Journal of Computerized Adaptive Testing

Volume 7 Number 2

September 2019

Time-Efficient Adaptive Measurement of Change

Matthew D. Finkelman and Chun Wang

DOI 10.7333/1909-0702015

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

©2019 by the Authors. All rights reserved.

This publication may be reproduced with no cost for academic or research use.
All other reproduction requires permission from the authors;
if the author cannot be contacted, permission can be requested from IACAT.

Editor

David J. Weiss, *University of Minnesota, U.S.A.*

Consulting Editors

John Barnard
EPEC, Australia
Juan Ramón Barrada
Universidad de Zaragoza, Spain
Kirk A. Becker
Pearson VUE, U.S.A.
Theo H. J. M. Eggen
Cito and University of Twente, The Netherlands
Andreas Frey
Friedrich Schiller University Jena, Germany
Kyung T. Han
Graduate Management Admission Council, U.S.A.
Matthew D. Finkelman, *Tufts University School of Dental Medicine, U.S.A.*
G. Gage Kingsbury
Psychometric Consultant, U.S.A.

Wim J. van der Linden
University of Twente, The Netherlands
Alan D. Mead
Illinois Institute of Technology, U.S.A.
Mark D. Reckase
Michigan State University, U.S.A.
Barth Riley
University of Illinois at Chicago, U.S.A.
Bernard P. Veldkamp
University of Twente, The Netherlands
Chun Wang
University of Washington, U.S.A.
Steven L. Wise
Northwest Evaluation Association, U.S.A.
Anthony R. Zara
Pearson VUE, U.S.A.

Technical Editor

Barbara L. Camm

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 7, Number 2, September 2019

DOI 10.7333/1909-0702015

ISSN 2165-6592

Time-Efficient Adaptive Measurement of Change

Matthew D. Finkelman

Tufts University School of Dental Medicine

Chun Wang

University of Washington

The adaptive measurement of change (AMC) refers to the use of computerized adaptive testing (CAT) at multiple occasions to efficiently assess a respondent's improvement, decline, or sameness from occasion to occasion. Whereas previous AMC research focused on administering the most informative item to a respondent at each stage of testing, the current research proposes the use of Fisher information per time unit as an item selection procedure for AMC. The latter procedure incorporates not only the amount of information provided by a given item but also the expected amount of time required to complete it. In a simulation study, the use of Fisher information per time unit item selection resulted in a lower false positive rate in the majority of conditions studied, and a higher true positive rate in all conditions studied, compared to item selection via Fisher information without accounting for the expected time taken. Future directions of research are suggested.

Keywords: computerized adaptive testing, adaptive measurement of change, item selection, Fisher information, response-time modeling

The measurement of individual change has been a well-known topic in the psychometric literature for decades (e.g., Cronbach & Furby, 1970; Lord, 1956, 1958; Rogosa & Willett, 1983; Willett, 1989). In medical contexts, assessment may be conducted on multiple occasions in order to ascertain whether the severity of a respondent's condition has improved, declined, or remained the same from one occasion to the next. Indeed, the measurement of individual change has been studied among chiropractic patients with chronic neck or low back pain (Hays, Spritzer, Sherbourne, Ryan, & Coulter, 2019), patients undergoing foot and ankle treatment (Hung et al., 2019), and children with chronic pain (Kashikar-Zuck et al., 2016). Psychological assessment may also be conducted on more than one occasion for a given respondent to determine whether a significant change in their beliefs, mental health, or behavior has occurred (Brouwer, Meijer, & Zevalkink, 2013; Kruyen, Emons, & Sijtsma, 2014; Wang & Weiss, 2018). In the field of education, testing might be repeated to gauge a student's degree of improvement (or decline) in a given domain over time (Wang & Weiss, 2018; Weiss & Von Minden, 2011).

When conducting the measurement of individual change, the amount of time between testing occasions is an important consideration. Such follow-up time might vary from setting to setting depending on the context of the testing scenario, the participants, the anticipated amount of time needed for an observable change to occur, and logistical concerns. In medical contexts, Hays et al. (2019) used a three-month follow-up, whereas Hung et al. (2019) stratified their results by follow-up time, with three months being the shortest follow-up and a period of over six months being the longest. In the contexts of psychological and educational assessment, respectively, Brouwer et al. (2013) and Wang and Weiss (2018) both examined results after a follow-up time of approximately one year.

Although a frequent subject of research, the measurement of individual change has also elicited controversy. Change scores have been characterized as exhibiting problematic properties such as low reliability, negative correlation between initial level and change score, and scaling difficulties (e.g., Bereiter, 1963; Cronbach & Furby, 1970; Embretson, 1995). Some authors have suggested that item response theory (IRT) approaches to measuring individual change are superior to approaches based on classical test theory (Brouwer et al., 2013; Doucette & Wolf, 2009; Wang & Weiss, 2018). IRT provides a framework for modeling respondent answers from the latent trait being measured and the parameters of the items administered. One benefit of IRT is that it facilitates the use of computerized adaptive testing (CAT), in which the item selected for a respondent at a given stage of testing is influenced by the respondent's answers to previous items in the assessment. Such tailoring of the assessment to each respondent allows more efficient measurement (van der Linden & Glas, 2000; Weiss, 2004), including within the context of measuring change (Kim-Kang & Weiss, 2008; Weiss, 2011; Weiss & Von Minden, 2011).

One concern that arises when assessment occurs at multiple occasions longitudinally is the cumulative respondent and administrative burden that is incurred as the result of repeated testing. Respondent burden refers to the extent to which individuals taking an assessment (or multiple assessments) feel that the experience is stressful, difficult, or time-consuming (Graf, 2008). Administrative burden refers to the extent to which costs are incurred through the supervision of the assessment process (Forbey & Ben-Porath, 2007). When assessment is conducted on multiple occasions, it is vital to keep instruments short in order to avoid requiring an inordinate amount of time from respondents and providers (Kruyen et al., 2014; Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012). The aforementioned benefit of CAT in improving the efficiency of measurement thus becomes particularly important for assessments that are repeated longitudinally. The use of CAT in evaluating a respondent's improvement, decline, or sameness from occasion to occasion is called the adaptive measurement of change (AMC; Kim-Kang & Weiss, 2008).

Traditionally, the respondent and administrative burden of a CAT has been quantified based on the number of items administered by the CAT. Accordingly, most item selection procedures in CAT have been developed with the goal of minimizing this number (assuming that "all else is equal" with respect to measurement properties). For instance, a classic item selection procedure is to maximize Fisher information at the respondent's current estimate of the latent trait (Lord, 1980); this method is designed to obtain a precise estimate with as few items as possible. In more recent work, some researchers have used an alternative paradigm in which it is not the number of items that is used as a proxy for respondent and administrative burden, but rather the total time taken by the assessment. Indeed, it is typical that some items in a bank tend to be answered more quickly than others; respondents and administrators might be willing to accept a larger number of "fast" items if the total amount of time taken by those items is reduced. In medical contexts, for example, it is important to make testing time as short as possible due to the limited time that providers can spend with each patient (Dugdale, Epstein, & Pantilat, 1999; Smits et al., 2012). In education, time efficiency is likewise critical because of the limited classroom time that teachers have with students (Welch & Frick, 1993).

For CAT settings in which the burden of an assessment is measured by the time required to complete it, a modification of the traditional Fisher information (FI) item selection procedure has been developed (Fan, Wang, Chang, & Douglas, 2012). Specifically, the proposed modification is to maximize Fisher information per time unit (FITU) rather than solely FI. The former item selection procedure has been found to enhance the time efficiency of assessment compared to the latter in the context of a single testing session (Fan et al., 2012).

Given the crucial need for time efficiency in the measurement of individual change, particularly considering the cumulative amount of time taken by its multiple test administrations, it would be natural to combine the FITU item selection procedure with AMC. This item selection procedure can be used in applications of AMC whether assessment is conducted in the educational, psychological, or medical context, and irrespective of the follow-up time between testing occasions. Indeed, it can be applied as long as the FI and expected time of each item can be determined. However, as will be discussed, some details (namely, how to calculate the expected time) might change depending on the context.

A detailed comparison of the two aforementioned item selection procedures (FITU and FI) would illuminate the magnitude of difference between these procedures in the AMC setting. However, no previous research appears to have explored FITU as an item selection method in AMC. The objective of this study was to fill this gap by proposing the use of FITU within the AMC framework and comparing it to FI under a variety of conditions in simulation.

Method

Hypothesis Testing in AMC

Although AMC can be used to detect intra-individual change at more than two occasions (Phadke, 2017), attention focuses on the case of two occasions, as is common in AMC research (Finkelman, Weiss, & Kim-Kang, 2010; Kim-Kang & Weiss, 2008). Letting θ_1 denote a given respondent's latent trait at the first occasion (hereafter Occasion 1), and letting θ_2 denote the respondent's latent trait at the second occasion (hereafter Occasion 2), the null and alternative hypotheses are $H_0: \theta_1 = \theta_2$ and $H_A: \theta_1 \neq \theta_2$, respectively (because the AMC method is intra-individual, notation indexing different respondents is not used). To conduct the hypothesis test, the likelihood-ratio test (Agresti, 1996) can be employed; this test has been used previously in the psychometric literature (Klauer & Rettig, 1990; Sinharay, 2017) including in AMC (Finkelman et al., 2010). The logic of this test is to compute the highest value that the joint likelihood of the data (pooling the respondent's answers from Occasion 1 and Occasion 2) can take when the null hypothesis is true (i.e., when θ_1 is constrained to equal θ_2) as well as computing the highest value that the joint likelihood can take when θ_1 and θ_2 are allowed to differ. If the ratio of these two likelihoods is small, that is, if the former joint likelihood is many times lower than the latter, there is strong evidence against the null hypothesis. Mathematically, the following statistic reflecting the above logic is calculated as

$$\Lambda = \frac{\text{maximum likelihood constrained under } H_0}{\text{unconstrained maximum likelihood}}. \tag{1}$$

In order to describe the computation of Equation 1, further notation is required. Let $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ denote the respondent's pooled data from both occasions, with $\boldsymbol{u}_1$ representing the respondent's vector of answers at Occasion 1 and $\boldsymbol{u}_2$ representing the vector of answers at Occasion 2. The pooled maximum likelihood estimate (MLE) under the null hypothesis, $\hat{\theta}_p$, is then defined as the single value of θ that maximizes the likelihood for the joint data $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$. Moreover, the numerator of Equation 1 is the corresponding likelihood (i.e., the likelihood when setting $\theta = \hat{\theta}_p$ and computing the likelihood based on the joint data $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$). To obtain the denominator, the latent trait is estimated separately at Occasion 1 and Occasion 2 using $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. The resulting final MLEs for Occasion 1 and Occasion 2 are denoted $\hat{\theta}_{1,final}$ and $\hat{\theta}_{2,final}$, respectively. Then Equation 1 can be expressed as

$$\Lambda = \frac{L(\hat{\theta}_p | \{\boldsymbol{u}_1, \boldsymbol{u}_2\})}{L(\hat{\theta}_{1,final} | \boldsymbol{u}_1) \times L(\hat{\theta}_{2,final} | \boldsymbol{u}_2)}. \tag{2}$$

To ascertain whether significant change has occurred from Occasion 1 to Occasion 2, the test statistic $-2\log\Lambda$ is computed, and the value of this test statistic is compared to a critical value of the chi-square distribution. H_0 is rejected if $-2\log\Lambda > \chi^2_{1-\alpha}$, where $\chi^2_{1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with one degree of freedom. In the current research, this hypothesis test was conducted at the end of the CAT to evaluate if significant intra-individual change has occurred. Other research has also investigated its application while assessment is underway as part of a variable-length testing paradigm (Finkelman et al., 2010).

Fisher Information

As noted above, CAT involves the selection of items for a respondent based on the answers to previous items, and a traditional procedure is to select the item with the greatest FI at the current estimate of the respondent's latent trait. For a dichotomous item j, the FI at θ is defined as

$$I_j(\theta) = \frac{[P'(\theta)]^2}{P_j(\theta)\left[1 - P_j(\theta)\right]}, \tag{3}$$

where $P_j^{'}(\theta)$ is the derivative of $P_j(\theta)$ with respect to θ. Intuitively, an item with a higher value of $I_j(\theta)$ is better able to differentiate θ from its neighboring values. For the 3-parameter logistic (3PL) model using the scaling parameter $D = 1.7,$ the FI at θ can also be written as

$$I_j(\theta) = 1.7^2 a_j^2 \frac{\left[1 - P_j(\theta)\right]}{P_j(\theta)} \left[\frac{P_j(\theta) - c_j}{1 - c_j}\right]^2 \tag{4}$$

(Lord, 1980).

To use FI as part of an AMC item selection procedure, the MLE $\hat{\theta}_k$ is first computed based on the k previously answered items at the current occasion (for simplicity, the occasion is suppressed in the notation). The value $I_j(\hat{\theta}_k)$ is then calculated for each item eligible to be administered. Finally, the item with the largest value of $I_j(\hat{\theta}_k)$ is administered to the respondent. This process is followed at each occasion.

Fisher Information per Time Unit

The FITU item selection procedure is identical to the procedure described for FI, with one alteration: in the FITU approach, the value $I_j(\hat{\theta}_k)$ is still computed for each eligible item $j,$ but this value is divided by the expected time required to complete the item (Fan et al., 2012). The item maximizing the resulting quotient is then chosen for administration. The logic behind this procedure is that by dividing by an item's expected value of required time, the selected item will be the one that is most time efficient in obtaining information about θ.

In order to operationalize the above procedure, a method for estimating the expected time taken by each item is required. Such a method is provided via response-time modeling. For example, van der Linden (2006) developed a lognormal model to relate a respondent's latent speed (denoted τ) to their response time to item j (denoted T_j). Under this model, the density of T_j given τ is

$$f(t_j \mid \tau) = \frac{\alpha_j}{t_j\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_j\left(\log t_j - \left(\beta_j - \tau\right)\right)\right]^2\right\}. \tag{5}$$

Here t_j denotes a realized value of the random variable T_j, and α_j and β_j are parameters of the item within the response-time model. A larger value of α_j indicates that $\log T_j$ exhibits less dispersion, and a larger value of β_j indicates that $\log T_j$ tends to be larger, on average. Specifically, the distribution of $\log T_j,$ given $\tau,$ is normal with mean $\beta_j - \tau$ and standard deviation α_j^{-1}. Moreover, the expected response time for item $j,$ given $\tau,$ is equal to

$$E(T_j \mid \tau) = \exp(-\tau)\exp(\beta_j + \alpha_j^{-2} / 2) \tag{6}$$

(van der Linden, 2011). Hence, combining Equations 3 and 6, the FI (evaluated at the MLE of θ following the administration of the kth item) per expected unit of time for item j is equal to

$$\frac{I_j(\hat{\theta}_k)}{E(T_j\,|\,\tau)}=\frac{[P_j'(\hat{\theta}_k)]^2 \Big/ \left[P_j(\hat{\theta}_k)(1-P_j(\hat{\theta}_k))\right]}{\exp(-\tau)\exp(\beta_j+\alpha_j^{-2}/2)}. \tag{7}$$

To utilize Equation 7 as an item selection procedure, a specific value of τ (for instance, its MLE) could be input into the term $\exp(-\tau)$ on the right-hand side of Equation 7, thereby providing a single value per item and enabling the items to be rank ordered. However, because $\exp(-\tau)$ does not depend on the item being evaluated (i.e., it is constant across all items j), this term is immaterial to the rank ordering of items. Therefore, item selection can be based on the equivalent criterion

$$C\times\frac{I_j(\hat{\theta}_k)}{E(T_j\,|\,\tau)}=\frac{[P_j'(\hat{\theta}_k)]^2 \Big/ \left[P_j(\hat{\theta}_k)(1-P_j(\hat{\theta}_k))\right]}{\exp(\beta_j+\alpha_j^{-2}/2)}, \tag{8}$$

where $C=\exp(-\tau)$ for an arbitrary τ. That is, the eligible item with the largest value of the right-hand side of Equation 8 is administered to the respondent—a rule that obviates the need for any calculations involving τ and is equivalent to maximizing FITU. As with FI, the FITU item selection procedure described above can be utilized at each occasion in AMC.

It is notable that although van der Linden's (2006) lognormal model is well known in the psychometric literature, it is used only as an example in the present research. The FITU procedure can easily be used with a different response-time model if such a model provides a better fit to the data, whether in the context of educational, psychological, or medical assessment. The expected time taken by an item is simply computed with regard to the model used. Moreover, Cheng, Diao, and Behrens (2017) recently developed a simplified version of the information per time unit approach that does not require fitting a response-time model. In their approach, an item's information is divided by the mean log-transformed response time to the item. Hence, if a response-time model has not been fit for a particular application, the methodology of the current research can still be implemented using Cheng et al.'s approach as long as response-time data are available for each item.

Simulation Design

A simulation study was conducted to compare the FI item selection procedure with the FITU procedure. The two criteria were tested under a variety of conditions. For each item selection procedure, the first item at Occasion 1 was selected assuming that the initial MLE was $\hat{\theta}_0=0$. The first item at Occasion 2 was selected assuming that the initial MLE was equal to the last MLE estimated at Occasion 1 (i.e., $\hat{\theta}_{1,final}$). All MLEs were bounded on the range [–4, 4].

Two hundred eighty-eight items with parameters conforming to the 3PL model were generated based on the methodology of a previous simulation study on AMC (Kim-Kang & Weiss, 2008). To generate their b_j (difficulty) parameters, Kim-Kang and Weiss divided the range [–4.5, 4.5] into 18 consecutive intervals, each of width 0.5. That is, the first interval was [–4.5, –4.0], the second interval was [–4.0, –3.5], and so forth, ending with the interval [4.0, 4.5]. Each of the six middle intervals (from [–1.5, –1.0] to [1.0, 1.5]) contained the b_j parameters of 24 items; each of the other 12 intervals (from [–4.5, –4.0] to [–2.0, –1.5] and from [1.5, 2.0] to [4.0, 4.5]) contained the b_j parameters of 12 items. In the current research, the same approach was taken, with the b_j parameters within each interval following the uniform distribution (with the lower and upper limits of the uniform distribution equal to the lower and upper limits of the given interval). The a_j (discrimination) parameters were randomly assigned to follow a normal distribution with mean 1.5 and standard deviation 0.15, based on one of the conditions of Kim-Kang and Weiss. Finally, as Kim-Kang and Weiss fixed the c_j (pseudo-guessing) parameter at 0.20 for all items, the same procedure was followed here. The scaling parameter $\mathrm{D} = 1.7$ was used.

In addition to the 3PL parameters, each item was randomly assigned parameters (namely, α_j and β_j) corresponding to the response-time model (Equation 5). Following the methodology of Sie, Finkelman, Riley, and Smits (2015), this assignment was conducted in two ways. In the first approach, α_j and β_j were both drawn independently of item j 's 3PL parameters; α_j followed a uniform distribution with a lower limit of 1.0 and an upper limit of 3.0, and β_j followed a uniform distribution with a lower limit of 3.0 and an upper limit of 5.0. When selecting these ranges, Sie et al. referred to van der Linden (2008), who noted that the ranges aligned with the empirical findings of van der Linden (2006) and van der Linden, Breithaupt, Chuah, and Zhang (2007) for these parameters. Items with parameters resulting from this procedure are referred to as "Item Bank 1." In the second approach, the α_j parameters were the same as those from Item Bank 1, but the β_j parameters were assumed to be positively correlated with the b_j parameters of the 3PL model (i.e., the more difficult items were assumed to take longer, on average). In particular, the β_j parameters were drawn under the assumption that b_j and β_j followed a bivariate normal distribution with a correlation of 0.65. Each β_j was randomly sampled from its conditional normal distribution, given b_j, with the marginal distribution of β_j assumed to exhibit the same mean and variance as in Item Bank 1 (which were 4.0 and 1/3, respectively). Items resulting from this second approach are referred to as "Item Bank 2."

Table 1 shows the matrix of θ_1 and θ_2 values that were examined in the study. As can be seen in the table, θ_1 ranged from –2.0 to 2.0. For each value of θ_1, four values of θ_2 were specified: θ_1, $\theta_1 + 0.5$, $\theta_1 + 1.0$, and $\theta_1 + 1.5$. For instance, when θ_1 was set to –2.0, the four values of θ_2 under study were –2.0, –1.5, –1.0, and –0.5. This procedure for defining θ_1 and θ_2 was identical to that of Finkelman et al. (2010) except that the previous research truncated the simulated values of θ_2 at 2.0, whereas the current study allowed values of θ_2 higher than 2.0 whenever such values were prescribed by the methodology outlined above (see Table 1).

Simulees' values of τ were generated in two different ways. In each, the marginal distribution of τ was taken to be normal with a mean of 0.0 and a standard deviation of 0.24 (Sie et al., 2015). In the first approach, values of τ were generated according to the above distribution, assuming that τ was independent of θ_1 and θ_2. The second approach assumed a positive correlation between a simulee's speed and initial ability; in particular, τ and θ_1 were assumed to follow the bivariate normal distribution with a correlation of 0.50 (Wang, Chang, & Douglas, 2013). Each τ was randomly sampled from its conditional normal distribution, given θ_1. In both approaches, each simulee's value of τ was assumed to remain constant between Occasion 1 and Occasion 2 (i.e., once a simulee's τ was generated for Occasion 1, the same value of τ was used for Occasion 2).

Table 1. Combinations of θ_1 and θ_2 Examined in the Simulation Study

	Value of θ_1								
Value of θ_2	–2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
3.5									X
3								X	X
2.5							X	X	X
2.0						X	X	X	X
1.5					X	X	X	X	
1.0				X	X	X	X		
0.5			X	X	X	X			
0.0		X	X	X	X				
-0.5	X	X	X	X					
-1.0	X	X	X						
-1.5	X	X							
-2.0	X								

Note. An "X" indicates that a given combination was examined.

The two item selection procedures under comparison (FI and FITU) were subjected to the same stopping rule. In particular, because the research was conducted assuming a paradigm in which respondent and administrative burden are functions of the time spent under assessment, a time limit was implemented as the stopping rule. That is, at each administration (Occasion 1 and Occasion 2), a given simulee's assessment proceeded until the cumulative time spent on the test reached a specified threshold (which was common to both occasions), upon which the assessment for that occasion was immediately terminated.

Two different time limits were examined. The first time limit was chosen to be approximately equal to the expected time spent by a simulee with $\tau = 0$ on 20 randomly selected items from Item Bank 1; the second time limit was chosen to be approximately equal to the expected time spent by a simulee with $\tau = 0$ on 30 randomly selected items from Item Bank 1. The expected time spent by a simulee with $\tau = 0$ on a randomly chosen item from this bank was calculated to be 72.3; as 20×72.3 = 1446 and 30×72.3 = 2,169, the two time limits were selected to be 1,450 and 2,175. Table 2 summarizes the different conditions of the study that were examined for every combination of θ_1 and θ_2. The table shows three factors (item bank, presence or absence of correlation between θ_1 and τ, and time limit), which were completely crossed. Hence, a total of $2^3 = 8$ conditions were tested for each combination of θ_1 and θ_2. Within every condition and combination of θ_1 and θ_2, responses and response times of 2,500 simulees were generated. For each simulee and occasion, answers to all items were generated according to the 3PL model, and response times for all items

Table 2. Conditions of the Simulation Study for Every Combination of θ_1 and θ_2

Condition	Item Bank	Correlation Between θ_1 and τ	Time Limit
1	1	Yes	1,450
2	1	No	1,450
3	2	Yes	1,450
4	2	No	1,450
5	1	Yes	2,175
6	1	No	2,175
7	2	Yes	2,175
8	2	No	2,175

were generated according to the lognormal model (Equation 5). The items chosen by each item selection procedure for the simulee (and responded to within the time limit) were then determined at each occasion. The simulee's answers to these selected items were subjected to the likelihood-ratio test to determine whether statistically significant change had occurred between the two occasions. The standard significance level of 0.05 was used in the hypothesis test. As in previous research on AMC (Finkelman et al., 2010; Kim-Kang & Weiss, 2008), no item was allowed to be selected twice at a given occasion, but an item could be selected once at Occasion 1 and once at Occasion 2.

The above design served to isolate the effect of the item selection procedure on the classification properties of the likelihood-ratio test for AMC, as all other elements of the simulation were standardized. For every condition and combination of θ_1 and θ_2, the false positive rate (FPR; i.e., "Type I error rate") or true positive rate (TPR; i.e., "power") was calculated for each item selection procedure as coupled with the likelihood-ratio test (the FPR was calculated for all combinations in which $\theta_1 = \theta_2$, and the TPR was calculated for all combinations in which $\theta_1 \neq \theta_2$). Note that for conditions in which $\theta_1 \neq \theta_2$, it was possible for the likelihood-ratio test (which is two-sided) to reject the null hypothesis $\theta_1 = \theta_2$, but in the "incorrect direction." That is, it was possible (albeit unlikely) for an increase in a simulee's true θ to be observed from Occasion 1 to Occasion 2 ($\theta_1 < \theta_2$), but for the likelihood-ratio test to detect significant change in the opposite direction (i.e., to obtain a p-value < 0.05 where the final MLE at Occasion 1 was greater than the final MLE at Occasion 2). In conditions such that $\theta_1 \neq \theta_2$, the TPR of each item selection procedure was calculated in two ways: first, as the proportion of simulees for whom the likelihood-ratio test was significant (regardless of the direction of the difference) and, second, as the proportion of simulees for whom the likelihood-ratio test was significant and the difference in MLEs was in the correct direction (i.e., aligned with the true difference in θ from Occasion 1 to Occasion 2). Results were similar regardless of which definition of TPR was used; therefore, only results corresponding to the second definition are presented. R version 3.5.1 (R Core Team, 2013) was used in the analysis.

Results

The FPR of each item selection procedure, for all combinations of θ_1 and θ_2 for which $\theta_1 = \theta_2$ and all conditions, is presented in Figure 1. Among the total of 72 comparisons of FI and FITU shown in the figure (across all eight panels), FITU exhibited a lower FPR than FI in 66 comparisons (91.7%), FI had a lower FPR in five comparisons (6.9%), and the two procedures had equal FPRs in one comparison (1.4%). The FPR for the FI procedure ranged from 5.1% to 13.7%, with a median of 7.7% and an interquartile range (IQR) of 6.6% to 9.0%. For FITU, the FPR ranged from 4.6% to 7.8%, with a median of 5.8% and an IQR from 5.4% to 6.3%. The difference between the two procedures' values (FPR of FITU – FPR of FI, for a given condition and combination of θ_1 and θ_2) had a minimum of -7.0%, a maximum of 0.6%, and a median of $-$ .7% with an IQR from -2.9% to -0.7% across the 72 comparisons, with negative differences indicating lower (superior) FPRs for FITU. As can be seen in Figure 1, the difference between the procedures tended to be most marked at lower levels of θ.

The TPR of each item selection procedure, for all combinations of θ_1 and θ_2 for all conditions when $\theta_2 = \theta_1 + 0.5$, is presented in Figure 2. The FITU procedure exhibited greater TPR than FI in all 72 comparisons shown in the figure. FITU's TPR ranged from 26.6% to 59.6%, with a median of 46.2% and IQR from 40.2% to 51.4%; FI's TPR ranged from 20.6% to 47.0%, with a median of 33.1% and IQR from 25.5% to 38.0%. The difference between the TPRs of FITU and FI had a minimum of 2.7%, a maximum of 20.0%, and a median of 13.5% (IQR from 10.8% to 15.5%) across the 72 comparisons, with positive differences indicating higher (superior) TPR for FITU. The difference between the procedures tended to be greater at low to moderate levels of θ than at the extremes, with the smallest difference tending to occur at the highest level of θ.

The TPR of both item selection procedures, for all conditions and combinations of θ_1 and θ_2 when $\theta_2 = \theta_1 + 1.0$, is presented in Figure 3. FITU again exhibited larger TPR than FI in all 72 comparisons. The TPR of FITU ranged from 74.6% to 98.8%, with a median of 94.5% (IQR from 89.9% to 96.4%). For FI, the range in TPR was from 50.4% to 95.8%; the median was 77.1% (IQR from 63.9% to 82.8%). The difference in TPR between FITU and FI exhibited a minimum of 1.7%, a maximum of 39.3%, and a median of 16.6% (IQR of 10.7% to 24.0%). Particularly in conditions in which there was a positive correlation between θ_1 and τ (Conditions 1, 3, 5, and 7), the difference between procedures tended to be greater at lower levels of θ than at higher levels (see Figure 3).

Figure 4 displays information analogous to Figures 2 and 3, but for combinations of θ_1 and θ_2 for which $\theta_2 = \theta_1 + 1.5$. As in the previous two figures, FITU had TPR superior to that of FI in all 72 comparisons. For the former procedure, the TPR ranged from 95.2% to 100%, with a median of 99.7% (IQR from 98.9% to 100.0%), whereas for the latter procedure, the TPR ranged from 64.8% to 99.9%, with a median of 91.8% (IQR from 84.0% to 95.9%). The difference in TPR between FITU and FI had a minimum of 0.1%, a maximum of 32.5%, and a median of 8.0% (IQR from 4.0% to 14.4%). As in Figure 3, the difference between procedures tended to be greater at lower levels of θ than at higher levels, particularly in conditions in which there was a positive correlation between θ_1 and τ (Conditions 1, 3, 5, and 7; see Figure 4).

Figure 1. FPR of AMC When Using FI and FITU Item Selection Criteria, by Condition and Value of $\theta_1 = \theta_2$

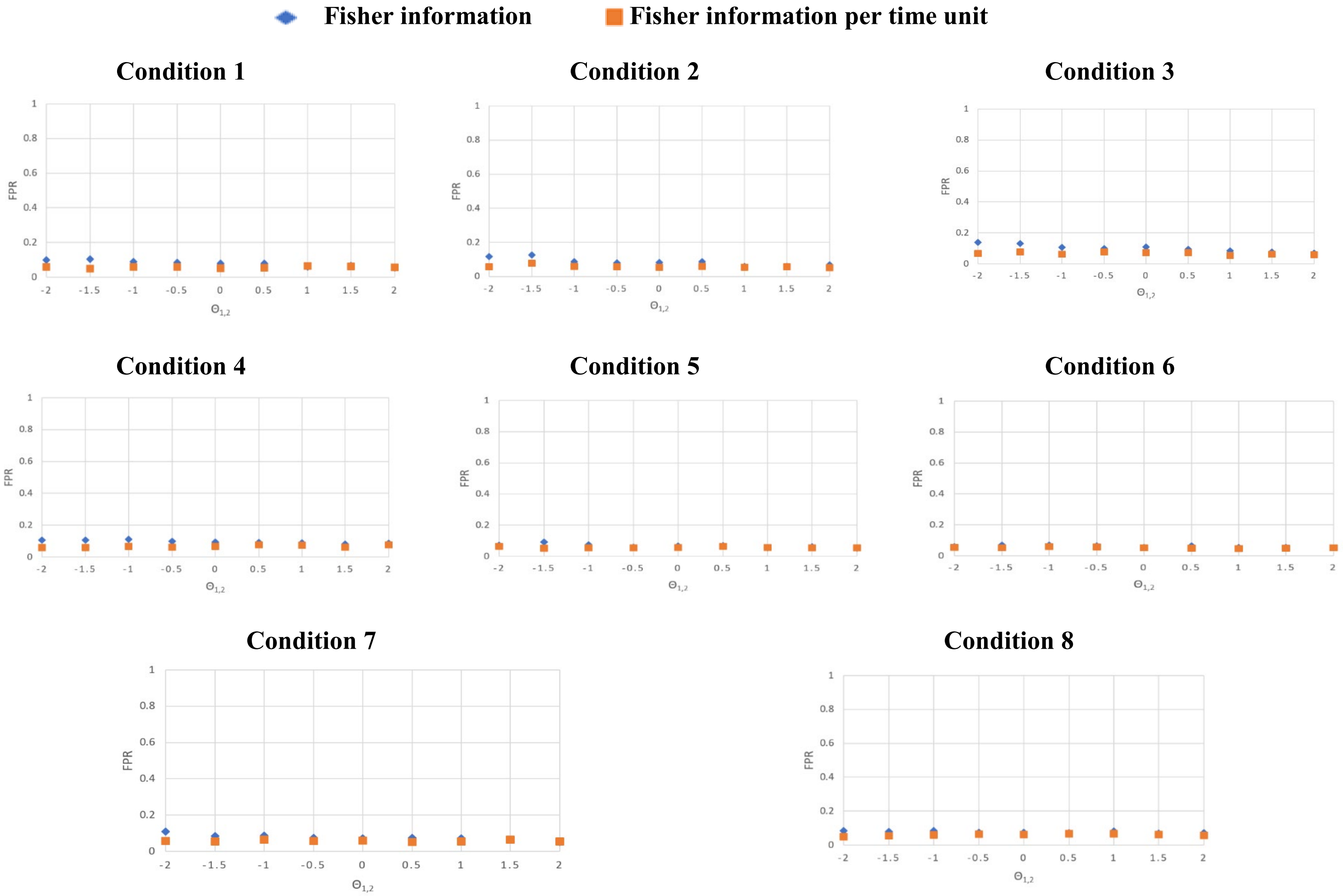


Figure 2. TPR of AMC When Using FI and FITU Item Selection Criteria, by Condition and Value of θ_1 with $\theta_2 = \theta_1 + 0.5$

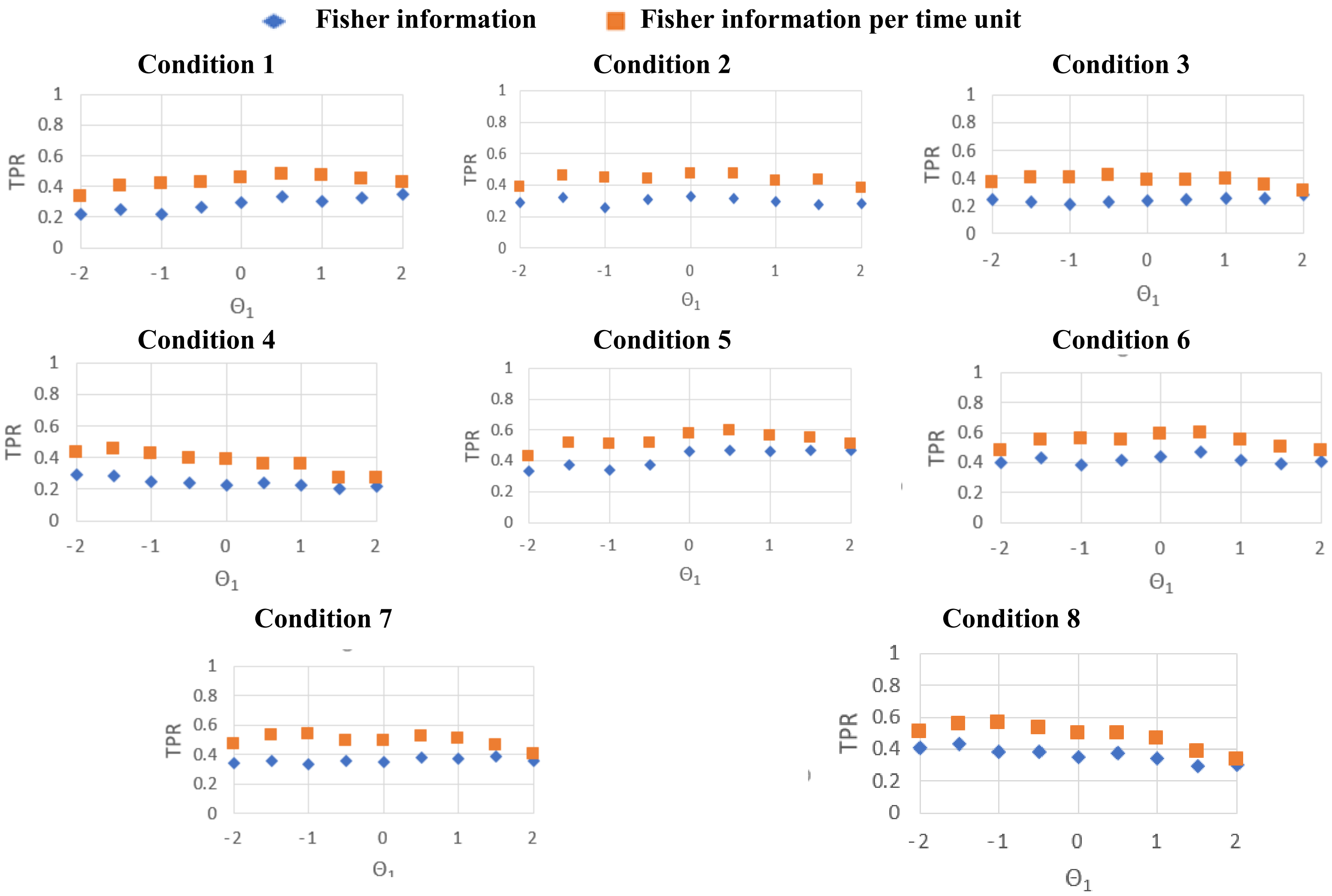
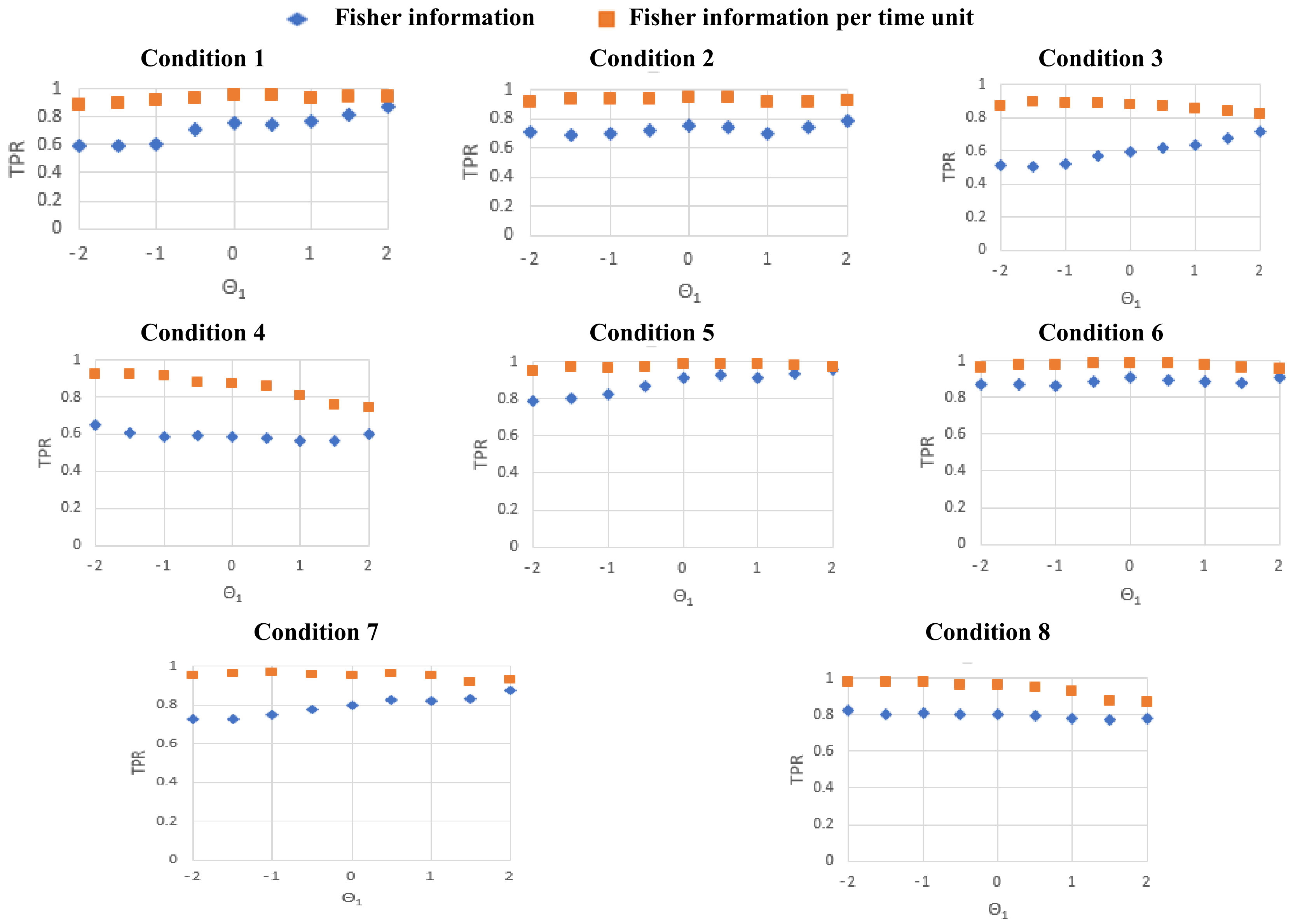


Figure 3. TPR of AMC When Using FI and FITU Item Selection Criteria, by Condition and Value of θ_1 with $\theta_2 = \theta_1 + 1.0$



Condition 1

Condition 2

Condition 3

Condition 4

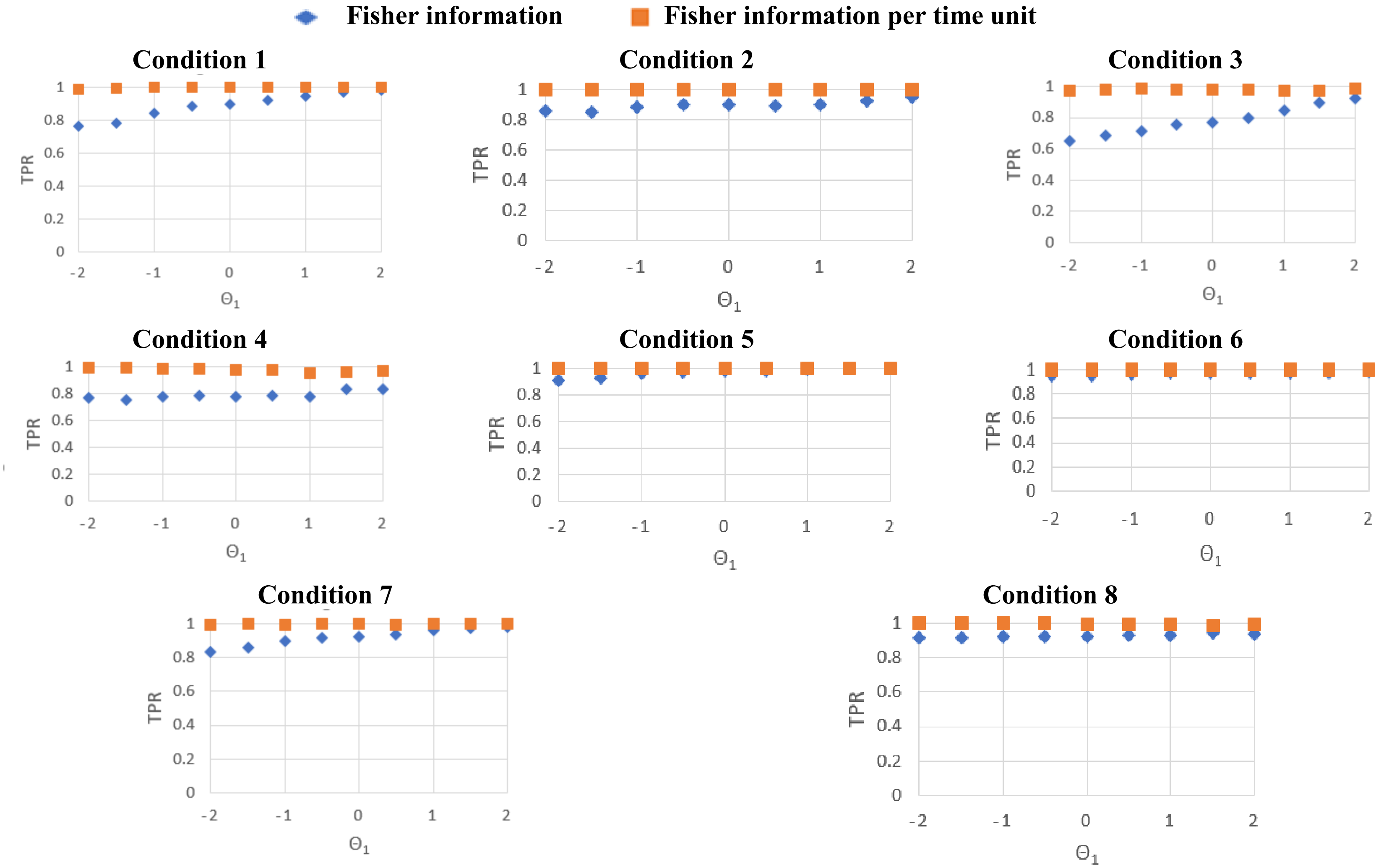
Condition 5

Condition 6

Condition 7

Condition 8

Figure 4. TPR of AMC When Using Fisher Information and Fisher Information per Time Unit Item Selection Criteria, by Condition and Value of θ_1 with $\theta_2 = \theta_1 + 1.5$



Discussion and Conclusions

The goal of this research was to propose the use of the Fisher information per time unit item selection procedure within the context of AMC and compare it to Fisher information item selection in simulation. The former criterion exhibited greater TPR than the latter in all comparisons for which $\theta_1 \neq \theta_2$ (216 comparisons in total when pooling the results of Figures 2 through 4). It also exhibited a lower (superior) FPR in more than 90% of the comparisons in which $\theta_1 = \theta_2$. Perhaps more surprising than the direction of the difference between the two item selection procedures is the magnitude of the difference. For both the set of comparisons in which $\theta_2 = \theta_1 + 0.5$ and those in which $\theta_2 = \theta_1 + 1.0,$ the median improvement in TPR from using Fisher information per time unit was more than 10% (with medians of 13.5% and 16.6%, respectively). For the latter set (in which $\theta_2 = \theta_1 + 1.0),$ the gain in TPR was greater than 10% in over three-quarters of comparisons, was greater than or equal to 24% in one-quarter of comparisons, and reached nearly 40% at its maximum. Regarding comparisons in which $\theta_2 = \theta_1 + 1.5,$ the median improvement in TPR from using Fisher information per time unit was lower, yet still substantial, at 8%, with a gain greater than or equal to 14% in over one-quarter of comparisons and a maximum gain of more than 30%.

In comparisons in which there was large individual change between occasions, the difference between procedures tended to be greater at lower levels of θ than at higher levels in conditions in which there was a positive correlation between θ_1 and τ. This pattern was due, at least in part, to a ceiling effect: The Fisher information per time unit procedure often exhibited TPR close to 100% in cases with large individual change, even in the scenarios where θ was low (for which it was more challenging to achieve high TPR in conditions with a positive correlation between θ_1 and τ because, by definition, such a correlation implied that simulees with lower θ tended to answer items more slowly, and thus answered fewer items within the time limit). As θ increased, the simulees' speed tended to increase, yielding more answered items and therefore greater TPR for the Fisher information criterion, whereas the Fisher information per time unit criterion could not appreciably increase its TPR given that its TPR was already close to 100%.

In sum, the findings suggest the utility of the Fisher information per time unit item selection procedure in AMC contexts in which greater TPR to detect change is desired and time is of the essence. In the fields of medicine and psychology, the importance of the measurement of individual change has been well documented (Brouwer et al., 2013; Hays et al., 2019; Hung et al., 2019; Kashikar-Zuck et al., 2016; Kruyen et al., 2014; Wang & Weiss, 2018), and efficiency might be particularly critical when assessment is performed more than once, given the limited time of respondents and providers (Kruyen et al., 2014; Smits et al., 2012). In education, the determination of a given student's improvement, decline, or stasis over time is fundamental information for the student as well as for parents and teachers (Wang & Weiss, 2018; Weiss & Von Minden, 2011); and the efficiency of assessment is again crucial when, for example, testing is conducted in the classroom setting. The focus of the present study was on the use of AMC with the 3PL model, which is frequently used in educational applications. However, the AMC procedure can be used with any IRT model because all that it requires is MLEs of θ which can be computed for an examinee's responses to any set of items with estimated item parameters at each measurement occasion.

The reason that Fisher information per time unit increased the TPR as compared to only Fisher information is that the former considers the element of time in the item selection process. As noted by Fan et al. (2012), the selection of items that provide high information relative to the amount of time that they take results in greater time efficiency in the collection of information. In particular, compared to an item selection procedure (such as Fisher information) that does not consider the expected time taken by a candidate item, Fisher information per time unit tends to administer more items within a given time frame. Through these additional items, it is able to amass more cumulative information in the same amount of time. As the current study imposed a common time limit for the two item selection procedures at each occasion, the Fisher information per time unit criterion achieved greater total information, through the administration of more items, before time expired at each occasion. This enhanced cumulative information translated into greater TPR.

To perform item selection via Fisher information per time unit, it is necessary to account for the expected time taken on an item. In the current research (as well as previous research outside of the AMC context; Fan et al., 2012), van der Linden's (2006) lognormal model was utilized as an example, for which the denominator of Equation 7 or, equivalently, the denominator of Equation 8 is used to quantify the expected time. The Fisher information per time unit procedure could easily be paired with a different response-time model; the expected time required for an item would merely be calculated with respect to that other model. Response times can follow different distributions in different contexts (Balota & Yap, 2011); for instance, the distribution of response times in psychological assessment settings might differ from the distribution in high-stakes educational testing settings, due to potential differences in the cognitive processes and motivations operating in those respective settings. Even within a given field, the distribution of response times can vary by test and population. Commonly found response-time distributions, other than the lognormal distribution, include the exponential distribution (Scheiblechner, 1979) and the Weibull distribution (Rouder, Sun, Speckman, Lu, & Zhou, 2003), among others.

When a response-time model is used, care should be taken to select a model that exhibits strong empirical evidence of fit in the testing context. When a parametric model does not fit the empirical response-time distribution well, a semi-parametric model that has weaker assumptions could be used instead (e.g., Wang, Chang, & Douglas, 2013; Wang, Fan, Chang, & Douglas, 2013). Also, as mentioned previously, Cheng et al. (2017) provided a version of information per time unit that does not require the fitting of a response-time model, and the time-based item selection procedure discussed in the present study can be used within their framework. As the desire to obtain time-efficient information might be present in educational, psychological, and medical assessment, and respondents might tend to answer some items more quickly than others in any of these contexts (differences in cognitive processes and motivations notwithstanding), Fisher information per time unit might prove to be a valuable tool to gain information quickly in CAT applications including diverse uses of AMC.

The simulation study of the present research included conditions in which the item difficulty parameter was positively correlated with the time-intensity parameter. As pointed out by Cheng et al. (2017), the difficulty parameter is also frequently positively correlated with the discrimination parameter. Hence, the a_j, b_j, and time-intensity parameters might all be positively associated; and in fact, Cheng et al. did find a positive correlation between such parameters empirically. Further simulation studies in which the correlation matrix of these parameters includes only positive values might be realistic and illuminating. In particular, it is well known that the Fisher information item selection procedure tends to select items with high discrimination; however, if such items also exhibit high time intensity, the Fisher information per time unit procedure might select these items less frequently. A comparison of the current simulation study with a study in which all three of the aforementioned parameters are correlated would provide insight into the scenarios in which Fisher information and Fisher information per time unit perform similarly, and those in which they perform differently.

Although the primary objective of the research was not to evaluate the likelihood-ratio test, an important secondary finding was that the FPR of this hypothesis test frequently exceeded the nominal rate of 5% in the context of AMC. In fact, when Fisher information was used as the item selection procedure, the FPR was greater than the desired 5% in all conditions studied, with a median of 7.7%. When Fisher information per time unit was employed, the median FPR was reduced to 5.8%, and sometimes fell to 5% or lower, yet remained greater than its nominal rate in the majority of conditions. Although Fisher information per time unit thus achieved an improvement in the FPR of the likelihood-ratio test, practitioners of AMC should be aware of this test's potential tendency to result in false positives at a proportion above the nominal rate.

One limitation of the study is that it used simulated data that followed the IRT model and response-time model; robustness to model misspecification was not explored. Additionally, as there were a total of 288 combinations of θ_1, θ_2, and the study conditions (9 values of $\theta_1 \times 4$ values of θ_2 per value of θ_1 × the 8 conditions of Table 2), further conditions were not explored. For instance, conditions with content balancing and exposure control were not simulated. The use of these features of CAT is frequently necessary, particularly in the context of high-stakes educational assessment (Leung, Chang, & Hau, 2003). Content balancing might be a key requirement when the same respondent is measured on multiple occasions because of the desire for standardization between occasions. In particular, if a respondent exhibits significant change from Occasion 1 to Occasion 2, it is important to be able to attribute such change to growth (or decline) in that respondent, and not to differences in content between the two test occasions. Further research examining the item selection procedures studied here in conjunction with content balancing (as well as exposure control) would be illuminating.

Also, as in previous AMC research (Finkelman et al., 2010; Kim-Kang & Weiss, 2008), the current study did not examine the case in which an item was barred for selection for a respondent at Occasion 2 if it had been administered to the respondent at Occasion 1. In many assessment settings, it would be necessary in practice to ensure that the same item is not administered to a respondent on more than one occasion. Particularly in high-stakes educational testing (although in other contexts as well), it would be undesirable to present a given item to a respondent more than once (even on different occasions), as the answer at Occasion 2 could be influenced by the respondent's having been administered the item at Occasion 1. For example, if a respondent answered an item incorrectly at Occasion 1, was motivated to find out the answer to that specific item, and then was administered the same item at Occasion 2, the respondent's θ estimate at Occasion 2 could be falsely inflated. Studying different item selection criteria under a constraint requiring each item to be administered a maximum of one time across all testing occasions would be instructive.

Future studies investigating the factors and scenarios listed above are warranted, as are studies using additional item banks and time limits. For example, in the current study, discrimination parameters were generated from the normal distribution with a mean of 1.5 and a standard deviation of 0.15, corresponding to the high discrimination condition of Kim-Kang and Weiss (2008). Further studies could generate the discrimination parameters from the lognormal distribution and/or utilize other means and standard deviations, such as using a mean of 0.5 or 1.0 (corresponding to the low discrimination and medium discrimination conditions, respectively, of Kim-Kang and Weiss). Finally, as in Fan et al. (2012), the estimation of θ was based only on simulees' item responses, not their response times. However, it might be possible under certain circumstances to enhance the precision of θ estimation by incorporating information about an examinee's response times, which might be correlated with ability, into the θ estimate (e.g., van der Linden, Klein Entink, & Fox, 2010). Such "borrowing strength" of information between ability and speed could be combined with the methodology proposed herein for AMC. All of these topics will be studied in future research to further address the goal of obtaining time-efficient information about examinees' individual change.

References

Agresti, A. (1996). *An introduction to categorical data analysis.* New York, NY: John Wiley.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science, 20*(3), 160-166. CrossRef

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison, WI: University of Wisconsin Press.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research, 23*(5), 489-501. CrossRef

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods, 49*(2), 502-512. CrossRef

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin, 74*(1), 68-80. CrossRef

Doucette, A., & Wolf, A. W. (2009). Questioning the measurement precision of psychotherapy research. *Psychotherapy Research, 19(4-5),* 374-389. CrossRef

Dugdale, D. C., Epstein, R., & Pantilat, S. Z. (1999). Time and the patient-physician relationship. *Journal of General Internal Medicine, 14*(Suppl. 1), S34-S40. CrossRef

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement, 32*(3), 277-294. CrossRef

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*(5)*,* 655-670. CrossRef

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34*(4), 238-254. CrossRef

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment, 19(1),* 14-24. CrossRef

Graf, I. (2008). Respondent burden. In Lavrakas, P. J. (Ed.), *Encyclopedia of Survey Research Methods.* Thousand Oaks, CA: SAGE Publications. Retrieved from http://srmo.sagepub.com/view/encyclopedia-of-survey-research-methods/n477.xml.

Hays, R. D., Spritzer, K. L., Sherbourne, C. D., Ryan, G. W., & Coulter, I. D. (2019). Group and individual-level change on health-related quality of life in chiropractic patients with chronic low back or neck pain. *Spine, 44*(9), 647-651. CrossRef

Hung, M., Baumhauer, J. F., Licari, F. W., Voss, M. W., Bounsanga, J., & Saltzman, C. L. (2019). PROMIS and FAAM minimal clinically important differences in foot and ankle orthopedics. *Foot & Ankle International, 40*(1), 65-73. CrossRef

Kashikar-Zuck, S., Carle, A., Barnett, K., Goldschneider, K. R., Sherry, D. D., Mara, C. A., et al. (2016). Longitudinal evaluation of patient-reported outcomes measurement information systems measures in pediatric chronic pain. *Pain, 157*(2), 339-347. CrossRef

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift für Psychologie / Journal of Psychology, 216*(1), 49-58. CrossRef

Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43*(2), 193-206. CrossRef

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2014). Assessing individual change using short tests and questionnaires. *Applied Psychological Measurement, 38*(3), 201-216. CrossRef

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment, 2*(5). Available from https://ejournals.bc.edu/index.php/jtla/article/view/1665

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16,* 421-437. CrossRef

Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement, 18*(3), 437-451. CrossRef

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Phadke, C. (2017). *Measuring intra-individual change at two or more occasions with hypothesis testing methods.* Unpublished doctoral dissertation, University of Minnesota.

R Core Team (2013). R: *A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*(4), 335-343. CrossRef

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*(4), 589–606. CrossRef

Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*(1), 18–38. CrossRef

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement, 39*(5), 389-405. CrossRef

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*(1), 46-68. CrossRef

Smits, N., Zitman, F. G., Cuijpers, P., den Hollander-Gijsman, M. E., & Carlier, I. V. E. (2012). A proof of principle for using adaptive testing in routine outcome monitoring: The efficiency of the Mood and Anxiety Symptoms Questionnaire—Anhedonic Depression CAT. *BMC Medical Research Methodology, 12(4)*. CrossRef van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2)*,* 181-204. CrossRef

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*(1)*,* 5-20. CrossRef

van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement, 35*(3)*,* 183-199. CrossRef

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement, 44*(2)*,* 117-130. CrossRef

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*(5)*,* 327-347. CrossRef

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 144-168. CrossRef

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381-417. CrossRef

Wang, C., & Weiss, D. J. (2018). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement, 42*(3)*,* 221-239. CrossRef

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2)*,* 70-84. CrossRef

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1)*,* 1-23. CrossRef

Weiss, D. J., & Von Minden, S. (2011). Measuring individual growth with conventional and adaptive tests. *Journal of Methods and Measurement in the Social Sciences, 2*(2)*,* 80-101. CrossRef

Welch, R. E., & Frick, T. W. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research and Development, 41*(3)*,* 47-62. CrossRef

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49*(3)*,* 587-602. CrossRef

Author Address

Matthew D. Finkelman, Tufts University School of Dental Medicine, 1 Kneeland St., Boston, MA 02111 U.S.A., matthew.finkelman@tufts.edu.