR OF THE MLE

By using the results in Theorem A.1, we will prove our main results, the weak consistency and asymptotic normality of the MLE $\hat{\theta}_n$. One major contribution here is the characterization of the Fisher information in Theorem 4.1, for which it can be used to construct confidence interval of θ_0, the true parameter.

Proof of Theorem 4.1. The proof will follow the argument outlined in Section 2. First, we show that (5) holds. Note that $X_0 = 0$, and

$$\begin{aligned} & E_{\theta_0}\left\{\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0)\bigg| X_1\right\} \\ = \; & E_{\theta_0}\left\{\frac{\partial f(X_1,X_2;\theta_0)/\partial\theta}{f(X_1,X_2;\theta_0)}\bigg| X_1\right\} \\ = \; & \sum_{x_2\in S}\frac{\partial}{\partial\theta}f(X_1,x_2;\theta_0). \end{aligned}$$

Differentiate both side of the equation

$$\sum_{x_2\in S}f(x_1,x_2;\theta)=1$$

with respect to θ leads to

$$\sum_{x_2\in S}\frac{\partial}{\partial\theta}f(x_1,x_2;\theta)=0.$$

This implies that $E_{\theta_0}\{\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0)\}=0$. Since

$$\frac{\partial}{\partial\theta}g(x_1,x_2;\theta)=\begin{cases}1/(1+e^{\theta-x_1}), & x_2=x_1+1,\\ -e^{\theta-x_1}/(1+e^{\theta-x_1}), & x_2=x_1-1,\end{cases}$$

we have

$$E_{\theta_0}\left\{\left|\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0)\right|\bigg| X_1\right\}\leq\frac{1}{4}.$$

By Theorem A.1 (a), (5) holds by

$$\begin{aligned} & \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\theta}g(X_{t-1},X_t;\theta)\bigg|_{\theta=\theta_0} \\ & \to E_{\theta_0}\left\{\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0)\right\}=0 \quad \text{in probability.} \end{aligned}$$

Next, we show that (6) holds. Twice differentiation of $\sum_{x_2\in S}f(x_1,x_2;\theta)=1$ with respect to θ leads to

$$\sum_{x_2\in S}\frac{\partial^2}{\partial\theta^2}f(x_1,x_2;\theta)=0,$$

and

$$\begin{aligned} & E_{\theta_0}\left\{\frac{\partial^2}{\partial\theta^2}g(X_1,X_2;\theta_0)\bigg| X_1\right\} \\ = \; & E_{\theta_0}\left\{\frac{\frac{\partial^2}{\partial\theta^2}f}{f}-\frac{(\frac{\partial f}{\partial\theta})^2}{f^2}\bigg| X_1\right\} \\ = \; & E_{\theta_0}\left\{\frac{\frac{\partial^2}{\partial\theta^2}f}{f}\bigg| X_1\right\}-E_{\theta_0}\left\{\frac{(\frac{\partial f}{\partial\theta})^2}{f^2}\bigg| X_1\right\} \\ = \; & -E_{\theta_0}\left\{\left[\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0)\right]^2\bigg| X_1\right\}. \end{aligned}$$

Again, for all given $x_1\in S$, we have that

$$E_{\theta_0}\left\{\left[\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0)\right]^2\bigg| X_1\right\}=\frac{e^{\theta_0-x_1}}{(1+e^{\theta_0-x_1})^2}\leq\frac{1}{4}.$$

Therefore (7) holds.

We also need to calculate $E_{\theta_0}\{|\frac{\partial^2}{\partial\theta^2}g(X_1,X_2;\theta_0)|\}$. For $x_2=x_1+1$ or $x_2=x_1-1$, we have

$$\left|\frac{\partial^2}{\partial\theta^2}g(X_1,X_2;\theta_0)\right|=\frac{e^{(\theta_0-x_1)}}{[1+e^{(\theta_0-x_1)}]^2}\leq\frac{1}{4},$$

and this implies that

$$E_{\theta_0}\left\{\left|\frac{\partial^2}{\partial\theta^2}g(X_1,X_2;\theta_0)\right|\right\}<\infty.$$

It follows from Theorem A.1 (b) that (6) holds by

$$\begin{aligned} & \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2}{\partial\theta^2}g(X_{k-1},X_k;\theta)\bigg|_{\theta=\theta_0} \\ & \to E_{\theta_0}\left\{\frac{\partial^2}{\partial\theta^2}g(X_1,X_2;\theta_0)\right\}=-\mathrm{I} \quad \text{in probability.} \end{aligned}$$

Denote

$$\begin{aligned} G(x_1,x_2) &:= \sup_{\theta\in R}\left|\frac{\partial^3}{\partial\theta^3}g(x_1,x_2;\theta)\right| \\ &= \sup_{\theta\in R}\left|\frac{e^{\theta-x_1}(1-e^{\theta-x_1})}{(1+e^{\theta-x_1})^3}\right|<1. \end{aligned}$$

There exists a constant M such that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}G(X_{t-1},X_t)=M \quad \text{in probability.} \tag{10}$$

By the mean value theorem, for some $|\alpha|<1$, we have

$$\begin{aligned} \frac{1}{n}\frac{\partial}{\partial\theta}L_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\theta}g(x_{t-1},x_t;\theta) \\ = & \frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\theta}g(x_{t-1},x_t;\theta_0) \\ & +\frac{1}{n}(\theta-\theta_0)\sum_{t=1}^{n}\frac{\partial^2}{\partial\theta^2}g(x_{t-1},x_t;\theta_0) \\ & +\frac{\alpha}{2n}(\theta-\theta_0)^2\sum_{t=1}^{n}G(x_{t-1},x_t). \end{aligned}$$

Let S^* denote the collection of $(x_1,\cdots,x_n)$ satisfying

$$\begin{aligned} & \left|\frac{1}{n}\sum_{k=1}^{n}\frac{\partial}{\partial\theta}g(x_{t-1},x_t;\theta_0)\right|<\delta^2, \\ & \frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2}{\partial\theta^2}g(x_{t-1},x_t;\theta_0)<-\mathrm{I}/2, \quad \text{and} \\ & \frac{1}{n}\sum_{t=1}^{n}G(x_{t-1},x_t)<2M. \end{aligned}$$

It follows from (5), (7) and (10) that, for all δ, ε, there exists an n_0 such that $P(S^*)>1-\varepsilon$ when $n>n_0(\delta,\varepsilon)$.

For $\theta=\theta_0\pm\delta$, choose $\delta<\frac{1}{2}\mathrm{I}/(M+1)$, then,

$$\frac{1}{n}\frac{\partial}{\partial\theta}L_n(\theta)\bigg|_{\theta=\theta_0+\delta}\le\delta^2-\frac{1}{2}(\mathrm{I}\cdot\delta)+M\delta^2<0,$$

if $(x_1,\cdots,x_n)\in S^*$. By the same argument, we have

$$\frac{1}{n}\frac{\partial}{\partial\theta}L_n(\theta)\bigg|_{\theta=\theta_0-\delta}>0.$$

Since $\frac{1}{n}\frac{\partial}{\partial\theta}L_n(\theta)$ is continuous, so for any $\delta,\varepsilon>0$, the likelihood equation will, with probability exceeding $1-\varepsilon$, have a root belongs to $(\theta_0-\delta,\theta_0+\delta)$ as long as $n>n_0(\delta,\varepsilon)$. We conclude that

$$\hat{\theta}_n\longrightarrow\theta_0 \quad \text{in probability.} \tag{11}$$

To prove 2), we first characterize the asymptotic variance $\mathrm{Var}\Big(\sum_{t=1}^{T_{x_0}}\frac{\partial}{\partial\theta}g(X_{t-1},X_t;\theta_0)\Big)$, where T_{x_0} is the first regeneration time to state x_0. Recall that $E_{\theta_0}\{\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0)\}=0$, and $E_{\theta_0}\{(\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0))^2\}=I$.

$$\begin{aligned} & \mathrm{Var}\left(\sum_{t=1}^{T_{x_0}}\frac{\partial}{\partial\theta}g(X_{t-1},X_t;\theta_0)\right) \\ = & E_{\theta_0}\Big(\sum_{k=1}^{T_{x_0}}\frac{\partial}{\partial\theta}g(X_{t-1},X_t;\theta_0)\Big)^2- \\ & \Big[E_{\theta_0}\Big(\sum_{t=1}^{T_{x_0}}\frac{\partial}{\partial\theta}g(X_{t-1},X_t;\theta_0)\Big)\Big]^2 \\ = & E_{\theta_0}\left(\sum_{k=1}^{T_{x_0}}\left[\frac{\partial}{\partial\theta}g(X_{t-1},X_t;\theta_0)\right]^2\right)+ \\ & 2\sum_{t'>t}E_{\theta_0}\Big(\frac{\partial}{\partial\theta}g(X_t,X_{t+1};\theta_0)\frac{\partial}{\partial\theta}g(X_{t'},X_{t'+1};\theta_0)\Big) \\ = & \frac{1}{\pi(x_0)}E_{\theta_0}\{(\frac{\partial}{\partial\theta}g(X_1,X_2;\theta_0))^2\}+ \\ & 2\sum_{t'>t}E_{\theta_0}\Big[E_{\theta_0}\Big(\frac{\partial}{\partial\theta}g(X_t,X_{t+1};\theta_0)\times \\ & \frac{\partial}{\partial\theta}g(X_{t'},X_{t'+1};\theta_0)|X_t,X_{t+1},X_{t'}\Big)\Big] \\ = & \frac{\mathrm{I}}{\pi(x_0)}+2\sum_{t'>t}E_{\theta_0}\Big[\frac{\partial}{\partial\theta}g(X_t,X_{t+1};\theta_0)\times \\ & E_{\theta_0}\Big[\frac{\partial}{\partial\theta}g(X_{t'},X_{t'+1};\theta_0)\Big|X_{t'}\Big]\Big] \\ = & \frac{\mathrm{I}}{\pi(x_0)}. \end{aligned}$$

By Theorem A.1, we have

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial}{\partial\theta}g(X_{t-1},X_t;\theta_0)\longrightarrow N(0,I) \text{ in distribution.}$$

Note that the score equation $n^{-1}\frac{\partial}{\partial\theta}L_n(\hat{\theta}_n)=0$ can be written as

$$\begin{aligned} 0 = & \frac{1}{n}\sum_{k=1}^{n}\frac{\partial}{\partial\theta}g(x_{t-1},x_t;\theta_0) \\ & +\frac{1}{n}(\hat{\theta}_n-\theta_0)\sum_{k=1}^{n}\frac{\partial^2}{\partial\theta^2}g(x_{t-1},x_t;\theta_0) \\ & +\frac{\alpha}{2n}(\hat{\theta}_n-\theta_0)^2\sum_{k=1}^{n}G(x_{t-1},x_t). \end{aligned}$$

We have

$$\sqrt{n}(\hat{\theta}_n-\theta_0)$$

$$= \frac{-n^{1/2}\sum_{t=1}^{n} \frac{\partial}{\partial\theta} g(X_{t-1},X_t;\theta_0)}{\sum_{t=1}^{n} \left[\frac{\partial^2}{\partial\theta^2} g(X_{t-1},X_t;\theta_0) + \frac{\alpha}{2}(\hat{\theta}_n-\theta_0)G(X_{t-1},X_t)\right]}$$
$$\longrightarrow N(0,I^{-1}) \quad \text{in distribution.} \qquad \square$$

Proof of Theorem A.1 (a). For simplicity, set x_0 to be the state 0, which is positive recurrent, and denote $m := \sum_{j=1}^{N} I_{x_0}(X_j)$ as the number of visits to state x_0 up to N. It is known (cf. Chung, 1967) that $m/N \to \pi(x_0)$ in probability, where $\pi(x_0) = 1/E(T_{x_0})$. Let $T_{x_0}^k$ be the kth regeneration time to state x_0, and denote

$$\eta_j(h) := \sum_{i=T_{x_0}^{j-1}+1}^{T_{x_0}^{j}} h(x_i)$$

as the jth regeneration epoch. Note that $\{\eta_j(h), j = 1, \cdots, m\}$ forms i.i.d. blocks due to strong Markov property of the underlying Markov chain. Write

$$(12) \qquad \frac{1}{N}\sum_{j=1}^{N} h(X_j)$$
$$= \frac{1}{N}\sum_{j=T_{x_0}^m+1}^{N} h(X_j) + \frac{1}{N}\left(\sum_{j=1}^{m}\eta_j(h) - \sum_{j=1}^{[N\pi]}\eta_j(h)\right)$$
$$+\frac{1}{N}\sum_{j=1}^{[N\pi]}\eta_j(h) := I_1 + I_2 + I_3.$$

By the law of large numbers for i.i.d. random variables, we have

$$\frac{1}{N}\sum_{j=1}^{[N\pi(x_0)]}\eta_j(h) = \frac{[N\pi(x_0)]}{N}\frac{1}{[N\pi(x_0)]}\sum_{j=1}^{[N\pi(x_0)]}\eta_j(h)$$
$$\longrightarrow \pi(x_0)E_\pi(\eta) = E_\pi(h) \quad \text{in probability.}$$

Next, we show that both I_1 and I_2 converge to zero in probability. For any $\varepsilon > 0$,

$$P_\pi\left\{\left|\sum_{j=T_{x_0}^m+1}^{N} h(X_j)\right| > \varepsilon N\right\}$$
$$\leq P_\pi\left\{\sum_{j=T_{x_0}^m+1}^{N} |h(X_j)| > \varepsilon N\right\}$$
$$\leq P_\pi\left\{\sum_{j=T_{x_0}^m+1}^{T_{x_0}^{m+1}} |h(X_j)| > \varepsilon N\right\}$$
$$= P_\pi\left\{\eta_1(|h|) > \varepsilon N\right\} \leq \frac{E_\pi[\eta_1(|h|)]}{\varepsilon N} = \frac{E_\pi(|h|)}{\varepsilon\pi(x_0)N}.$$

The last inequality follows from Markov inequality and $E_\pi(|h|)/[\varepsilon\pi(x_0)N] \to 0$ as $N \to \infty$ by $E_\pi(|h|) < \infty$. This implies that $I_1 \to 0$ in probability.

Since $m/N \longrightarrow \pi(x_0)$ in probability, we have that for all $\varepsilon > 0$, there exists N_0 such that for $N > N_0$, $P_\pi\{|m - [N\pi(x_0)]| > N\varepsilon^2\} < \varepsilon$. Then, for $N > N_0$, we have

$$P_\pi\left\{\left|\sum_{j=1}^{m}\eta_j(h) - \sum_{j=1}^{[N\pi(x_0)]}\eta_j(h)\right| > \varepsilon N\right\}$$
$$\leq P_\pi(|m - [N\pi(x_0)]| > N\varepsilon^2)$$
$$+P_\pi\left\{\max_{|r-[N\pi(x_0)]|\leq\varepsilon^2 N}\left|\sum_{j=[N\pi(x_0)]+1}^{r}\eta_j(h)\right| > \varepsilon N\right\}$$
$$< \varepsilon + 2P_\pi\left\{\max_{1\leq r\leq\varepsilon^2 N}\left|\sum_{j=1}^{r}\eta_j(h)\right| > \varepsilon N\right\}$$
$$= \varepsilon + 2P_\pi\left\{\sum_{j=1}^{r}|\eta_j(h)| > \varepsilon N\right\}$$
$$< \varepsilon + \frac{2\varepsilon^2 N E(|\eta_1|)}{\varepsilon N} = \left(1 + \frac{2E_\pi(|h|)}{\pi(x_0)}\right)\varepsilon.$$

Therefore, $I_2 \to 0$ in probability. We conclude the proof of (a). □

Proof of Theorem A.1 (b). Using the same argument as in the proof of (a), we have

$$(13) \qquad \frac{\sqrt{N}}{\sigma\sqrt{\pi(x_0)}}\left(\frac{\sum_{j=1}^{N} h(X_j)}{N} - \frac{E_\pi[h(X_1)]}{\pi(x_0)}\right)$$
$$= \frac{1}{\sigma\sqrt{N\pi(x_0)}}\sum_{j=T_{x_0}^m+1}^{N} h(X_j)$$
$$+\frac{1}{\sigma\sqrt{N\pi(x_0)}}\left(\sum_{j=1}^{m}\eta_j(h) - \sum_{j=1}^{[N\pi(x_0)]}\eta_j(h)\right)$$
$$+\frac{\sqrt{N}}{\sigma\sqrt{\pi(x_0)}}\left(\frac{\sum_{j=1}^{[N\pi(x_0)]}\eta_j(h)}{N} - \frac{E_\pi[h(X_1)]}{\pi(x_0)}\right)$$
$$:= II_1 + II_2 + II_3.$$

First, we consider II_3. Note that η_j are i.i.d. random blocks. Under the condition of $E_\pi(|h|^2) < \infty$ and $\sigma^2 := \mathrm{Var}(\sum_{t=1}^{T_{x_0}} h(X_t)) < \infty$, by standard central limit theorem for i.i.d. random variables, we have

$$(14) \qquad II_3 = \frac{\sqrt{N\pi(x_0)}}{\sigma}\left(\frac{\sum_{j=1}^{[N\pi(x_0)]}\eta_j(h)}{N\pi(x_0)} - \frac{E_\pi[h(X_1)]}{\pi(x_0)}\right)$$
$$\longrightarrow N(0,1) \quad \text{in distribution.}$$

It remains to show that II_1 and II_2 converges to zero in probability. For any $\varepsilon > 0$, we have

$$\begin{aligned}
&P_\pi\left\{\left|\sum_{j=T^m_{x_0}+1}^{N} h(X_j)\right| > \varepsilon\sigma\sqrt{N\pi(x_0)}\right\} \qquad (15)\\
\leq\ & P_\pi\left\{\sum_{j=T^m_{x_0}+1}^{N} |h(X_j)| > \varepsilon\sigma\sqrt{N\pi(x_0)}\right\}\\
\leq\ & P_\pi\left\{\sum_{j=T^m_{x_0}+1}^{T^{m+1}_{x_0}} |h(X_j)| > \varepsilon\sigma\sqrt{N\pi(x_0)}\right\}\\
=\ & P_\pi\left\{\eta_1(|h|) > \varepsilon\sigma\sqrt{N\pi(x_0)}\right\}\\
\leq\ & \frac{E_\pi[\eta_1(|h|)]}{\varepsilon\sigma\sqrt{N\pi(x_0)}} = \frac{E_\pi(|h|)}{\varepsilon\pi(x_0)\sigma\sqrt{N\pi(x_0)}}.
\end{aligned}$$

Hence, II_1 converges to 0 in probability as $N \to \infty$ by assumption.

Since $m/N \longrightarrow \pi(x_0)$ in probability, we have for all $\varepsilon > 0$, there exists N_0 such that $N > N_0$, $P_\pi(|m - [N\pi(x_0)]| > N\varepsilon^3) < \varepsilon$. Clearly, for such N, we have

$$\begin{aligned}
&P_\pi\left(\left|\sum_{j=1}^{m}\eta_j(h) - \sum_{j=1}^{[N\pi(x_0)]}\eta_j(h)\right| > \varepsilon\sigma\sqrt{N\pi(x_0)}\right)\\
\leq\ & P_\pi(|m - N\pi(x_0)| > N\varepsilon^3) +\\
& P\left\{\max_{|r-N\pi(x_0)|\leq\varepsilon^3 N}\left|\sum_{j=N\pi(x_0)+1}^{r}\eta_j(h)\right| > \varepsilon\sigma\sqrt{N\pi(x_0)}\right\}\\
<\ & \varepsilon + 2P_\pi\left\{\max_{1\leq r\leq\varepsilon^3 N}\left|\sum_{j=1}^{r}\eta_j(h)\right| > \varepsilon\sigma\sqrt{N\pi(x_0)}\right\}\\
<\ & \varepsilon + \frac{2\varepsilon^3 N\sigma^2}{\varepsilon^2\sigma^2 N\pi(x_0)} = \varepsilon\left(1 + \frac{2}{\pi(x_0)}\right).
\end{aligned}$$

This proves that II_2 converges to 0 in probability as $N \to \infty$. We conclude the proof of (b). □

Received 31 October 2018

REFERENCES

[1] Binet, A. and Simon, T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année psychologique*, 11:191–244, 1904.

[2] Bruce, R. D. An up-and-down procedure for acute toxicity testing. *Fundamental and Applied Toxicology*, 5(1):151–157, 1985.

[3] Chang, H. and Ying, Z. A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23:211–222, 1999.

[4] Chang, H. and Ying, Z. Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37:1466–1488, 2009. MR2509080

[5] Chang, H., Qian, J., and Ying, Z. A-stratified multistage cat with b-blocking. *Applied Psychological Measurement*, 25:333–341, 2001. MR1863509

[6] Chung, K. L. *Markov chain with stationary transition probabilities.* New York: Springer-Verlag, 1967. MR0217872

[7] Csiksezentmihalyhi, M. and Larson, R. Validity and reliability of experience-sampling method. *Journal of Nervious and Mental Disease*, 175:526–536, 1987.

[8] Dixon, W. J. and Mood, A. M. A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association*, 43:109–126, 1948.

[9] Dodd, B. G., De Ayala, R. J., and Koch, W. R. Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19:5–22, 1995.

[10] Triantafilloum E., Georgiadou E., and Economides A. A. The design and evaluation of computerized adaptive test on mobile devices. *Computers & Education*, 50:1319–1330, 2008.

[11] Gibbons, R. D., Weiss, D. J., Pikonis, P. A., Frank, E., Moore, T., Kim, J. B., and Kupfer, D. J. Development of a computerized adaptive test for depression. *Archive of General Psychiatry*, 69:1104–1112, 2012.

[12] Huang, Y., Lin, Y., and Cheng, S. An adaptive testing system for supporting versatile educational assessment. *Computers & Education*, 52:53–67, 2009.

[13] Hulin, C. L., Drasgow, F., and Parsons, C. K. *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin, 1983.

[14] Lee, Y. H., Ip, E. H., and Fuh, C.-D. A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, 68:215–232, 2008. MR2427266

[15] Lord, M. F. Some test theory for tailored testing. In Holtzman W.H., editor, *Computer-assisted instruction, testing and guidance.* New York: Harper and Row, 1970.

[16] Lord, M. F.. Robbins-monro procedures for tailored testing. *Educational and psychological Measurement*, 31:3–31, 1971.

[17] Lord, M. F. A theoretical study of two-stage testing. *Psychometrika*, 36:227–242, 1971.

[18] Lord, M. F. *Applications of item response theory to practical testing problem.* Hillsdale, NJ: Lawrence Erlbaum, 1980.

[19] Nahum-Shani, I., Smith, S. N., Witkiewitz, K., Collins, L. M., Spring, B., and Murphy, S. A. Just-in-time adaptive interventions (jitais): an organizing framework for ongoing health behavior support. Technical report, The Methodology Center, Penn State University, 2014.

[20] Nocedal, J. and Wright, S. *Numerical Optimization.* Springer Science & Business Media, 2006. MR2244940

[21] Oron, A. and Hoff, P. The k-in-a-row up-and-down design, revisited. *Statistics in Medicine*, 28:1805–1820, 2009. MR2751599

[22] Owen, R. J. A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70:351–356, 1975. MR0381185

[23] Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., and Lawrence, S. M. Validation of the depression item bank from the patient-reported outcomes measurement information system (promis) in a three-month observational study. *Journal of Psychiatry Research*, 56:112–119, 2014.

[24] Rizopoulos, D. ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17:1–25, 2006. URL http://www.jstatsoft.org/v17/i05/.

[25] Segall, D. O. Multidimensional adaptive testing. *Psychometrika*, 61:331–354, 1996.

[26] Shiffman, S., Stone, A. A., and Hufford, M. R. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, 2008.

[27] Thompson, G. Computer adaptive testing, big data and algorithmic approaches to education. *British Journal of Sociology of Education*, 38:8278–840, 2017.

[28] Warmerdam, L., Riper, H., Klein, M., van den Ven, P., Rocha, A., Henriques, M. R., Tousset, E., Silva, H., Anderson, G., and Cujjpers, P. Innovative ict solutions to improve treatment outcomes for depression: The ict4depression project.

Annual Review of Cybertherapy and Telemedicine, 181:339–343, 2012.

[29] Weiss, D. J. Adaptive testing research in minnesota: Overview, recent results, and future directions. In L. Clark C., editor, *Proceedings of the first conference on computerized adaptive testing*, pages 24–35. United States Civil Service Commission, Washington DC, 1976.

Cheng-Der Fuh
Fanhai International School of Finance
Fudan University
China
E-mail address: cdffuh@gmail.com

Edward Haksing Ip
Department of Biostatistics and Data Science
Wake Forest School of Medicine
USA
E-mail address: eip@wakehealth.edu

Shyh-Huei Chen
Department of Biostatistics and Data Science
Wake Forest School of Medicine
USA
E-mail address: schen@wakehealth.edu