Behavior Research Methods (2019) 51:1305–1320
https://doi.org/10.3758/s13428-018-1068-x

Comparing computer adaptive testing stopping rules under the generalized partial-credit model

Rose E. Stafford[1] · Christopher R. Runyon[1] · Jodi M. Casabianca[2] · Barbara G. Dodd[1]

Published online: 20 June 2018
© Psychonomic Society, Inc. 2018

Abstract

An important consideration of any computer adaptive testing (CAT) program is the criterion used for ending item administration—the stopping rule, which ensures that all examinees are assessed to the same standard. Although various stopping rules exist, none of them have been compared under the generalized partial-credit model (Muraki in *Applied Psychological Measurement, 16*, 159–176, 1992). In this simulation study we compared the performance of three variable-length stopping rules—standard error (SE), minimum information (MI), and change in theta (CT)—both in isolation and in combination with requirements of minimum and maximum numbers of items, as well as a fixed-length stopping rule. Each stopping rule was examined under two termination criteria—one a more lenient requirement (SE = 0.35, MI = 0.56, CT = 0.05), and one more stringent (SE = 0.30, MI = 0.42, CT = 0.02). The simulation design also included content-balancing and exposure controls, aspects of CAT that have been excluded in previous research comparing variable-length stopping rules. The minimum-information stopping rule produced biased theta estimates and varied greatly in measurement quality across the theta distribution. The absolute-change-in-theta stopping rule had strong performance when paired with a lower criterion and a minimum test length. The standard error stopping rule consistently provided the best balance of measurement precision and operational efficiency and was based on the fewest number of administered items necessary to obtain accurate and precise theta estimates, particularly when it was paired with a maximum-number-of-items stopping rule.

Keywords CAT · Computer adaptive testing · Stopping rules · Termination criteria

Computer adaptive testing (CAT) is a measurement approach that uses item response theory (IRT; Lord & Novick, 1968) to generate tailored tests for examinees in real time on the basis of their responses to previous items (Lord, 1971, 1980). By administering items that are the most informative for an examinee, CATs provide precise measurement of an examinee's proficiency with relatively few items. However, CATs require the consideration of many practical components, making its implementation relatively complicated when compared to paper-and-pencil administrations (Parshall, Spray, Kalohn, & Davey, 2002; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). These required CAT components include: (a) an item pool with known item characteristics, (b) a response model appropriate for the item type, (c) an item selection algorithm, (d) an ability estimation procedure, and (e) some termination criteria to end item administration (Dodd, De Ayala, & Koch, 1995; Reckase, 1989; Weiss & Kingsbury, 1984). In addition, the algorithm used for item selection may be constrained to include content balancing and exposure control mechanisms, particularly in high-stakes testing (Boyd, Dodd, & Choi, 2010). Content-balancing methods ensure that each test administered covers multiple domains according to predetermined specifications. Exposure control procedures are designed to enhance test security by protecting items from overexposure. Both procedures place constraints on maximum information item selection, which consequently increases test length and decreases measurement efficiency (Weiss, 2004). Therefore, CAT components must be carefully selected so that appropriate content coverage and test security are provided while maintaining a psychometrically sound estimate of an examinee's ability (Weiss, 2004).

CATs have two primary advantages over conventional linear test forms, which administer the same items to all examinees. One advantage is that CATs provide precise measurement of all examinees throughout the proficiency range (Lord,

✉ Rose E. Stafford
rose.stafford@utexas.edu

[1] University of Texas at Austin, Austin, TX, USA

[2] Educational Testing Service, Princeton, NJ, USA

1971). This is due to the item selection algorithm used in CATs, which selects the most informative item for an examinee after each item administration. All examinees have an initial ability estimate (usually the population mean) that is updated after the administration of each item. This interim ability estimate is used in the item selection procedure, so that an appropriate item is selected out of the item pool for the next item to be administered. Another benefit of CATs is increased measurement efficiency in comparison to conventional tests. Efficient measurements use the fewest possible items to gain the most information possible about an examinee. The efficiency of CATs is related to the selection of informative items on the basis of an examinee's current ability estimate, such that items that are too easy or too hard are not administered. This item selection procedure has the potential to reduce test length by 80% (De Ayala, 2009).

These two fundamental benefits define the goal of CAT administration—to simultaneously maximize measurement precision and efficiency. Though measurement precision and efficiency are both increased by the information provided by selected items, they are inversely related to one another due to their differing relationships with test length. Although measurement precision increases along with test length, efficiency requires that traits be measured with as few items as possible. Measurement precision and efficiency must be balanced so that only as many items needed to gain a sufficiently reliable estimate of an examinee's proficiency are administered.

Thus, an important consideration in CATs is how many items to administer before estimating the final ability level. This is determined by a termination criterion, or *stopping rule*, which ends each examinee's test once they have been assessed equivalently according to some prespecified standard. There are several stopping rules for CATs, which differ in the criteria used to indicate that an examinee has been measured sufficiently; however, their relative performance under the generalized partial-credit model (GPCM; Muraki, 1992) has not been studied. This simulation study provides a comprehensive examination of the performance of several variable-length stopping rules (i.e., standard error, minimum information, and absolute change in theta estimate) under the GPCM. Importantly, in contrast to previous research examining the performance of stopping rules in CAT, we employ content balancing and item exposure controls to simulate high-stakes testing conditions.

Stopping rules

Stopping rules can be categorized as either fixed-length (FL) or variable-length, which refers to whether test length is equal or varied across examinees. FL stopping rules are the most straightforward as they end an examinee's test after a predetermined number of items are administered, regardless of their ability. However, using a FL test results in a lack of consistent measurement precision (i.e., the standard error of the θ ability estimate) for all examinees across the range of abilities (Leroux & Dodd, 2014). These imprecise ability estimates are problematic, especially in high-stakes testing (e.g., licensure or certification testing) in which they can have detrimental implications. Although administering a greater number of items can provide more precise ability estimates, doing so decreases the efficiency of using CAT and can lead to item overexposure. Similarly, some examinees may be measured with high precision after responding to only a few items, so administering additional items unnecessarily exposes items to examinees and increases examinee burden.

Variable-length stopping rules are designed to provide equivalent measurement precision across examinees by ending item administration after a prespecified measurement standard has been satisfied. These tests are of variable length because examinees may take a different number of items before the criterion for test termination is met. Researchers have developed several variable-length stopping rules, which all aim to administer as few items as needed to obtain a psychometrically sound estimate of an examinee's ability but differ in the criteria used to indicate that an examinee's ability has been measured adequately. One variable-length stopping rule is the *standard error* (SE) stopping rule, which terminates item administration when a prespecified standard error of the present ability estimate has been reached (Dodd, Koch, & De Ayala, 1989). After each interim ability level has been estimated, the standard error associated with the examinee's current ability estimate is evaluated. If this standard error is below some prespecified value, then item administration ends. If not, item administration continues until the standard error associated with the interim ability estimate falls below the criterion value. Once item administration terminates, the most recent interim ability estimate becomes the final ability estimate.

Rather than evaluating the precision of the ability estimate of the examinee, the *minimum-information* (MI; Dodd et al., 1989) stopping rule focuses on the quality of the items in terms of the information they provide. This stopping rule determines when a test is completed by evaluating the information of the available items remaining in the pool after each item is administered. If eligible items in the item pool provide some sufficiently high level of information on the basis of the interim ability estimate, then item administration continues. When the information of the remaining items falls below the specified level, item administration ends.

A more recently developed stopping rule is the *absolute-change-in-theta* (CT) stopping rule. This variable-length stopping rule regulates test length using the absolute change in an examinee's theta estimate ($\hat{\theta}$) after an item is administered (Babcock & Weiss, 2012). During a CAT administration, an examinee's $\hat{\theta}$ generally changes with each additional item administered, though the size of this change lessens as the number of administered items increases. The CT stopping rule

evaluates the absolute change in the $\hat{\theta}$ after each item is administered to an examinee. When this expected change falls below a specified value, item administration ends.

FL and variable-length stopping rules are often combined, such that a variable-length termination criterion is used as the stopping criteria until the examinee is administered a maximum number of items. Using a maximum test length in addition to a variable-length stopping rule is beneficial when there is a mismatch between the item pool distribution and the examinee's ability level. In consequence, the examinee could be administered all items in the item pool because no items remain that can satisfy the variable-length stopping criterion. Thus, in order to keep CATs efficient, as well as reduce examinee burden and item exposure, a secondary FL termination criterion is often used to stop item administration after a certain number of items.

Performance of CAT stopping rules

Previous research has examined the performance of different stopping rules for polytomous IRT models. The performance of the stopping rule can differ on the basis of its interaction with other aspects of CAT administration, such as item pool characteristics, the match of the distribution of the item pool shape to the examinees, and whether item are dichotomously or polytomously scored (Boyd et al., 2010). CAT has greater measurement efficiency when used with polytomous items because each response category within an item provides additional information (Dodd et al., 1995). Because of the greater amount of information polytomous items provide, fewer polytomous items are needed to obtain the same level of measurement precision as dichotomous items.

Much of the previous research on stopping rules with polytomously scored items has examined the performance of the SE and MI stopping rules (with FL as a secondary termination criterion). Dodd, Koch, and De Ayala (1989) examined these stopping rules with the graded-response model (GRM; Samejima, 1969), the partial-credit model (PCM; Masters, 1982) in later research (Dodd, Koch, & De Ayala, 1993), and Dodd (1990) used the Andrich's Rating Scale Model (RSM; Andrich, 1978). In these studies the SE stopping rule generally outperformed the MI stopping rule—fewer items were administered, the correlations between known and estimated ability levels were higher, and fewer nonconvergent cases resulted. The only simulation conditions in which the MI stopping rule had superior performance to the SE rule were those in which the information distribution of the item pool did not align with the trait distribution of the test taker population. This was due to the SE stopping rule administering more items than the MI stopping rule to examinees with ability levels that were far from the peak of the item pool shapes.

More recently, Babcock and Weiss (2012) examined the performance of 14 different stopping rules on dichotomously scored items using the three-parameter logistic (3PL) IRT model. The researchers examined various cutoff values for the SE stopping rule, the MI stopping rule, combinations of SE and MI stopping rules, the CT stopping rule, and several fixed-length stopping rules. The authors concluded that the SE stopping rule works well in most cases, but recommended including a minimum number of items or using it in combination with an MI stopping rule. The information structure of the item bank had less impact on the performance of the CT stopping rule in comparison to other termination methods, meaning the CT rule might have increased utility when the information distribution of an item bank does not cover the range of examinee abilities in a sample (Babcock & Weiss, 2012).

Content balancing and exposure control

Exposure control and content balancing methods are commonly implemented in large-scale high-stakes testing scenarios, which often require that test-takers be measured across multiple domains within a single ability continuum, such as addition, subtraction, multiplication, and division within an arithmetic achievement assessment (Weiss, 2004). Content balancing methods are implemented to ensure that examinees answer a sufficient number of items from each domain so that items are equally distributed across the domains for all examinees, meaning that their ability estimates will be determined from similar content.

Large-scale assessments frequently use content balancing concurrently with an exposure control procedure. During CAT administration, different items in the item bank will naturally be used at different rates. Reducing item exposure rates is important in high-stakes testing scenarios when there is incentive for examinees to have access to items prior to testing. Exposure control methods improve test security by controlling the exposure rate of items across a group of examinees. Exposure control and content balancing methods are both implemented by modifying the maximum information item selection procedure that places constraints on what items can be administered (Weiss, 2004). The incorporation of these constraints can decrease measurement precision (Davis, 2004; Davis & Dodd, 2003). Implementing exposure control and content balancing may also increase the test length and reduce the efficiency of CATs, as more items are required to obtain the prespecified degree of precision specified by variable-length stopping rules (Leroux & Dodd, 2016; Leroux, Lopez, Hembry, & Dodd., 2013).

Despite the wide use of content balancing and exposure control methods in large-scale assessments, no prior research has been conducted that compares variable-length stopping rules under their implementation. CATs incorporating both

exposure control and content balancing have been studied with fixed-length tests (Boyd, Dodd, & Fitzpatrick, 2013). There has also been research comparing exposure control methods using content balancing for dichotomous (i.e., Leroux et al., 2013) and polytomous (i.e., Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003; Leroux & Dodd, 2016) IRT models that included separate FL and SEFL stopping rule conditions but did not compare variable-length stopping rules. Variable-length stopping rules have only been compared without the use of content balancing or exposure control (e.g., Choi, Grady, & Dodd, 2010; Dodd et al., 1989, 1993; Leroux & Dodd, 2014). Due to the relationship between stopping rules and exposure control and content balancing mechanisms, it is important to extend previous research of CAT stopping rules to the frequently encountered scenarios in which these constraints are implemented.

Purpose of study

The CT stopping rule is a recently developed approach to CAT termination, which has demonstrated strong performance under a dichotomous IRT model (3PL; Babcock & Weiss, 2012), but the generalization of this finding is limited to the specific set of conditions previously researched. The present study continues this examination of the performance of SE, MI, and CT stopping rules by extending it to a scoring model for polytomous items, the GPCM (Muraki, 1992). Furthermore, we utilize exposure control and content balancing in our study to mirror conditions in high-stakes testing, as previous research examining variable-length stopping rules has excluded both of these CAT components and stopping rules may be sensitive to their use.

Method

Study design overview

In the present study, we compare variable-length CAT stopping rules under the GPCM (Muraki, 1992) as shown in Eq. 1. The GPCM models the probability of scoring in category x on item i out of $m_i + 1$ response categories for an individual with a given trait level, θ as:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x} a_i(\theta - b_{ij})\right]}{\sum_{r=0}^{m_i}\exp\left[\sum_{j=0}^{r} a_i(\theta - b_{ij})\right]} \quad (1)$$

where a_i is the discrimination or slope of the item, b_{ij} is the step difficulty parameter associated with score category j (j = 1, . . . , m_i), and $\Sigma\, a_i\,(\theta - b_{ij}) = 0$ when $j = 0$. Item discrimination, a_i, varies across items but not within items across categories. The GPCM requires that the steps within an item be completed in order, though the step difficulties of the ordered categories, b_{ij}, are not required to be in sequential order.

We examined 21 different stopping rules, which consisted of variations of SE, MI, and CT stopping rules, as well as their combination with a FL stopping rule. These are discussed in detail in a following section. This study used a repeated measures design in which each simulated examinee was administered the CAT 21 times, once using each stopping rule. All data generating procedures and analyses were conducted in SAS statistical software (version 9.4 for Windows).

Item pool description and data generation

The item pool for this study was generated using the item parameter values of a large-scale educational assessment previously calibrated using the GPCM (Davis, 2004). The means (*SD*, minimum, maximum) of the item discrimination (a) and step difficulty ($b_{1–4}$) parameters were as follows: $a = 0.92$ ($SD = 0.19$, 0.54, 1.52), $b_1 = -0.99$ (0.90, – 3.13, 1.50), $b_2 = 0.18$ (0.99, – 1.81, 3.57), $b_3 = -0.19$ (0.76, – 1.48, 1.51), $b_4 = -0.12$ (0.90, 2.36, 2.34). The pool consisted of 157 polytomously scored items with varying numbers of response categories—99 items with three response categories, 29 items with four categories, and 29 items with five categories. Each item assessed one of three content areas—61 items assessed content area A, 59 items assessed B, and 37 items assessed C. The numbers and proportions of items assessing each content area by the number of item categories are presented in Table 1. Item and test information based on the GPCM was calculated using the SAS macro IRTINFO (Fitzpatrick, Choi, Chen, Hou, & Dodd, 1994). Figure 1 shows the information function of the 157-item pool, which indicates adequate information coverage across the range of θ values, and maximal information at θ = – 0.6.

Item responses for a sample of 1,000 simulees were generated from the true generating item parameter values using the

Table 1 Numbers and proportions of items by number of categories and content area

Content Area	3 Categories		4 Categories		5 Categories	
	N	*P*	*N*	*P*	*N*	*P*
A	42	.27	10	.06	9	.06
B	42	.27	6	.04	11	.07
C	15	.10	13	.08	9	.06

P = proportion of items fitting content and category criteria out of 157-item pool. Rounded proportions are presented, but exact proportions were used as the target proportions in content balancing procedure in CAT implementation

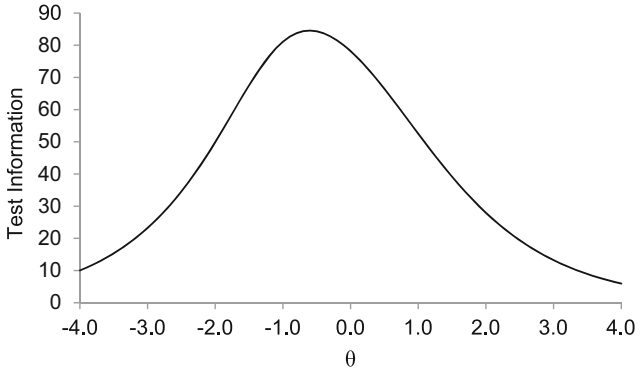

Fig. 1 Item pool information curve

SAS macro IRTGEN (Whittaker, Fitzpatrick, Williams, & Dodd, 2003). The true θ levels for each sample of simulees were drawn from a normal distribution with a mean of 0 and a standard deviation of 1. There were 1,000 replications used in this simulation study, and thus, each of the 1,000 generated datasets (of N = 1,000 simulees) underwent 21 CATs (with different stopping rules), for a grand total of 21,000 CAT simulations.

CAT procedure

CATs were administered to the simulees in SAS software using a program similar to that of the commercially available software (SIMPOLYCAT; Chen & Cook, 2009). We used a modified version of a SAS program that simulated CAT implementation according to the GPCM (Davis, 2004), which was altered to implement the 21 stopping rule conditions. For each CAT administration, all simulees started with an initial θ estimate of zero. A variable step size procedure (Koch & Dodd, 1989) was used to estimate θ until the simulee made responses in two different categories, after which point maximum likelihood estimation (MLE) was used to obtain ability estimates. The variable step size approach changed the $\hat{\theta}$ by half the distance between the initial $\hat{\theta}$ and either of the two extreme step-difficulty parameter estimates depending on whether the response to the previous item was in the lower or upper half of the response scale. Only items that met the content balancing were used for the variable step size.

Content balancing was employed using Kingsbury and Zara's (1989) constrained CAT (CCAT) content-balancing method. The goal of this procedure is to have the proportions of items that an examinee answers closely match the prespecified desired proportions of each content area (i.e., target proportions). After an item was administered to a simulee, the proportion of items answered in each content area out of all those administered was computed, and then the discrepancy between the true proportions and the target proportions was calculated. The content area with the largest discrepancy was selected for the subsequent item administration. Two item characteristics were used to define the areas and their target proportions—content area and the number of response categories. The combination of these two factors stratified the item pool into nine target areas. The target proportion for each of the nine item types was set equal to the proportion of items in the 157-item pool that belonged to that combination of the content area and scale length factors, which are presented in Table 1. Item administration then proceeded using the CCAT procedure.

The randomesque exposure control procedure (Kingsbury & Zara, 1989) was also incorporated into item selection. In this method, the item administered is randomly selected from a group of items that are the most informative given an examinee's current $\hat{\theta}$. This study used a group size of three, which provides high measurement precision and exposure control (Davis, 2004). Content balancing was given precedence over exposure control. Therefore, the content area of the item to be administered to a simulee was first identified. Then, the three most informative items in that content area based on the simulee's current $\hat{\theta}$ were identified. The CAT program then randomly selected one of those three items and administered it to the simulee. The program then calculated the discrepancy in actual and target content area proportions to determine the next content area and selected an informative item for the simulee using the randomesque procedure based on their updated $\hat{\theta}$. This item selection and administration process continued until the stopping rule was satisfied and the CAT program terminated, giving the simulee a final θ estimate for each of the 21 stopping rule conditions.

Implemented stopping rules

We investigated an FL stopping rule and several variations of three variable-length stopping rules: (a) SE, (b) MI, and (c) CT (Table 2). The FL stopping rule administered 20 items to all simulees. This is a sufficient test length for obtaining an accurate estimate of θ, as was indicated by prior simulation research using the same item pool (Davis, 2004; Leroux & Dodd, 2016) and across CAT research in general (Dodd, 1990; Dodd et al., 1989, 1993; Gorin, Dodd, Fitzpatrick, & Shieh, 2005; Koch & Dodd, 1989; Lee & Dodd, 2012).

To increase the generalizability of our findings, each variable-length stopping rule was implemented twice using different prespecified criterion values for the stopping rule. The SE stopping rule ended a test when the SE of the simulee's $\hat{\theta}$ was less than the criterion value—being either 0.30 (in SE[.30] conditions) or 0.35 (in SE[.35] conditions), which are both commonly used SE criteria (e.g., Dodd, 1990; Dodd et al., 1993; Leroux & Dodd, 2014; Leroux et al., 2013). Equivalent SE and MI criteria were used to aid in comparisons between these stopping rules. MI criterion values were selected using the well-known relationship between the standard

Table 2 Summary of stopping rule conditions

Stopping Rule	Description
FL	Fixed length of 20 items
SE[.30]	$SE_{\hat{\theta}}$ below 0.30
SEFL[.30]	$SE_{\hat{\theta}}$ below 0.30 or a maximum of 20 items
SE[.35]	$SE_{\hat{\theta}}$ below 0.35
SEFL[.35]	$SE_{\hat{\theta}}$ below 0.35 or a maximum of 20 items
MI[.42]	Information of nonadministered items below 0.42
MI9[.42]	Information of nonadministered items below 0.42, with a minimum of nine items
MIFL[.42]	Information of nonadministered items below 0.42 or a maximum of 20 items
MIFL9[.42]	Information of nonadministered items below 0.42 or a maximum of 20 items, with a minimum of nine items
MI[.56]	Information of nonadministered items below 0.56
MI9[.56]	Information of nonadministered items below 0.56, with a minimum of nine items
MIFL[.56]	Information of nonadministered items below 0.56 or a maximum of 20 items
MIFL9[.56]	Information of nonadministered items below 0.56 or a maximum of 20 items, with a minimum of nine items
CT[.02]	Absolute change in $\hat{\theta}$ less than 0.02
CT9[.02]	Absolute change in $\hat{\theta}$ less than 0.02, with a minimum of nine items
CTFL[.02]	Absolute change in $\hat{\theta}$ less than 0.02 or a maximum of 20 items
CTFL9[.02]	Absolute change in $\hat{\theta}$ less than 0.02 or a maximum of 20 items, with a minimum of nine items
CT[.05]	Absolute change in $\hat{\theta}$ less than 0.05
CT9[.05]	Absolute change in $\hat{\theta}$ less than 0.05, with a minimum of nine items
CTFL[.05]	Absolute change in $\hat{\theta}$ less than 0.05 or a maximum of 20 items
CTFL9[.05]	Absolute change in $\hat{\theta}$ less than 0.05 or a maximum of 20 items, with a minimum of nine items

error of θ and the total information, specifically that the SE is equal to the inverse of the square root of the information for a given θ. The MI stopping rule ended the CAT program when all nonadministered items had Fisher information less than either 0.42 (MI[.42] conditions) or 0.56 (MI[.56] conditions) for the simulee's current $\hat{\theta}$, which are equivalent to SE[.35] and SE[.30] conditions, respectively. Finally, the CT stopping rule ended a simulee's test when the absolute change in the simulee's $\hat{\theta}$ was less than either 0.02 (CT[.02] conditions) or 0.05 (CT[.05] conditions). These values were selected due to their usage in previous CT research (i.e., Babcock & Weiss, 2012) to aid in cross-study comparisons. The SE[.30], MI[.42], and CT[.02] represented more *stringent* criteria, in that they required more items in order to be satisfied in comparison to *lenient* criteria (SE[.35], MI[.56], CT[.05]), which produced a relatively shorter test.

Variable-length stopping rules were studied both under isolation and in combination with the FL stopping rule (i.e., a maximum number of items). When used in isolation, as previously described, the CAT program continued until the termination criteria was met or until no items remained. Variable-length stopping rules are frequently paired with a maximum number of items in real-world CAT applications to prevent the administration of more items than is necessary to obtain estimates with high measurement precision (Boyd et al., 2010; Dodd et al., 1995). Additional conditions were included that combined each variable-length stopping rule and criterion value with a FL maximum of 20 items (SEFL[.30], SEFL[.35], MIFL[.42], MIFL[.56], CTFL[.02], and CTFL[.05]), which ended the CAT program when either the variable-length criteria were reached or 20 items had been administered.

Preliminary CAT trials revealed that conditions using MI and CT stopping rules had high rates of nonconvergent cases, particularly in the MI conditions. We found that these nonconverged cases had usually only answered an average of four items, indicating that the criteria for these stopping rules were being satisfied before the program could obtain an acceptable θ estimate. Therefore, we included additional conditions that had the requirement that at least nine items be administered (MI9[.42], MIFL9[.42], MI9[.56], MIFL9[.56], CT9[.02], CTFL9[.02], CT9[.05], and CTFL9[.05]). Nine was selected as the minimum test length because it was equal to the number of content areas, meaning that all content areas could potentially be covered before termination of a test. SE stopping rules were not studied using a minimum test length because the SE stopping rules were already delivering at least

nine items across all simulees and replications. The addition of these eight minimum test length conditions produced a total of 21 stopping rule conditions.

Data analyses

We compared CAT stopping rules using several criteria that are important in adaptive testing. We recorded the number of nonconvergent cases and calculated descriptive statistics of their frequency for each stopping rule. Only the estimates of converged cases were used in the following analyses. We also examined summary statistics of final trait estimates ($\hat{\theta}$), the standard error of the trait estimate ($SE_{\hat{\theta}}$), and the number of items administered. The descriptive statistics for these criteria were calculated by finding their average within each replication across simulees, and then calculating the minimum, maximum and mean (i.e., grand mean) of these averages across the 1,000 replications of each stopping rule condition. We computed descriptive statistics for the Pearson correlations between true and estimated θ values, bias, and root mean square error (RMSE). Bias and RMSE were calculated using the following formulas:

$$\text{Bias} = \frac{\sum_{k=1}^{n}\left(\hat{\theta}_k - \theta_k\right)}{n} \tag{2}$$

and

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{n}\left(\hat{\theta}_k - \theta_k\right)^2}{n}}, \tag{3}$$

where $\hat{\theta}_k$ is the estimated trait level for simulee k, θ_k is the known trait level for simulee k, and n is the total number of simulees. In addition, we examined plots of test length, $SE_{\hat{\theta}}$, bias, and RMSE conditional on θ to detect whether the stopping rules differed in their parameter recovery depending on a simulee's true θ. The values on these plots were created by grouping simulees into 13 groups along the continuum of known θ from – 3 to + 3 and plotting the average test length, $SE_{\hat{\theta}}$, bias, and RMSE for each θ group.

We also evaluated the stopping rules in relation to the exposure control and content balancing constraints imposed across all conditions. Minimum, maximum, and mean item exposure rates were calculated for each item by dividing the number of times an item was administered by the total number of simulees. These were averaged across the 157-item pool to evaluate the relative exposure rates of each stopping rule. Pool utilization was examined by the percentage of items that were never administered to each replicated sample of 1,000 simulees. We report the minimum, maximum, and mean percentage of items not administered across replications. In addition, we calculated the differences between the proportions of items administered in each content area and their targeted content area proportions to examine adherence to content balancing constraints across stopping rule conditions.

Results

Number of nonconvergent cases

Table 3 presents descriptive statistics for the frequency of nonconvergent cases across the 1,000 replications . The lowest nonconvergence rates occurred when using a fixed-length test of 20 items (FL) and in conditions using SE-based stopping rules (i.e., SE and SEFL), where nonconvergence occurred in an average of 1.5% of simulees. As was previously noted, stopping criteria that relied on minimum information (MI and MIFL) or absolute change in $\hat{\theta}$ (CT and CTFL) produced excessive numbers of inestimable traits when a criterion for minimum number of items administered was not included. The MI and MIFL conditions had the greatest number of nonconvergent cases, with an average of 6.8% of simulees when using information of 0.42, and an average of 11.4% when using 0.56 as the minimum information value. Investigation of these nonconverging cases revealed that the majority of these simulees were only administered three or four items and that maximum likelihood estimation had never been reached.

The average numbers of nonconvergent cases for CT and CTFL were less than half those for the MI and MIFL conditions, though they were still higher than would usually be expected, at about 2.8% of simulees for both 0.02 and 0.05 change in $\hat{\theta}$ conditions. Nonconvergent cases in the CT and CTFL conditions usually answered five to six items before the CAT program ended, but all of the simulees' responses were in either the highest or the lowest response category, meaning that MLE could not be used (indicated by $\hat{\theta} \leq -4$ or $\hat{\theta} \geq 4$). The addition of a requirement that a minimum of nine items be administered mitigated these convergence issues, reducing the number of nonconvergent cases to an average of 1.8% of simulees in these conditions (i.e., MI9, MIFL9, CT9, and CTFL9).

Trait estimation and number of items administered

Table 3 also presents descriptive statistics for the averages of trait estimates ($\hat{\theta}$), standard errors of trait estimates ($SE_{\hat{\theta}}$), and numbers of items administered (NIA) by CATs for each stopping rule. The grand means of $\hat{\theta}$ were uniformly close to 0 across conditions, though the use of MI stopping rules tended to overestimate θ, particularly when using a more lenient MI value (MI = 0.56) and not including a minimum NIA requirement ($\hat{\theta}$ = 0.08). As would be expected, when more stringent values were used for the stopping rules—meaning when MI

Table 3 Descriptive statistics of trait estimation and numbers of items administered

Stopping Rule	*N* NonCon Mean (Min, Max)	$\hat{\theta}$ Grand Mean (Min, Max)	$SE_{\hat{\theta}}$ Grand Mean (Min, Max)	NIA Grand Mean (Min, Max)
FL	15.37 (5, 29)	0.01 (– 0.09, 0.12)	0.29 (0.28, 0.29)	20.00 (20.00, 20.00)
SE[.30]	14.62 (3, 29)	0.00 (– 0.10, 0.09)	0.30 (0.29, 0.30)	19.42 (18.50, 20.51)
SE[.35]	14.79 (3, 29)	– 0.01 (– 0.11, 0.10)	0.34 (0.34, 0.34)	14.03 (13.47, 14.85)
SEFL[.30]	15.32 (5, 29)	0.01 (– 0.09, 0.12)	0.31 (0.31, 0.31)	16.92 (16.70, 17.21)
SEFL[.35]	15.38 (3, 29)	0.00 (– 0.10, 0.10)	0.35 (0.34, 0.35)	13.15 (12.84, 13.59)
MI[.42]	68.05 (47, 96)	0.04 (– 0.09, 0.20)	0.34 (0.31, 0.38)	45.79 (42.64, 48.69)
MI[.56]	114.39 (86, 153)	0.08 (– 0.04, 0.23)	0.51 (0.48, 0.55)	15.94 (14.54, 17.20)
MI9[.42]	18.17 (5, 34)	0.04 (– 0.08, 0.16)	0.24 (0.22, 0.25)	51.13 (48.86, 53.72)
MI9[.56]	18.41 (6, 34)	0.04 (– 0.07, 0.16)	0.31 (0.30, 0.33)	21.91 (20.93, 22.82)
MIFL[.42]	68.12 (48, 91)	0.04 (– 0.07, 0.18)	0.41 (0.39, 0.44)	16.13 (15.50, 16.69)
MIFL[.56]	114.39 (85, 152)	0.08 (– 0.06, 0.20)	0.53 (0.49, 0.56)	12.12 (11.32, 12.87)
MIFL9[.42]	18.21 (6, 33)	0.03 (– 0.08, 0.13)	0.31 (0.30, 0.33)	18.49 (18.10, 18.81)
MIFL9[.56]	18.24 (6, 35)	0.03 (– 0.07, 0.16)	0.33 (0.32, 0.35)	16.84 (16.34, 17.23)
CT[.02]	28.20 (13, 50)	– 0.02 (– 0.13, 0.07)	0.37 (0.36, 0.38)	14.02 (13.30, 14.83)
CT[.05]	28.57 (13, 42)	– 0.03 (– 0.16, 0.08)	0.46 (0.45, 0.47)	8.96 (8.49, 9.36)
CT9[.02]	18.82 (7, 34)	0.00 (– 0.11, 0.10)	0.32 (0.32, 0.33)	16.33 (15.64, 17.12)
CT9[.05]	18.97 (4, 34)	0.01 (– 0.09, 0.11)	0.37 (0.36, 0.37)	12.13 (11.81, 12.49)
CTFL[.02]	27.83 (13, 45)	– 0.02 (– 0.12, 0.08)	0.38 (0.37, 0.39)	12.70 (12.15, 13.16)
CTFL[.05]	28.74 (14, 45)	– 0.03 (– 0.14, 0.09)	0.46 (0.45, 0.47)	8.91 (8.44, 9.35)
CTFL9[.02]	18.60 (6, 33)	0.00 (– 0.09, 0.12)	0.33 (0.33, 0.34)	14.74 (14.39, 15.10)
CTFL9[.05]	18.77 (5, 32)	0.01 (– 0.09, 0.13)	0.37 (0.36, 0.37)	12.05 (11.76, 12.44)

N NonCon = number of nonconvergent cases out of 1,000 simulees. The mean of *N* NonCon is given across 1,000 replications. NIA = number of items administered

was higher (MI = 0.42), CT was lower (CT = 0.02), or SE was lower (SE = 0.30)—the average NIA was higher and $SE_{\hat{\theta}}$ was lower than we observed under the comparative stopping rule condition using a more lenient value (i.e., MI = 0.56, CT = 0.05, SE = 0.35). The results also indicate that including a fixed-length component to a variable-length stopping rule successfully decreased test length, most dramatically so in MI conditions using a 0.42 information value, due to these conditions having a large drop in NIA (MIFL[.42], NIA = 16.13; MIFL9[.42], NIA = 18.49) from very high values when a maximum of 20 items was not in place (MI[.42], NIA = 45.79; MI9[.42], NIA = 51.13). Except for these two conditions, attaching a restriction of a maximum of 20 items to a variable-length stopping rule only increased $SE_{\hat{\theta}}$ by about 0.01.

Figure 2 depicts the average NIA (i.e., test length) for each variable-length stopping rule, conditional on the simulees' known θ. The MI stopping rules administer the largest number of items to simulees with θs in the center of the distribution, where the majority of informative items exist. The SE stopping rules display the opposite behavior, giving more items to simulees with extreme θs, whereas those in the center of the distribution are measured precisely with fewer items. The addition of a fixed-length component curtails these tendencies for delivering long tests for MI and SE stopping rules, though the CT stopping rule performs similarly with and without an FL component. When using a CT criterion of 0.05, the CT stopping rule ends item administration before ever reaching 20 items. The SE and CT stopping rules deliver similar numbers of items across the θ scale when a maximum number of items is incorporated in the stopping rule.

MI[.56] and MIFL[.56] had the poorest measurement precision out of the stopping rules, as is demonstrated by the standard errors of their θ estimates in Table 3. Though the average $SE_{\hat{\theta}}$ of the MI conditions was lower when an MI criterion value was used, the $SE_{\hat{\theta}}$ remained fairly high in the MIFL[.42] condition. Out of all stopping rule conditions, MI9[.42] had the greatest measurement precision, which was due to this stopping criterion delivering a very high number of items. CT stopping rules also had relatively large $SE_{\hat{\theta}}$ s, with all conditions having grand mean $SE_{\hat{\theta}} \geq 0.37$, with the exception of CT9[.02] and CTFL9[.02]. These two conditions had the highest average NIAs out of the CT conditions, delivering an average of 16 and 15 items, respectively. The CT conditions had the poorest measurement precision when using a 0.05 criterion and no minimum number of items, which is a result of these conditions only administering an

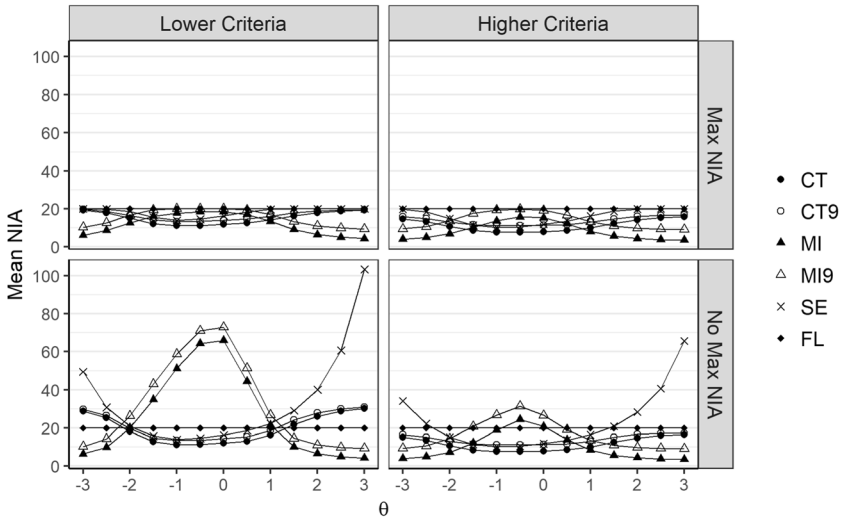

Fig. 2 Mean numbers of items administered (NIA), conditional on known trait level (θ) by stopping rule condition, separated by criterion value and whether they included a maximum-NIA component. The FL condition is included in each panel for comparison. Lower criteria: CT = 0.02, MI = 0.42, SE = 0.30. Higher criteria: CT = 0.05, MI = 0.56, SE = 0.35

average of nine items. CT stopping rules generally produced the shortest tests, and using an MI stopping rule produced the most variable test lengths. SE conditions also produced relatively short tests, particularly when used with using a higher SE criterion. As would be expected, the average $SE_{\hat{\theta}}$ in the SE conditions were close to the SE criteria used to end the CAT program. The FL condition always administered 20 items, which resulted in a low standard error.

Figure 3 displays the average $SE_{\hat{\theta}}$ of each variable-length stopping rule conditional on known θ. Although the MI and MI9 conditions produce low average $SE_{\hat{\theta}}$ s for simulees with θ between – 1.0 and 0.0, all MI conditions increased rapidly, the farther the simulees' true θ was from the peak of the item pool's information function (θ = – 0.6) and were excessively high for simulees in the upper and lower regions of the θ range. SE conditions without a maximum number of items were consistently at the minimum SE value used for this stopping rule. When using an FL component, the $SE_{\hat{\theta}}$ increased when the maximum of 20 items was reached before the minimum standard error criteria was met. CT conditions generally had higher $SE_{\hat{\theta}}$ s in the center of the θ range than did the other conditions, though $SE_{\hat{\theta}}$ s were comparable to those in the SE conditions toward θ extremes.

Latent trait recovery

Table 4 presents descriptive statistics for bias, RMSE, and correlations between the known and estimated θs. The biases

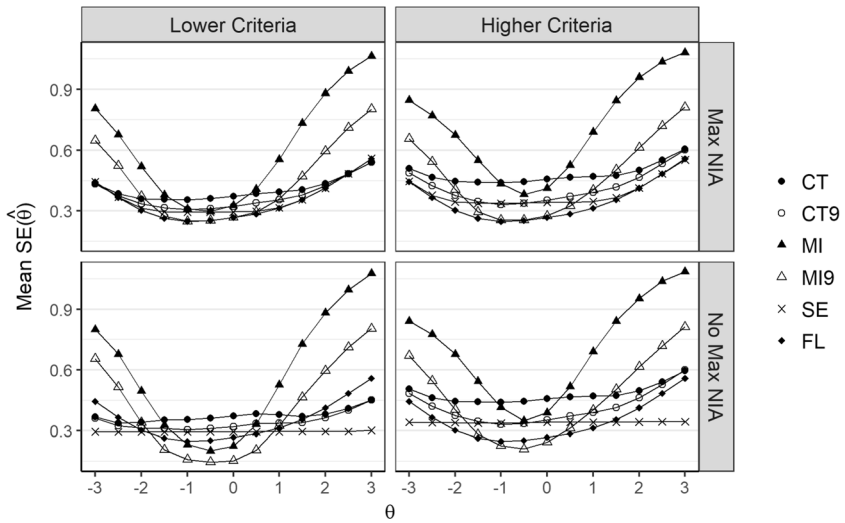

Fig. 3 Grand means of the standard errors of trait estimates ($\hat{\theta}$), conditional on known trait level (θ) by stopping rule condition, separated by criterion value and whether they included a maximum-NIA component. The FL condition is included in each panel for comparison. Lower criteria: CT = 0.02, MI = 0.42, SE = 0.30. Higher criteria: CT = 0.05, MI = 0.56, SE = 0.35

Table 4 Latent trait parameter recovery statistics by stopping rule

Stopping Rule	Bias Mean (Min, Max)	RMSE Mean (Min, Max)	Correlation Mean (Min, Max)
FL	0.01 (– 0.02, 0.05)	0.29 (0.27, 0.32)	.96 (.95, .97)
SE[.30]	– 0.01 (– 0.04, 0.02)	0.29 (0.27, 0.34)	.96 (.94, .97)
SE[.35]	– 0.01 (– 0.04, 0.02)	0.34 (0.31, 0.39)	.95 (.93, .96)
SEFL[.30]	0.00 (– 0.04, 0.04)	0.32 (0.29, 0.34)	.95 (.94, .96)
SEFL[.35]	0.00 (– 0.04, 0.03)	0.35 (0.32, 0.37)	.94 (.93, .95)
MI[.42]	0.07 (0.01, 0.13)	0.48 (0.41, 0.56)	.91 (.89, .94)
MI[.56]	0.07 (0.01, 0.14)	0.61 (0.55, 0.69)	.86 (.83, .89)
MI9[.42]	0.04 (0.01, 0.07)	0.30 (0.25, 0.36)	.96 (.96, .97)
MI9[.56]	0.04 (0.00, 0.07)	0.36 (0.31, 0.40)	.95 (.94, .96)
MIFL[.42]	0.07 (0.02, 0.12)	0.51 (0.45, 0.59)	.90 (.87, .92)
MIFL[.56]	0.07 (0.00, 0.13)	0.61 (0.55, 0.67)	.86 (.82, .89)
MIFL9[.42]	0.03 (– 0.01, 0.06)	0.35 (0.31, 0.40)	.95 (.94, .96)
MIFL9[.56]	0.03 (– 0.01, 0.08)	0.37 (0.33, 0.43)	.94 (.93, .96)
CT[.02]	– 0.01 (– 0.04, 0.02)	0.37 (0.34, 0.41)	.93 (.92, .94)
CT[.05]	– 0.01 (– 0.06, 0.03)	0.46 (0.42, 0.51)	.90 (.87, .92)
CT9[.02]	0.00 (– 0.03, 0.03)	0.33 (0.29, 0.35)	.95 (.94, .96)
CT9[.05]	0.01 (– 0.03, 0.05)	0.37 (0.34, 0.40)	.94 (.92, .95)
CTFL[.02]	0.00 (– 0.05, 0.04)	0.38 (0.34, 0.41)	.93 (.91, .94)
CTFL[.05]	– 0.01 (– 0.07, 0.03)	0.46 (0.42, 0.50)	.90 (.88, .92)
CTFL9[.02]	0.00 (– 0.03, 0.04)	0.34 (0.31, 0.36)	.95 (.93, .96)
CTFL9[.05]	0.01 (– 0.03, 0.04)	0.37 (0.34, 0.40)	.94 (.92, .95)

of the lower and higher stopping criteria values were nearly identical for all stopping rules, and bias was slightly decreased when an FL component was attached to variable length. All FL, SE, and CT stopping rule conditions produced very low bias. Positive bias was present in all MI stopping rule variations. Overestimation of ability was particularly an issue in MI conditions that did not include a minimum NIA requirement. Including this requirement in MI conditions decreased bias by about 50%, producing lower results when used in combination with a FL stopping rule than when there was no maximum-number-of-items requirement. Though the mean bias was very low across all CT conditions, the greatest negative bias found across all simulated datasets was found in lenient CT and CTFL conditions. There was zero average bias when using an SEFL stopping rule, as well as when using a CT stopping rule when it was paired with a lower criterion or a minimum and/or maximum NIA component (i.e., CT9[.02], CTFL[.02], and CTFL9[.02]). Figure 4 presents plots of the mean bias for each stopping rule conditional on known θ. The MI-based conditions had the greatest fluctuation in bias throughout the θ range. Conditions using SE and CT stopping rules demonstrated low levels of bias across θ, particularly when used with the lower criteria. The conditions with a minimum number of items had smaller bias for CT and MI stopping rules across the θ continuum. Though the FL stopping rule produced very little bias, the CT9 and SE stopping rules produced less bias across the θ scale.

As would be expected because of increased test length, variable-length stopping rules using more stringent (i.e., lower) values as termination criteria had lower average RMSEs than did equivalent conditions using more lenient criteria. The RMSEs of variable-length stopping rules were also lower when they were not used in conjunction with a fixed-length component (i.e., maximum NIA), as well when a minimum-NIA requirement was in place in the MI and CT stopping rule conditions. The lowest RMSEs were observed when using the FL stopping rule or the 0.30 SE stopping rule without a maximum-number-of-items component, though MI9[.42] had a similarly low RMSE. The additional constraint of a maximum of 20 items in MIFL9 increased the RMSE, though it was still low, and RMSE further increased by using a lenient termination criterion. The highest RMSE was seen in MI conditions without a minimum-NIA requirement, particularly when using lenient termination criteria. Though using a lower MI criterion improved parameter recovery, the RMSEs of these MI conditions remained quite high. As in the MI conditions, CT stopping rules had better performance (i.e., lower RMSE) when used with minimum and maximum test length

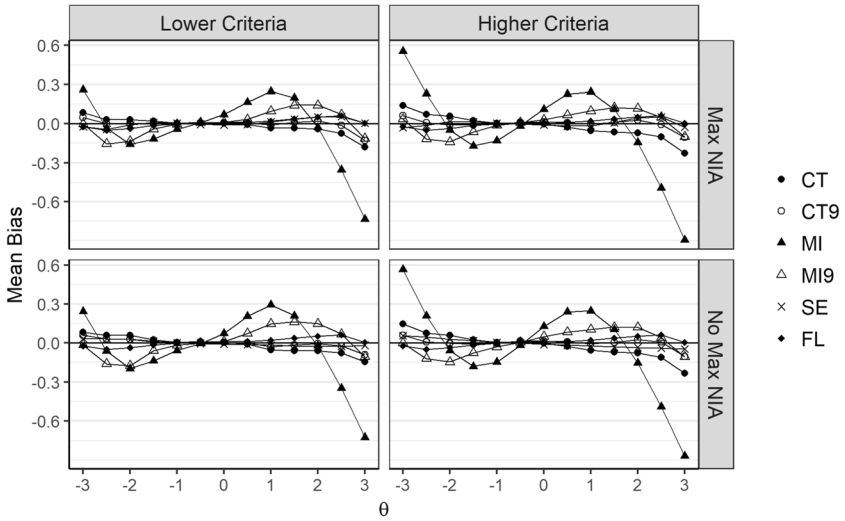

Fig. 4 Bias conditional on known trait level (θ) by stopping rule condition, separated by criterion value and whether they included a maximum-NIA component. The FL condition is included in each panel for comparison. Lower criteria: CT = 0.02, MI = 0.42, SE = 0.30. Higher criteria: CT = 0.05, MI = 0.56, SE = 0.35

components, particularly when used with lower criteria. SE stopping rules yielded low RMSEs across conditions, with the largest average RMSE being produced in the lenient SEFL condition.

Figure 5 depicts the average RMSE conditional on a simulee's known theta. As can be seen in the conditional-bias plot (Fig. 4), the θ recovery of the MI stopping rule varied greatly depending on a simulee's true θ. When used without a maximum number of items, the SE stopping rule produced low and consistent RMSEs across the proficiency range, though it was outperformed by MI9 and FL conditions in the center of the θ distribution, because of their administration of a greater number of items. SE and CT9 had the best performance across θs out of the variable-length stopping rule conditions, though SE led to slightly lower RMSE levels than did CT9.

Examination of the correlations between known and estimated θs revealed the same pattern of results detected for RMSE, in terms of relative parameter recovery ability across stopping rule conditions, and the influence of more stringent (i.e., lower) criteria as well as minimum and maximum NIA. Again, it is apparent that the CT and MI stopping rules are improved by a minimum NIA, especially when more lenient termination criteria are used. MI stopping rules without this requirement produced the lowest θ correlations, particularly when using a higher MI criterion. However, when MI was paired with a minimum NIA, it performed equivalently to the SE stopping rules, whose correlations ranged from .94 to

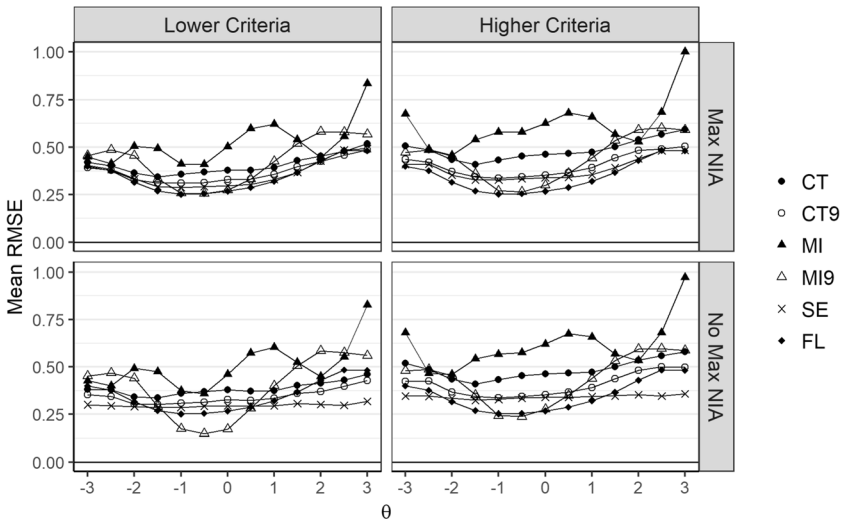

Fig. 5 RMSE conditional on known trait level (θ) by stopping rule condition, separated by criterion value and whether they included a maximum-NIA component. The FL condition is included in each panel for comparison. Lower criteria: CT = 0.02, MI = 0.42, SE = 0.30. Higher criteria: CT = 0.05, MI = 0.56, SE = 0.35

.96. The highest correlations found across CT conditions were only marginally lower than those of the best-performing MI and SE conditions. Coinciding with previous results, the highest correlations were seen in the FL condition as well as in the SE[.30], SEFL[.30], and MI9[.42] variable-length conditions.

Item exposure, pool utilization, and content coverage

Table 5 presents descriptives of item exposure, pool utilization, and content area coverage. The majority of stopping rule conditions produced mean and maximum item exposure rates that were less than the commonly used target maximum exposure rates of 0.2 (e.g., Cheng, Diao, & Behrens, 2017; Wang, Chang, & Douglas, 2012) and 0.3 (e.g., Leroux & Dodd, 2014; Moyer, Galindo, & Dodd, 2012). The exposure rates closely aligned with the number of items administered (Table 3). CT stopping rules, which had the shortest average test length, produced the lowest rates of item exposure. The only exceptions to these low exposure rates were under the MI[.42] and MI9[.42] stopping rules, which had maximum exposure rates surpassing 0.30, as well as the longest test lengths of the studied stopping rules.

Pool utilization was assessed by examining the percentage of the item pool that was not administered across all simulees for each stopping rule condition. The SE stopping rule conditions without a fixed-length component had very high item pool utilization and administered virtually all 157 items across examinees. Using an FL stopping rule led to an average of 24.3% of the item pool not being administered. All variable-length stopping rules with a maximum test length exceeded this percentage, because of their shorter average test lengths. SEFL and CTFL conditions behaved similarly, though the SEFL stopping rules had the greatest pool utilization out of the fixed-length stopping rules. MI and CT stopping rules without a maximum test length had low mean percentages of nonadministered items when they were used with lower criterion values. However, the maximum percentages of unutilized items reveal that these conditions (MI[.42], MI9[.42], CT[.02], and CT9[.02]) did not perform uniformly well across replications, and at times used less than 5% of the items in the pool. The stopping rules that led to greatest number of

Table 5 Descriptive statistics of exposure control and content balancing

Stopping Rule	Exposure Rate Mean (Min, Max)	% of Pool Not Administered Mean (Min, Max)	Difference in Content Area Proportions (Min, Max)
FL	0.13 (0.10, 0.15)	24% (22%, 78%)	(– .03, .02)
SE[.30]	0.14 (0.11, 0.17)	0% (0%, 0%)	(– .01, .02)
SE[.35]	0.10 (0.07, 0.12)	0% (0%, 1%)	(– .02, .03)
SEFL[.30]	0.11 (0.09, 0.13)	25% (24%, 78%)	(– .02, .02)
SEFL[.35]	0.08 (0.07, 0.11)	26% (24%, 77%)	(– .02, .04)
MI[.42]	0.27 (0.24, 0.31)	12% (6%, 97%)	(.00, .01)
MI[.56]	0.09 (0.07, 0.11)	32% (23%, 84%)	(– .02, .01)
MI9[.42]	0.32 (0.28, 0.36)	9% (4%, 97%)	(.00, .01)
MI9[.56]	0.14 (0.11, 0.16)	22% (17%, 87%)	(– .01, .02)
MIFL[.42]	0.10 (0.08, 0.12)	38% (36%, 67%)	(– .02, .02)
MIFL[.56]	0.07 (0.06, 0.09)	42% (40%, 63%)	(– .03, .02)
MIFL9[.42]	0.12 (0.10, 0.14)	31% (29%, 71%)	(– .03, .02)
MIFL9[.56]	0.11 (0.09, 0.13)	32% (31%, 69%)	(– .02, .03)
CT[.02]	0.09 (0.07, 0.11)	16% (11%, 95%)	(– .02, .03)
CT[.05]	0.06 (0.04, 0.07)	30% (25%, 81%)	(– .03, .04)
CT9[.02]	0.10 (0.08, 0.13)	15% (10%, 95%)	(– .01, .03)
CT9[.05]	0.08 (0.06, 0.10)	28% (25%, 83%)	(– .03, .05)
CTFL[.02]	0.08 (0.06, 0.10)	28% (25%, 77%)	(– .02, .03)
CTFL[.05]	0.06 (0.04, 0.07)	31% (27%, 78%)	(– .03, .04)
CTFL9[.02]	0.09 (0.07, 0.12)	27% (24%, 78%)	(– .01, .03)
CTFL9[.05]	0.08 (0.06, 0.10)	30% (26%, 77%)	(– .03, .05)

Exposure rate values were calculated as average across item pool. Min and max difference in content area proportions are the smallest and largest discrepancy between targeted and actual proportions of items administered across the nine content areas

nonadministered items were MIFL conditions without a minimum-number-of-items requirement.

Coverage of content areas was evaluated by finding the discrepancy between the targeted and actual proportions of items to be administered from each content area across all simulees and replications within each stopping rule condition. Table 5 presents the largest discrepancies between the targeted and actual content area proportions across the nine content areas. The average difference between the true and actual content area proportions is not presented because it was zero across conditions. The CCAT content balancing procedure appears to have worked well across stopping rule conditions, since the absolute value of the greatest difference in proportions was less than .05 for each stopping rule. The largest discrepancies in targeted and actual content area coverage were in CT conditions using a more lenient criterion value (i.e., CT = 0.05) and with a minimum test length, which had maximum content area discrepancies around .05. The lowest content area discrepancies were in conditions that had the longest tests, MI[.42] and MI9[.42], which had a maximum difference in content area proportions around .01.

Discussion

This study compared the performance of several stopping rules variations, including minimum information, minimum standard error, and absolute change in $\hat{\theta}$ under the GPCM. Our results extend prior research of these termination criteria using a dichotomous model (i.e., Babcock & Weiss, 2012) to a polytomous model and include content balancing and exposure control procedures. Developers of high-stakes tests have high motivation to create fair assessments and maintain test security. Therefore, they frequently use content balancing and exposure control procedures in combination, which ensure equal representation of content areas across examinees and limit item use to decrease the likelihood of item disclosure, respectively. The constraints these procedures place on the item selection process prevent the most informative item from always being selected, thereby increasing the number of items needed to satisfy the precision required by variable-length stopping rules. Despite the wide use of these techniques and their effect on measurement precision and efficiency, limited research has implemented both constraints simultaneously, and the present study is the first to our knowledge that compares variable-length stopping rules while controlling for item exposure and content balancing.

In general, most CAT stopping rule procedures demonstrated that they could arrive at precise and accurate estimates of θ, though the MI stopping rule demonstrated either inefficiency or poor θ recovery in the majority of conditions investigated. Furthermore, it is apparent that in order to reach optimal results, the MI- and CT-based stopping rules required an additional prerequisite that a minimum number of items be administered. All variable-length stopping rules were more efficient when paired with a fixed-length component (i.e., maximum number of items), as there were only negligible differences in ability parameter recovery, accompanied by often large decreases in test length. Terminating a test on the basis of either variable- or fixed-length termination criteria allows for equal measurement precision across the majority of examinees, while also preventing item overexposure. Our study used 20 items for FL conditions due to previous research indicating this number to produce low levels of bias and high precision in previous research, while also providing a short and efficient test (Dodd, 1990; Dodd et al., 1989, 1993; Gorin et al., 2005; Koch & Dodd, 1989; Lee & Dodd, 2012). However, if a larger number of items were to be used as the maximum test length, then differences in efficiency would be less pronounced. There are a number of considerations when making the decision on maximum test length, most importantly the relative importance of limiting item exposure for test security and the degree of precision desired by the testing scenario. Appropriate test length will vary across item pools and testing contexts, which may require empirically supported maximum test lengths determined through simulation (see Thompson & Weiss, 2011).

The importance of including a minimum number of items for MI and CT conditions is apparent throughout results. Although this inevitably increased the number of items administered and item exposure, these conditions saw meaningful gains in θ measurement. This was particularly true in MI conditions, which saw dramatic decreases in $SE_{\hat{\theta}}$, bias, and RMSE and increases in θ correlations. This can be partially attributed to increased average test length, but is also indicative of these stopping rules tending to be satisfied before obtaining an accurate and reliable estimate of an individual's ability. Though using a higher MI criterion decreased the average test length, this produced unacceptable decreases in measurement precision and accuracy. The MI stopping rule administers fewer items to examinees the farther they are from the center of the θ distribution, since there are increasingly fewer items that provide the information required to meet the MI criteria (Dodd et al., 1989). This proclivity toward shorter tests for examinees with extreme θs is amplified when including exposure control and content-balancing constraints on item selection as the most informative item in the pool may not be available for administration. Item exposure rates, pool utilization, and content area coverage were closely aligned with test length. Although shorter tests produced lower exposure rates, they also left a greater proportion of the item pool unused and greater discrepancies in actual and

targeted content area proportions. Ideally, all the items in the pool would be used and have an equal exposure rate, to prevent a waste of resources and enhance test security, respectively. CT and SE stopping rules generally maintained equilibrium between item exposure and pool utilization, though SE stopping rules without a fixed-length component clearly excelled in having the fewest unused items.

As can be seen in the conditional θ plots, measurement efficiency and θ recovery of MI stopping rules fluctuate greatly depending on the location of an examinee's true ability on the θ scale. Without a minimum number of items requirement, the MI rules frequently ended item administration before obtaining an accurate and precise θ estimate, particularly for examinees with θs further from the center of the θ distribution. These overly short tests for these examinees also contributed to the high nonconvergence rates. Though the inclusion of minimum and maximum numbers of items improved the balance of measurement efficiency and precision on average, there is still great disparity in test lengths and θ recovery for examinees across the θ distribution. The MI stopping rule can have consequences for both examinees in the center of the θ distribution center and those at the extremes. Those at the center have increased testing burden and are administered more items than necessary for measurement and those in the extremes are likely to be receive an inaccurate and imprecise proficiency score. Our results indicate that the MI stopping rule should be used with caution due to difficulty in finding a suitable criterion value to attain balance in efficiency and quality of measurement, especially when using exposure control and content balancing.

CT9 and CTFL9 shared nearly identical results when used with a 0.05 criterion. This indicates that the change in $\hat{\theta}$ stopping rule was usually reached before 20 items were administered, as was seen in previous research (Babcock & Weiss, 2012). CT conditions with a 0.05 criterion had the shortest average test lengths, though it appears that tests may have been shorter than required for accurate and reliable θ estimates in many cases. CT had a better performance when used with a 0.02 criterion, particularly when used with a minimum number of items component. Attaching a maximum number of items to the CT stopping rule did not have much of an impact and produced only slight differences in measurement when used with the lower criterion. Babcock and Weiss (2012) used the same CT criteria but found very different average test lengths, bias, and RMSE due to their use of a different item pool, dichotomous IRT model, uniformly distributed simulee θ distribution, as well as their lack of content balancing and exposure control constraints on item selection. The authors stated that CT may be preferred over the SE stopping rule for exams that include a variety of item banks with varying information structures, but the differences in our results indicate that the performance of the CT stopping rule is dependent on the components of the CAT, including the response model, item bank, and item selection procedure.

When viewed holistically, SE stopping rules generally maintained the greatest balance of efficiency and precision across conditions. This result is consistent with previous research comparing SE and MI stopping rules with different IRT models (e.g., Babcock & Weiss, 2012; Dodd, 1990; Dodd et al., 1989, 1993), indicating that the SE stopping rule is preferable to the MI stopping rule across multiple scoring models, as well as when exposure control and content balancing are implemented. Unlike the MI and CT conditions, SE and SEFL had no convergence issues that required an additional criterion for a minimum number of items and performed similarly when used in isolation. As expected, administering a fixed length of 20 items provided more precise measurement of θ due to an increased test length in comparison to the variable-length stopping rules. FL had increased precision in the center of the θ distribution in which examinees were administered more items than necessary for an acceptably precise $SE_{\hat{\theta}}$, meaning that using a FL stopping rule is inefficient for the majority of examinees. SE and SEFL administered fewer items to these examinees and maintained their low predefined $SE_{\hat{\theta}}$.

Combining SE with a fixed-length component led to an efficient test in which only the small number of examinees with relatively high or low ability saw increases in $SE_{\hat{\theta}}$. Both SE criteria values performed well, with the 0.30 criterion producing more precise $\hat{\theta}$ but having slightly increased test length. A researcher's decision between a higher or lower SE termination criterion should be based on whether efficiency or precision is the more desirable trait. If testing burden is the primary concern, then a higher criterion should be used to produce a shorter average test length. A lower criterion may be used if optimal measurement precision and accuracy is of greater importance. However, the differences between these criteria were quite minimal and both provided excellent balance between efficiency and quality of measurement.

Though the SE stopping rule generally outperformed the CT stopping rule, the CT stopping rule performed well when using a 0.02 criterion paired with minimum and maximum test length constraints. Furthermore, as described by Babcock and Weiss (2012), using a change in θ termination criterion may be more appropriate than a SE stopping rule when maintaining a low standard error of measurement across the majority of examinees is unlikely. Such circumstances could arise from using a small item bank or having mismatch between the item pool and trait distributions. Babcock and Weiss's work indicated that the CT stopping rule is less affected by changes in the item pool's information structure, a property not manipulated in this study. Given that CT is a relatively new stopping rule, there is considerable room for research on its utility under different simulated conditions. Another avenue for future research is the study of variable-length stopping rules used in combination. As was suggested by Babcock and Weiss

(2012), CT and SE stopping rules may work well when used in combination, with the CT stopping rule ending a test for examinees that are in ranges of θ where the minimum SE criterion cannot be reached.

Some limitations of our study could limit the generalizability of our results. Although numerous stopping rule components were manipulated, we used a single item pool and simulee population across conditions, which had similar information structures. It is possible that the relative performance of the stopping criteria may change when used with different item banks and θ distributions. Future research possibilities include examining the effect of differently shaped and sized item pools on stopping rule performance, as well as the degree of mismatch between the item pool information and the trait distribution.

Though exposure control and content balancing were not a primary focus, the results of our study may be dependent on the specific methods used for constraining item selection. Future research may investigate the effectiveness of SE, MI, and CT stopping rules across various methods designed to reduce item exposure and match targeted content area distributions. Though several stopping rules modifications were used in this study, our findings are limited to the specific criteria used. Although the test information function of an item pool can be helpful in determining the criteria used for MI and SE stopping rules, there is no such possible method for CT. Therefore, it may be difficult to judge the degree of measurement quality and expected test length when choosing between CT criteria. Further investigation should consider the use of alternative CT criteria, as different values may give this index superior performance relative to the SE stopping rule. Future research could conduct a more thorough investigation of the impact of various termination criterion values across different IRT models, item pools, and examinee populations. As was described by Thompson and Weiss (2011), simulation studies are necessary to determine the appropriate CAT properties (e.g., test length and termination criteria) required to ensure the degree of precision and test efficiency required in operational testing scenarios.

References

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561–573.

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*, 1–18.

Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering & R. Ostini (Eds.), Handbook of polytomous item response theory models (pp. 229–255). New York, NY: Routledge.

Boyd, A. M., Dodd, B., & Fitzpatrick, S. (2013). A comparison of exposure control procedures in CAT aystems based on different measurement models for testlets. *Applied Measurement in Education, 26*(2), 113–135.

Chen, S.-K., & Cook, K. F. (2009). SIMPOLYCAT: An SAS program for conducting CAT simulation based on polytomous IRT models. *Behavior Research Methods*, *41*, 499–506. doi:https://doi.org/10.3758/BRM.41.2.499

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*, 502–512. doi:https://doi.org/10.3758/s13428-016-0712-6

Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71*(1), 37–53.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, *28*, 165–185.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*, 335–356.

Davis, L. L., Pastor, D. A., Dodd, B. G., Chiang, C., & Fitzpatrick, S. J. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement*, *4*, 24–42.

De Ayala, R. J. (2009). The theory and practice of item response theory. New York, NY: Guilford Press.

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement*, *14*, 355–366.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *19*, 5–22.

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

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, *53*, 61–77.

Fitzpatrick, S. J. Choi, S. W., Chen, S., Hou, L., & Dodd, B. G. (1994). IRTINFO: A SAS macro program to compute item and test information. *Applied Psychological Measurement*, *18*, 380.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement*, *29*, 433–456.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.

Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive tests. *Applied Measurement in Education*, *2*, 335–357.

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for cat using the partial credit model. *Educational and Psychological Measurement*, *72*, 159–175.

Leroux, A. J., & Dodd, B. G. (2014). A comparison of stopping rules for computerized adaptive screening measures using the rating scale model. *Journal of Applied Measurement*, *15*, 213–226.

Leroux, A. J., & Dodd, B. G. (2016). A comparison of exposure control procedures in cats using the GPC model. *Journal of Experimental Education*, *84*, 666–685.

Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, *73*, 857–874.

Lord, F. M. (1971). Robbins–Monro procedures for tailored testing.

Educational and Psychological Measurement, *31*, 3–31.

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Moyer, E. L., Galindo, J. L., & Dodd, B. G. (2012). Balancing flexible constraints and measurement precision in computerized adaptive testing. *Educational and Psychological Measurement*, *72*, 629–648.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Practical considerations in computer-based testing. New York, NY: Springer.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice*, *8*, 11–15.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1), 1–9.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). Computerized adaptive testing: A primer (2nd). Mahwah, NJ: Routledge.

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*, 95–109. doi:https://doi.org/10.3758/s13428-011-0143-3

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

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, *27*, 299–300.