Journal of Computerized Adaptive Testing

Volume 11 Number 2
October 2024

Evaluating the Effectiveness of the Standard Error of Score Estimation as a CAT Termination Criterion

Chansoon (Danielle) Lee
American Board of Internal Medicine

Kyung (Chris) T. Han
Graduate Management Admission Council

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

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

Journal of Computerized Adaptive Testing
Volume 12, Number 2, October 2024
DOI 10.7333/2410-1102013
ISSN 2165-6592

Evaluating the Effectiveness of the Standard Error of Score Estimation as a CAT Termination Criterion

Chansoon (Danielle) Lee
American Board of Internal Medicine
Kyung (Chris) T. Han
Graduate Management Admission Council

In computerized adaptive testing (CAT), the standard error of score estimation (SEE) value is sometimes used as a test termination criterion since test forms with the same level of SEE are believed to exhibit equivalent measurement precision even when they vary in the number of items. This study aimed to evaluate the appropriateness and effectiveness of using SEE as an approximation for actual standard error of measurement (SEM) under various conditions. First, this study illustrated a potential issue with the inconsistent relationship between the SEE and actual SEM at different lengths in equivalent tests. A series of simulation studies were then conducted to investigate the relationship among SEE, SEM, and test lengths in a variable-length CAT administration. For this purpose, various combinations of score estimation methods and item selection methods were considered. The study found that using the SEE as a sole criterion to terminate a CAT administration might result in a less than ideal level of control in measurement precision when the test was either extremely short or administered in a variable-length CAT. The findings and discussion of this study provide insights and means to better control measurement quality and precision, especially in cases where controlling SEE alone might not be sufficient across tests of varying lengths.

Keywords: *computerized adaptive testing, CAT termination, standard error of ability estimation, standard error of measurement, test construction, test reliability*

In computerized adaptive testing (CAT), test forms differ across individual examinees both in the composition of items and the number of items. This is because test items are selected based on an individual's responses to previously administered items. For example, the National Council Licensure Examination for registered nurses (NCLEX-RN) is a variable-length CAT-based exam that can administer between 75 and 145 items (NCSBN, 2019, p. 42). When there are significant

variations in test lengths, it becomes crucial to ensure that all adaptively individualized sets of items, administered from the item bank, exhibit the same level of score precision. In CAT, it is presumed that the equivalence of score precision across different test forms is achieved by controlling the test information function (TIF) or standard error of score (θ) estimation (SEE). SEE is approximately the reciprocal of the square root of the TIF at a given score (θ). According to this assumption, when different test forms yield the same SEE at a given θ value, the test forms are expected to possess the same measurement precision for individuals at that θ point.

However, the testing landscape has undergone significant changes in response to the pandemic and educational inequality issues, leading to a high demand for shortening assessments across various industries. For instance, the Executive Assessment, which evaluates a person's readiness for business schools, has transitioned to comprising three sessions with 12 to 14 items for each session (mba.com, n.d.). Similarly, GMAT has announced dramatically shortened versions, which can contain fewer than 20 scored items. Additionally, both GRE and SAT have adjusted their formats to meet the demands from educational institutions and examinees. Such a shift toward shorter tests has prompted testing organizations to explore variable-length CAT as a potential option for maintaining measurement precision while reducing test length. Considering the demand for shorter assessments, this study aimed to investigate the consistent relationship between analytically computed SEE and empirically measured SEM given the examinee's ability estimate, both in low-stakes and high-stakes assessment scenarios.

In item response theory (IRT), SEE is often viewed as an analytical approximation of the conditional standard error of measurement (CSEM) at given estimated θ values (e.g., Yen & Fitzpatrick, 2006). Given this perspective, it seems reasonable to assume that the CSEMs of the θ estimates would also be comparable, as long as their SEEs are comparable. Following this line of thought, even when examinees respond to different sets of items, potentially with varying numbers of items, one can maintain the belief that their score comparability is preserved as long as the test forms exhibit equivalent SEEs. Consequently, it has been considered a reasonable practice to terminate a CAT administration when the SEE of θ reaches a predetermined criterion value. This approach permits the use of variable test lengths across individual examinees. While the assumption regarding the approximation of SEE has been widely embraced in the field of measurement for decades, there remains a lack of comprehensive exploration into the distinctions between SEE and CSEM, as well as their interrelationship. Although previous research has delved into the impact of the different score estimation methods on SEE, SEM, or test length (Han, 2016; Sulak & Kelecioğlu, 2019; S. Wang & T. Wang, 2001; Wang & Vispoel, 1998; Yi et al., 2001), only a limited number of studies have dedicated their investigations to the relationship among SEE, SEM, and test length.

The primary objective of this study was to conduct a thorough examination of these relationships, particularly within the context of variable-length CAT applications involving exceptionally short tests intended for both low-stakes and high-stakes situations. The implications of the findings from this study extend beyond these short CATs and can be extrapolated to longer CATs, which are commonly encountered in high-stakes situations. The insights gleaned from this study provide practical guidelines of significant value to practitioners, including test developers and users. These guidelines focus on the effective utilization of SEE to maintain control over the precision of test scores, particularly within the domain of extremely short CATs designed usually for low-stakes assessments, but expanded for high-stakes assessments due to the changing test landscape. Furthermore, the findings from this study illuminate the potential interplays among other critical factors within CAT that impact the relationship among SEE, SEM, and test lengths. This, in turn,

advances our understanding of measurement precision across diverse testing scenarios.

The subsequent section provides an overview of the background of the SEE and elucidates its significance and relationship with SEM. The paper then directs its attention toward a potential concern: the incongruent association between the SEE calculated analytically and the SEM measured empirically, given the estimated ability level ($\hat{\theta}$). This incongruity manifests across different test lengths within the context of a variable-length CAT framework.

Standard Error of Score Estimation (SEE)

The standard error of score estimation (SEE) serves as a commonly used termination rule in CAT. When given $\hat{\theta}$, which is obtained through maximum likelihood estimation (MLE) of θ, SEE can be determined using the reciprocal of the square root of the test information function (TIF) (Wainer et al., 1990). This relationship is expressed as follows:

$$\mathrm{SEE}(\hat{\theta}) \approx \frac{1}{\sqrt{\mathrm{TIF}(\hat{\theta})}}. \tag{1}$$

Under the framework of the two-parameter logistic model (2PL), the TIF associated with a given $\hat{\theta}$ can be calculated as:

$$\mathrm{TIF}(\hat{\theta}) = \sum_{i=1}^{n} \mathrm{IIF}_i(\hat{\theta}) = \sum_{i=1}^{n} \frac{[P_i'(\hat{\theta})]^2}{P_i(\hat{\theta})Q_i(\hat{\theta})} = \sum_{i=1}^{n} a_i^2 P_i(\hat{\theta})Q_i(\hat{\theta}). \tag{2}$$

Here, the index i pertains to the item (i = 1, 2, …, n). IIF stands for the item information function, while $P_i(\hat{\theta})$ and $Q_i(\hat{\theta})$ denote the probabilities of responding correctly and incorrectly, respectively, to item i given $\hat{\theta}$. Moreover, $P_i'(\hat{\theta})$ signifies the first derivative of $P_i(\hat{\theta})$ with respect to $\hat{\theta}$, and a_i represents the item discrimination parameter. As demonstrated in Equation 2, the TIF is the aggregate sum of the IIF values for n items.

Numerous studies have undertaken the evaluation and comparison of the SEE-based termination rule with alternative termination methods across various conditions. For instance, Babcock and Weiss (2012) conducted an examination of different termination rules concerning the precision of θ estimation in CAT scenarios involving distinct item banks. They found that employing a CAT termination criterion featuring a low SEE, such as 0.315, could yield high quality measurements when applied using a substantial item bank. Choi et al. (2011) brought to light that both the SEE-based rule and minimum information rule exhibited insensitivity to the relationship between the θ level and the information function of the item bank. Meanwhile, Dodd et al. (1989, 1993) contrasted the impact of termination rules between the graded response model (Samejima, 1969) and the partial credit model (Masters, 1982). In both of these studies, it was observed that the SEE-based termination rule yielded more robust outcomes, resulting in increased correlations between true and estimated θ values, all while administering fewer items in comparison to the minimum item information rule.

In practical applications, some practitioners mistakenly equate SEE with CSEM, using the terms interchangeably. For example, within CAT practices, test developers often employ centered root mean square error to assess the equivalence of test score reliability, operating under the

assumption that root mean squared error (RMSE) can be managed by controlling SEE during CAT administrations. However, it is crucial to recognize that SEE functions as an analytical tool for estimating conditional SEM given $\widehat{\theta}$, while CSEM represents an empirical summary of observed errors. To illustrate, two commonly used statistics for comparing SEE with actual CSEM are conditional mean absolute error (CMAE) and CRMSE. CMAE computes the mean absolute difference between the estimated θ and true θ at each data point (k) along the θ scale, accounting for all examinees (j = 1, 2, …, N) at that point:

$$\mathrm{CMAE}\left(\hat{\theta}_k\right) = \frac{\sum_{j=1}^{N_k}\left|\hat{\theta}_k^j - \theta_k^j\right|}{N_k}. \tag{3}$$

Similarly, CRMSE represents the standard deviation of the mean square difference between $\hat{\theta}$ and θ at each data point (k):

$$\mathrm{CRMSE}\left(\hat{\theta}_k\right) = \sqrt{\frac{\sum_{j=1}^{N_k}\left(\hat{\theta}_k^j - \theta_k^j\right)^2}{N_k}}. \tag{4}$$

Additionally, to comprehend the systematic aspect of the estimation errors, practitioners also consider the conditional bias (Cbias; Harwell, 2018). Cbias quantifies the signed mean difference between $\hat{\theta}$ and θ at each data point (k), providing insights into estimation accuracy:

$$\mathrm{Cbias}\left(\hat{\theta}_k\right) = \frac{\sum_{j=1}^{N_k}\left(\hat{\theta}_k^j - \theta_k^j\right)}{N_k}. \tag{5}$$

In addition to these measures, metrics from classical test theory (CTT), such as reliability and CTT-based SEM (CTT-SEM), are often used to communicate with test users (Haertel, 2006). It is important to note that these metrics summarize measurement errors differently and should not be used interchangeably without proper justification. Reliability in CTT quantifies the ratio of the true score variance to observed score variance (σ_T^2 / σ_X^2), where the observed score variance is composed of true score variance and error variance ($\sigma_X^2 = \sigma_T^2 + \sigma_E^2$). CTT-SEM represents the standard error of measurement. The reliability (r) in this study was calculated using weights (ω_k) associated with each data point (k) in a composite score (Kolen, 2006):

$$\text{Reliability } (r_{xx}) \;=\; 1 \;-\; \sum_{\mathrm{k}}\left(\frac{\omega_k \sigma_{E\,k}^2}{\omega_k \sigma_{T\,k}^2 \;+\; \omega_k \sigma_{E\,k}^2}\right). \tag{6}$$

The weights reflect the relative contribution of each data point (k) to the overall score in a linear combination (Feldt & Brennan, 1989). The CTT-SEM is then derived using the standard deviation of θ (s) and the computed reliability (r):

$$\text{CTT-SEM} = \; s\sqrt{(1-r)}. \tag{7}$$

Given these distinctions, it becomes evident that some practitioners erroneously treat SEE and CSEM as equivalent, potentially leading to misinterpretations.

Illustrative Example Showing the Relationship Between SEE and CSEM

As part of the illustrative example, a set of 15 test forms was generated, each form featuring variable test lengths ranging from 2 to 20 items. To create what we refer to as "theoretically equivalent forms" at a specific θ point, all the forms were designed to possess an identical TIF value of 7.0 (equivalent to SEE of 0.38) at a common θ point ($\theta = 0$) based on the 2PL model. This was achieved by adjusting the item discrimination values (a parameter) while keeping the item difficulty constant ($b = 0$) for all items. A summary of the test forms is provided in Table 1.

Table 1
Test Forms Illustrating a Potential Issue

Form	Number of Items	IRT a parameter	IIF ($\theta = 0$)	TIF ($\theta = 0$)
1	2	2.198	3.500	7
2	3	1.795	2.333	7
3	4	1.554	1.750	7
4	5	1.390	1.400	7
5	6	1.269	1.167	7
6	7	1.175	1.000	7
7	8	1.099	0.875	7
8	9	1.036	0.778	7
9	10	0.983	0.700	7
10	11	0.937	0.636	7
11	12	0.897	0.583	7
12	13	0.862	0.538	7
13	14	0.831	0.500	7
14	15	0.803	0.467	7
15	20	0.695	0.350	7

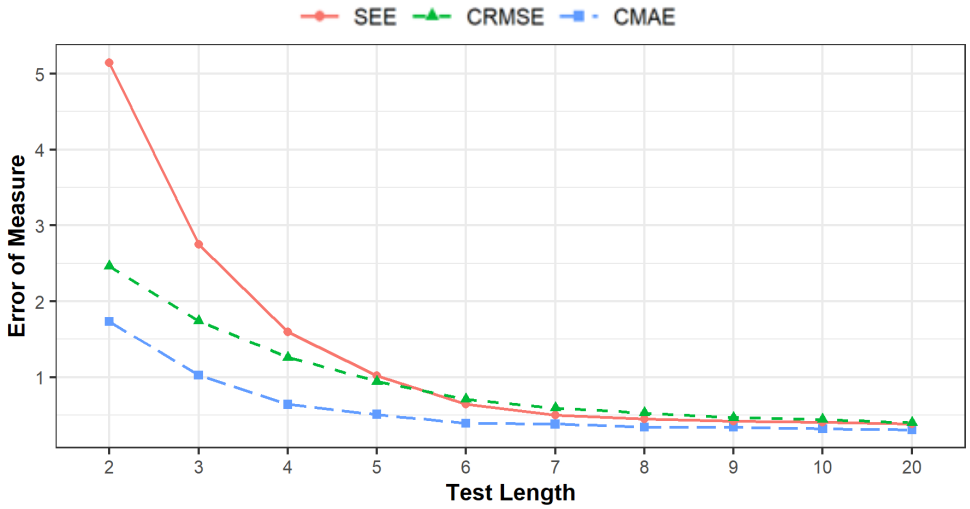
The simulation study involved 1,000 simulated examinees, each assigned θ values set at the same value as the b parameter ($\theta = 0$). These θ values were then estimated using maximum likelihood estimation (MLE; Birnbaum, 1968). The θ estimates were constrained to fall within the range of –3.5 to 3.5 through truncation. The simulation was replicated 25 times, and the results were subsequently aggregated to provide an average outcome. The analysis of CMAE and CRMSE were conducted to assess the precision of measurement and its relationships with SEE.

Figure 1 depicts a comparison of conditional error metrics (SEE, CRMSE, and CMAE) across different test lengths, ranging between 2 and 20 items. A significant observation from this example was the distinctive declining patterns exhibited by the three metrics as the number of test items increased. Among these measures, SEE exhibited the largest errors when the test length was extremely short (≤ 5). Meanwhile, CRMSE consistently displayed substantially smaller measurement errors than SEE, and its values rapidly stabilized in administrations featuring six or more items. CMAE, in this example, consistently remained smaller than CRMSE, and its rate of

decline in value as the test length increased was notably slower. Also, a few additional scenarios with different TIF and SEE values were examined (e.g., SEE of .32 and .20), and they all showed results similar to the example in Figure 1.

A natural follow-up question would be whether the practice of using SEE as the test termination criterion for CAT applications ensures comparable CSEM, signifying equivalent score precision, across CAT administrations with different test lengths, particularly in situations where

Figure 1
Measurement Errors Under Different Test Lengths



the test length is short. Additionally, many CAT-based test programs incorporate different θ estimation methods within the same administration. For instance, some Bayesian methods such as expected a posteriori (EAP; Bock & Aitkin, 1981) are used at the commencement of a CAT administration, while MLE is employed for the final θ estimation. It is widely recognized that these distinct θ estimation methods yield divergent SEEs due to potential shrinkage or stretching of the θ scale (Yen & Fitzpatrick, 2006). Consequently, the choice of θ estimation method can also contribute to determining the relationship between SEE and CSEM.

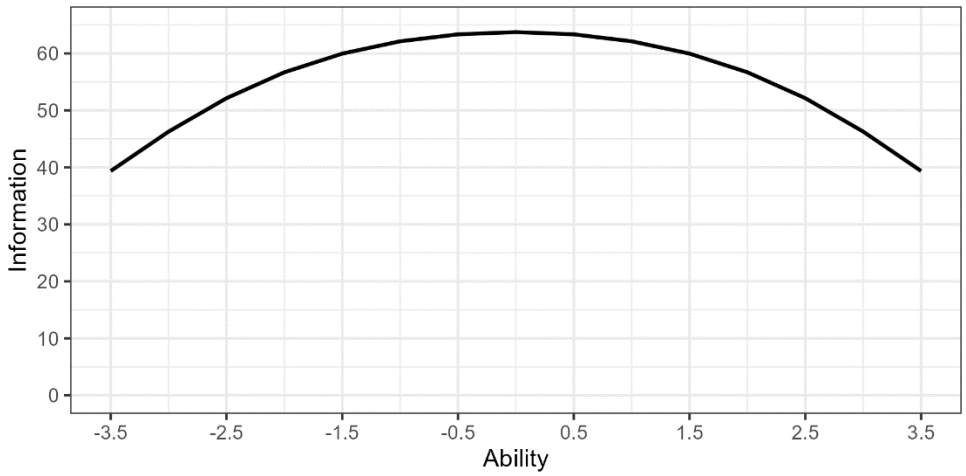
To address these questions, this study investigated the relationship among SEE, CSEMs, and test lengths within the context of a variable-length CAT framework, considering various conditions that include controlled SEE, different score estimation methods, and different item selection approaches. This investigation was conducted through a series of simulation studies, with the primary objectives being twofold: (1) scrutinizing whether the control of SEE in a variable-length CAT results in comparable CSEM across CAT administrations, and (2) understanding the ramifications of additional factors within CAT, such as the choice of θ estimation method and item selection criterion.

Method

The study was based on 1,000 simulated examinees. Each examinee was assigned an identical θ value for each of 15 predetermined evaluation points, ranging from –3.5 to 3.5 with increments of 0.5. All examinees began a CAT with an initial $\hat{\theta}$ set at 0 to mitigate initial variation. The item bank utilized for this study included a total of 750 simulated items. Among these, 50 items were

associated with each of the evaluation points distributed between −3.5 and 3.5. To minimize the influence of the item bank, an abundant availability of items across all difficulty levels, including extreme areas, was ensured to facilitate item selection for any given data values. Figure 2 illustrates the information function for the item bank. The IRT model used was the 2PL model with D = 1.702. The *a* parameters were drawn from a uniform distribution (U[0.3, 1.2]), while the *b* parameters were drawn from U[−3.5, 3.5], with an increment of 0.5. The CAT administration was designed to terminate once SEE reached a predefined value, without enforcing a minimum item requirement. Conversely, the maximum number of items was set at 100. Throughout the simulation studies, neither item exposure control nor item content balancing mechanism were implemented. The simulations used the software *SimulCAT* (Han, 2012b).

Figure 2
Information Function for the Item Bank



The study varied four key elements through simulation: (1) SEE criteria for CAT termination, (2) θ estimation methods, and (3) item selection methods. Specifically, three distinct SEE criteria (0.4, 0.5, and 1.0) were employed for the termination of CAT administrations. Each criterion carries varying practical implications for test design. For instance, SEE values of 0.4 or 0.5 could be considered reasonable targets for a CAT-based examination aiming to achieve a CSEM comparable to a fixed or linear test form consisting of approximately 30 to 40 items. Conversely, a SEE of 1.0 might be an appropriate selection for short testlets designed for formative assessments or quizzes intended for diagnostic purposes.

θ Estimation Methods

Four θ estimation methods were studied: two maximum likelihood methods—MLE (Birnbaum, 1968) and maximum likelihood estimation with fences (MLEF; Han, 2016)—and two Bayesian methods—expected a posteriori (EAP; Bock & Aitkin, 1981) and maximum a posteriori (MAP; Samejima, 1969). Interim and final θ estimates were truncated between −3.5 and 3.5. For the EAP and MAP methods, a prior of the standard normal distribution N(0, 1) was used. MLEF employed lower and upper imaginary fences at −3.5 and 3.5, respectively, to handle extreme response patterns, such as all-correct, all-incorrect, all-harder-items-correct, and all-easy-items-incorrect responses, not accommodated by MLE. The ML-based methods were known to have

lower bias and higher SEE and CSEM than the Bayesian-based methods (Han, 2016; Sulak & Kelecioğlu, 2019; S. Wang & T. Wang, 2001; Wang & Vispoel, 1998). Furthermore, the ML-based methods resulted in relatively consistent CSEM values across the θ scale compared to the Bayesian-based methods (Han, 2016; Wang & Vispoel, 1998).

Item Selection Methods

Similarly, the choice of item selection methods in CAT can influence measurement performance. The analysis involved three distinct conditions encompassing various item selection methods—Fisher information (MFI; Weiss, 1982), Kullback-Leibler information (KLI; Chang & Ying, 1996), and efficiency balanced information (EBI; Han, 2012a). Previous research has highlighted that MFI has a propensity to yield relatively small SEEs compared with the other studied item selection methods (Han, 2012a; Sulak & Kelecioğlu, 2019). Additionally, Sulak and Kelecioğlu (2019) found that the utilization of MFI in CAT configurations led to the shortest test length among the studied methods when aiming for a predefined SEE of 0.4. In the realm of item selection strategies, the KLI method (also recognized as a global information method) has been acknowledged for attaining smaller mean squared errors in θ estimation during the early stages of CAT, in contrast to MFI (Chang & Ying, 1996; Eggen, 1999). Furthermore, the EBI method, which makes use of items with lower discrimination without requiring item pool stratification, has demonstrated a higher correlation between true and estimated ability levels when compared to MFI (Han, 2012a).

The assessment metrics employed in this study included Cbias, CRMSE, conditional test length, CTT-based reliability with equal weights for evaluation points, and CTT-SEM. These measures collectively offered valuable insights into the nature of estimation errors, test length, and score precision under the varied factors examined in the simulation studies.

Results

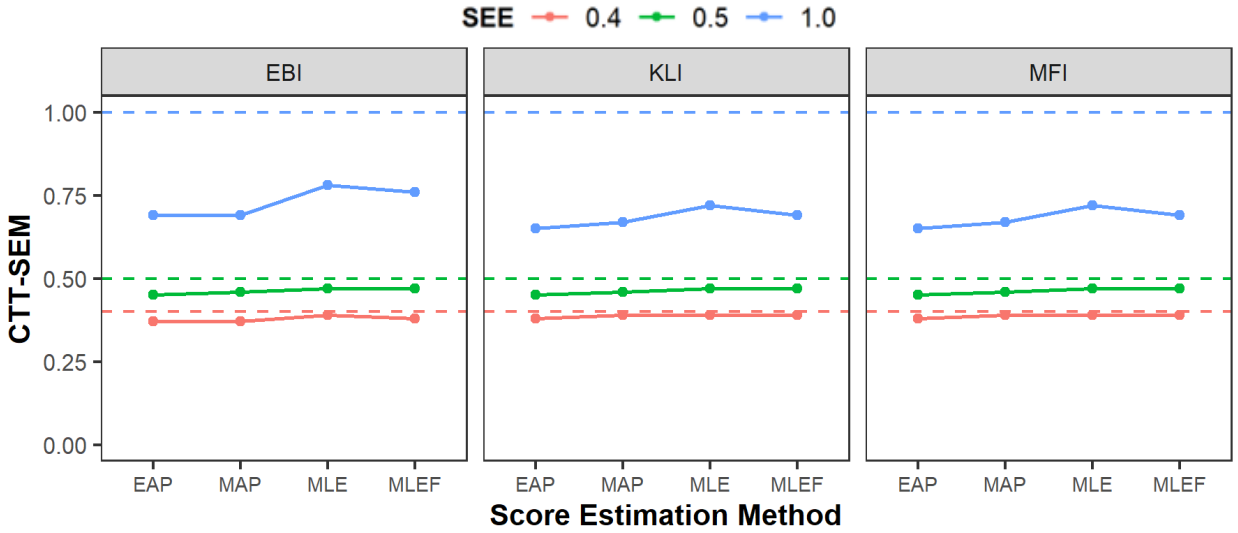
CTT-SEM and reliability were calculated for all combinations of the examined factors across the range of θ from –2 to 2, and these outcomes are summarized in Table 2. As per Equation 7, consistent reliabilities yielded comparable CTT-SEMs across the majority of the simulation scenarios when SEE was controlled. It was observed that CTT-SEMs remained relatively unaffected by choice of score estimation methods and item selection methods when SEE was set at 0.4 or 0.5. However, under conditions of SEE = 1.0, slightly smaller CTT-SEMs were evident in Bayesian-based score methods and MFI or KLI item selection methods when contrasted with ML-based score methods and EBI item selection method. The relatively small CTT-SEMs in Bayesian-based methods can be attributed to the prior value, which was set equal to the true θ value, and its item selection algorithm that initiates around the prior value and progressively extends to the extreme θ region through incorrect responses. Figure 3 provides a visual representation of the overestimation of SEE across all simulation conditions—indicating that SEE consistently surpassed CTT-SEM. This overestimation became more pronounced with higher SEE values. For instance, at SEE = 1.0, CTT-SEM ranged from 0.65 to 0.67 for EAP and MAP score methods, and slightly higher at 0.69 to 0.78 for MLE and MLEF score methods. This underscored that SEE could be set higher than anticipated to more closely align with the expected CTT-SEM, even when confronted with shorter test length.

Table 2
CTT-SEM and Reliability Averaged Across 25 Replications

Item Selection	θ Estimation	SEE = 0.4		SEE = 0.5		SEE = 1.0	
		CTT-SEM	r_{xx}	CTT-SEM	r_{xx}	CTT-SEM	r_{xx}
EBI	EAP	0.37	0.92	0.45	0.88	0.69	0.72
	MAP	0.37	0.92	0.46	0.87	0.69	0.72
	MLE	0.39	0.91	0.47	0.87	0.78	0.63
	MLEF	0.38	0.91	0.47	0.87	0.76	0.65
KLI	EAP	0.38	0.91	0.45	0.88	0.65	0.75
	MAP	0.39	0.91	0.46	0.87	0.67	0.73
	MLE	0.39	0.91	0.47	0.86	0.72	0.69
	MLEF	0.39	0.91	0.47	0.87	0.69	0.71
MFI	EAP	0.38	0.91	0.45	0.88	0.65	0.75
	MAP	0.39	0.91	0.46	0.87	0.67	0.73
	MLE	0.39	0.91	0.47	0.87	0.72	0.69
	MLEF	0.39	0.91	0.47	0.87	0.69	0.71

Note. Standard deviations were below 0.005.

Figure 3
Classical Test Theory-Based SEM for Three Item Selection Methods



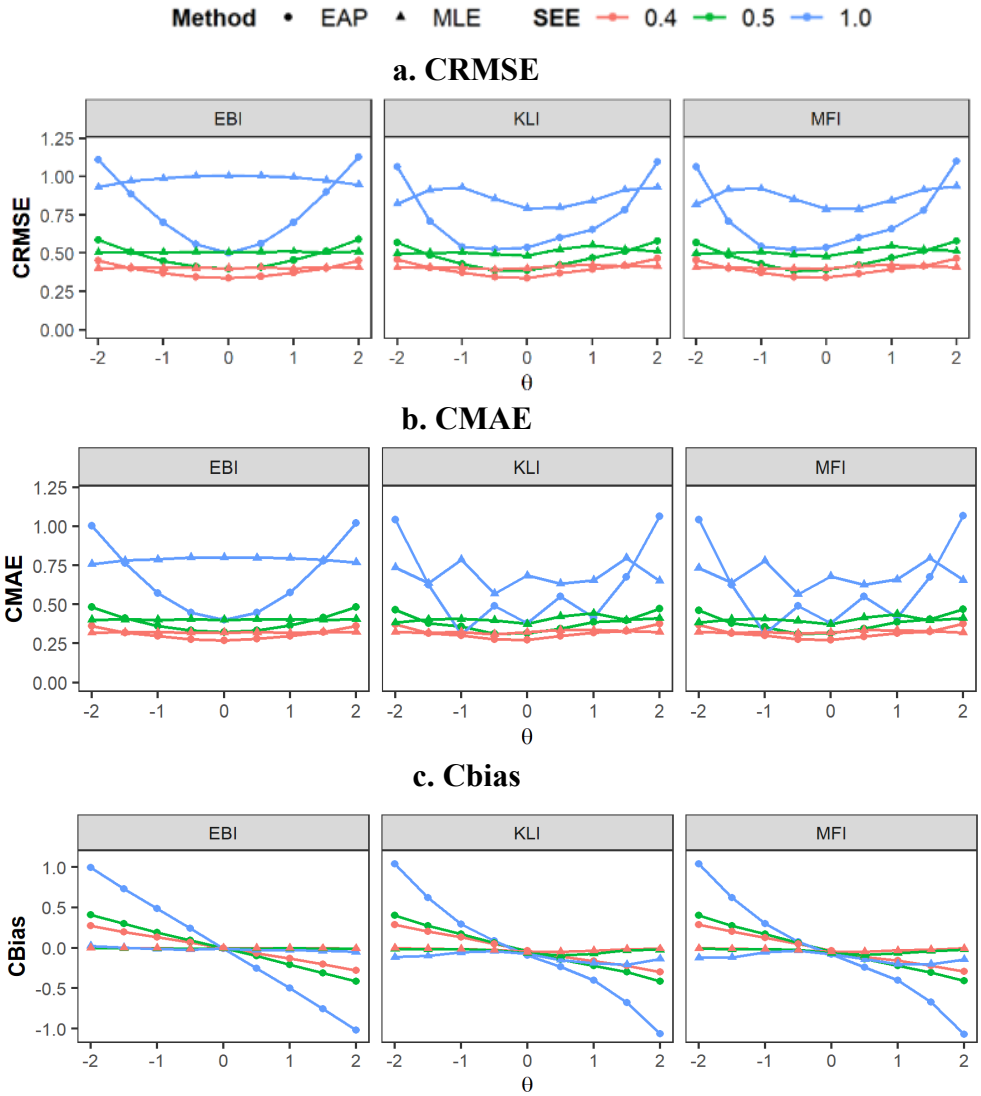
Note. Dots indicate the average CTT-SEMs across 25 replications for different score estimation methods to achieve SEE values of 0.4 (red), 0.5 (green) and 1.0 (blue). The horizontal dashed lines signify the predefined thresholds of SEE = 0.4, 0.5, and 1.0.

Beyond the CTT-based metrics, this study delved into the exploration of IRT-based conditional error measures including CRMSE, CMAE, Cbias, and test lengths across the θ spectrum. In IRT and CAT contexts, CRMSE and CMAE often serve, in simulations, as alternative indicators of CSEM. To maintain conciseness, this paper presents selected outcomes from the comprehensive simulation study. Consequently, the subsequent figures exclusively show the EAP and MLE score

estimation methods as the observed patterns of outcomes between EAP and MAP, as well as between MLE and MLEF, exhibited notable similarities.

Figure 4 illustrates the plots of CRMSE and CMAE against θ values for the three distinct SEE criteria. The trends observed for CRMSE (Figure 4a) and CMAE (Figure 4b) remained similar across each condition. However, it is noteworthy that, as detailed in Table 3, CMAEs consistently exhibited smaller values than CRMSEs throughout the entire θ spectrum, regardless of the simulation condition. This phenomenon arises due to the squaring of deviations between estimated and true θ values in CRMSE, leading to an increased weight on larger deviations and subsequently resulting in more significant differences even after taking the square root. Consequently, the two metrics are more aptly compared based on their patterns rather than their absolute values.

Figure 4
Conditional RMSE, MAE, and Bias Given θ for Three Item Selection Methods



In terms of trends, the EAP method displayed U-shaped patterns for both CRMSEs and CMAEs. This pattern indicated larger errors at the extreme θ areas (–2 or +2) and comparably

smaller errors—potentially even smaller than the SEE criterion—in the θ region closer to the center of the prior. As the SEE criterion increased, this U-shaped pattern became more pronounced, leading to significant disparities between SEE and CRMSE or CMAE. These disparities escalated as the θ values approached 0. Conversely, the MLE method showed relatively consistent CRMSEs and CMAEs across the entire θ range. Additionally, the MLE method exhibited smaller CMAEs than the predefined SEE values within the θ range from θ = –2 to +2, irrespective of the chosen item selection methods. These outcomes closely aligned with the results from CTT-SEM (see Figure 3).

When SEE = 0.4, the results showed that all item selection methods with MLE yielded impressively close CRMSEs of 0.39 to 0.42, precisely matching the targeted SEE. In contrast, when SEE = 1.0, the choice of item selection methods had varying impacts on CRMSE and CMAE. To illustrate, EBI combined with the MLE method yielded CRMSE and CMAE that were more consistent and similar to SEE across the entire θ range compared to MFI or KLI paired with the MLE method—this distinction was particularly pronounced when SEE = 1.0. The relatively diminished CRMSE and CMAE values observed for the MLE method at the extreme θ can be attributed to the truncation of the θ scale beyond –3.5 and 3.5. In essence, as revealed by CTT-SEM, SEE was found to be overestimated, subsequently revealing disparities between SEE and CSEM. The two distinct score estimation approaches exhibited varying impacts on CSEM across the θ spectrum, independent of the selected item selection methods and SEE criteria. Moreover, the choice of item selection methods exerted an influence on the correlation between SEE and CSME throughout the θ region, and this influence appeared to magnify as SEE values escalated.

The trends evident in the conditional biases presented in Figure 4c highlight distinct behaviors between the EAP and MLE methods. Specifically, the MLE method consistently estimated θ values closely aligned with the true θ across all θ regions. Conversely, the EAP method exhibited an overestimation bias within the lower θ region and an underestimation bias within the higher θ region. These fluctuations in biases across the θ spectrum contributed to significant shrinkage of the actual θ scale, detrimentally impacting measurement reliability (Han, 2012a). With the escalation of the SEE criterion, a concurrent amplification of conditional biases occurred at the extremes of the θ scale, regardless of the selected item selection methods. As SEE values increased, the biases associated with the MFI and KLI methods when paired with the MLE method manifested as negative biases. In other words, the θ values were underestimated within the framework of the MFI and MLE combination.

Figure 5 presents the conditional test lengths within a variable-length CAT setting. The distinct combinations of item selection methods and score estimation methods resulted in varying test lengths. For all other simulation factors, both MFI and KLI produced shorter test lengths in comparison to EBI. To elucidate, when aiming to attain an SEE of 0.4, the test lengths associated with MFI ranged from 7 to 10 items, while those linked with EBI ranged from 22 to 26 items (as detailed in Table 4). Furthermore, the EAP method yielded slightly shorter test lengths than the MLE method across all simulation conditions. As is widely recognized, the escalation of the SEE criterion correlated with a reduction in test lengths.

An analysis of the interplay among SEE, CSEM (specifically CRMSE and CMAE), and test length within the context of variable-length CAT simulation has unveiled significant relationships among these measures. Under the same SEE conditions, a positive relationship was evident between test length and CSEM. Specifically, when utilizing the EAP method, an increase in test length was associated with higher CSEMs as θ moved further away from the central prior. In

contrast, the MLE method showed consistent test lengths and CSEMs across diverse θ regions, as long as SEE was held constant. Thus, the findings suggest that controlling SEE in isolation demonstrated effectiveness in managing CSEMs across θ regions, while the range of test lengths remained relatively consistent or even decreased.

Table 3a
CRMSE and CMAE for SEE = 0.4

Item Selection	Score	CRMSE θ −2.0	−1.0	0.0	1.0	2.0	CMAE θ −2.0	−1.0	0.0	1.0	2.0
EBI	EAP	.45	.37	.34	.37	.45	.36	.30	.27	.30	.36
	MAP	.46	.37	.34	.37	.46	.37	.29	.27	.30	.37
	MLE	.40	.40	.40	.40	.41	.32	.32	.32	.32	.32
	MLEF	.39	.40	.40	.40	.40	.31	.32	.32	.32	.32
KLI	EAP	.46	.37	.34	.39	.47	.37	.30	.27	.32	.38
	MAP	.48	.37	.34	.40	.49	.39	.30	.27	.32	.40
	MLE	.41	.41	.40	.42	.41	.32	.32	.32	.34	.32
	MLEF	.40	.40	.41	.42	.40	.32	.32	.33	.33	.32
MFI	EAP	.46	.37	.34	.39	.47	.37	.30	.27	.32	.38
	MAP	.48	.37	.34	.40	.49	.39	.30	.27	.32	.40
	MLE	.41	.40	.40	.42	.41	.32	.32	.32	.33	.32
	MLEF	.39	.40	.40	.42	.40	.32	.32	.33	.33	.32

Table 3b
CRMSE and CMAE for SEE = 1.0

Item Selection	Score	CRMSE θ −2.0	−1.0	0.0	1.0	2.0	CMAE θ −2.0	−1.0	0.0	1.0	2.0
EBI	EAP	1.11	.70	.50	.70	1.13	1.00	.57	.40	.58	1.02
	MAP	1.12	.69	.50	.71	1.12	1.01	.57	.40	.58	1.02
	MLE	.93	.99	1.01	1.00	.95	.76	.79	.80	.80	.77
	MLEF	.86	.96	1.00	.97	.88	.70	.78	.80	.78	.71
KLI	EAP	1.06	.54	.54	.66	1.09	1.04	.31	.38	.41	1.06
	MAP	1.16	.54	.48	.66	1.19	1.14	.37	.36	.46	1.16
	MLE	.82	.93	.79	.84	.93	.74	.79	.68	.65	.65
	MLEF	.67	.83	.78	.94	.76	.49	.77	.64	.85	.53
MFI	EAP	1.06	.54	.54	.66	1.10	1.04	.31	.38	.42	1.07
	MAP	1.16	.54	.48	.66	1.18	1.14	.37	.36	.47	1.16
	MLE	.82	.92	.79	.85	.94	.73	.78	.68	.66	.66
	MLEF	.68	.84	.79	.94	.75	.49	.77	.65	.85	.53

Figure 5
Conditional Test Length for Three Item Selection Methods

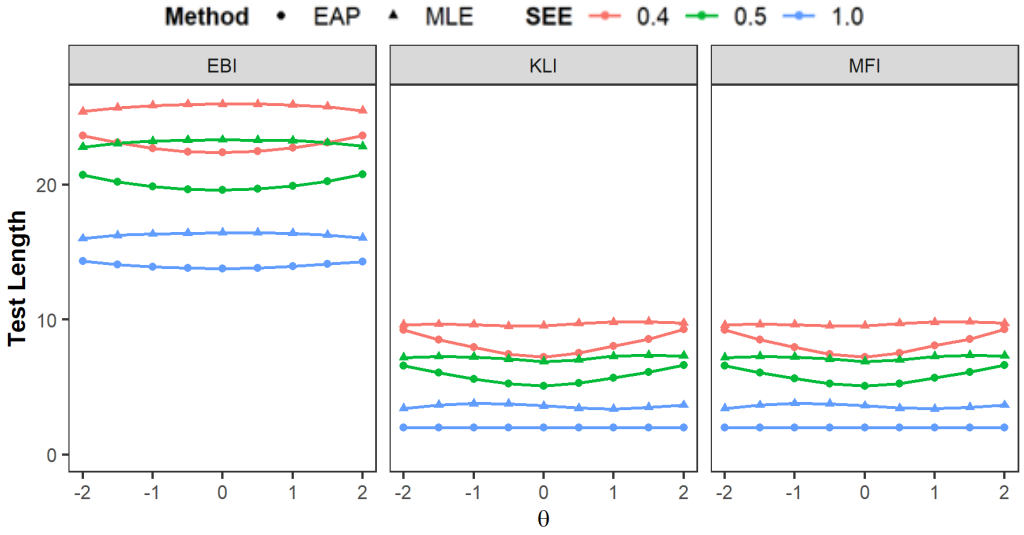


Table 4
Conditional Test Length for SEE = 0.4 and 1.00

Item Selection	Score	SEE = 0.4 θ −2.0	−1.0	0.0	1.0	2.0	SEE = 1.0 θ −2.0	−1.0	0.0	1.0	2.0
EBI	EAP	24	23	22	23	24	14	14	14	14	14
	MAP	24	23	22	23	24	14	14	14	14	14
	MLE	25	26	26	26	25	16	16	16	16	16
	MLEF	25	25	26	25	25	16	16	16	16	16
KLI	EAP	9	8	7	8	9	2	2	2	2	2
	MAP	9	8	7	8	9	2	2	2	2	2
	MLE	10	10	10	10	10	3	4	4	3	3
	MLEF	9	9	9	9	9	3	3	3	3	3
MFI	EAP	9	8	7	8	9	2	2	2	2	2
	MAP	9	8	7	8	9	2	2	2	2	2
	MLE	10	10	10	10	10	3	4	4	3	3
	MLEF	9	9	9	9	9	3	3	3	3	3

Discussion

The simulation suggests that the inconsistent relationship between SEE and empirical CSEM measures (i.e., CRMSE and CMAE), as illustrated by the extreme scenario in Figure 1 where SEE values were set at 0.38 (and similarly at 0.32 and 0.20), could manifest in other, more realistic conditions with different CAT settings. SEE tended to be an overestimation of CRMSE and CMAE for shorter tests but transitioned to underestimating CRMSE as tests lengthened (typically at six

items or more). When the test was longer than 20 items, however, SEE closely approximated CRMSE and CMAE. These results underscore the consistency of observed patterns across different SEE values, emphasizing the robustness of these relationships in variable-length CAT scenarios. These observations carry several important implications for variable-length CAT designs founded on the assumption that score precision is controlled by SEE. For short CATs, relying solely on SEE might lead to inconsistencies in actual score precision, suggesting the consideration of different SEE criterion values for varying test lengths.

For high-stakes exams, where the SEE criterion is typically set very rigorously (e.g., 0.3 or smaller), the SEE is expected to be a reliable approximation of the actual CSEM according to the results of this study. However, for low-stakes or brief quiz-style CATs, using an SEE criterion at a much higher level, such as 1.0, might be overshooting if the goal is to have CRMSE or CMAE at about 1.0, as seen in this study's examples. Resulting in a CRMSE or CMAE lower than the target itself might not be a bad thing. However, if another goal of the CAT-based test design was to minimize the test length, using the SEE alone as the termination rule might not offer the most optimal outcome. In such circumstances, considering the test length itself in addition to the SEE criterion might be advisable. For example, if one's test design objective is to keep CMAE under 1.0 while minimizing the test length, the SEE criterion when the test length is less than five items can be set at a much higher value, like 2.5, given the insights about the SEE-CMAE relationship from simulations like the one behind the example illustrated in Figure 1.

Also, it is important to make a distinction between CMAE and CRMSE, the two commonly used measures for assessing empirical score precision in simulation studies, rather than treat them interchangeably or directly compared. While the mathematical distinctions between these two measures are well-established, they are frequently treated as similar in practice, leading to inappropriate interpretations and comparisons. Beyond the theoretical inappropriateness of such practices, the study's results suggest that CRMSE values tend to closely align with the SEE value when CAT termination is based on SEE. In other words, when developing a CAT administration with SEE-based termination rules, CRMSE can be a more effective and intuitive method for evaluating and refining empirical score precision, given the chosen SEE criterion.

The appropriateness and utility of using a CTT-based score reliability index value for evaluating CAT quality and performance is a subject of debate. Nonetheless, it is often an anticipated outcome in CAT development, likely because a single-value score reliability index is easily communicable to test users and stakeholders. The results from this study, as shown in Table 2, generally indicate that the CTT-based SEM and reliability index values tend to be reasonably well-controlled when the SEE termination criterion is small (e.g., ≤ 0.5). However, with a larger SEE criterion (e.g., > 0.5), the CTT-SEM tends to increase at a significantly higher rate than the magnitude of SEE increase. This result can be attributed to the compound effect of the choice of SEE criterion and the test length, which is again primarily influenced by the SEE termination rule. In practical CAT simulations, such as the one conducted in this study, it is challenging to disentangle the effect of test length alone on CTT-SEM and score reliability from the compounded effect of test length and SEE. Nevertheless, as Figure 1 illustrates, the effect of test length on CTT-SEM (and score reliability) might be too pronounced to be effectively controlled by SEE alone, particularly when a CAT is designed to terminate quickly (e.g., test length < 7). This scenario is not uncommon in test modules for formative assessments and adaptive learning applications. Therefore, if the paramount objective is to ensure the CTT-SEM and score reliability, as reported to test users and stakeholders, it might be necessary to control both the test length and SEE simultaneously.

The longstanding debate between MLE and EAP, or any other Bayesian-based score estimation methods, continues in this study, along with their varying effects on the relationship between SEE and empirical CSEMs (CRMSE and CMAE). With MLE, as illustrated in Figure 4, the CSEMs were effectively and consistently controlled by SEE. However, the well-known tendency of EAP to introduce estimation biases toward the center of the prior distribution when the estimates are farther away from the prior center still persisted, even when the tests were adjusted to have consistent SEE values across the θ scale (Figure 4). This observation has an important implication: the systematic portion of estimation error, measured by Cbias, is not directly controlled by SEE. This is likely because the SEE is not directly related to the role of the prior, or in other words, the information contributed by the prior. Nevertheless, it should be noted that this does not imply that SEE is entirely unrelated to CSEMs when EAP is used. The influence of the prior with EAP diminishes as the test length increases. Given the apparent relationship between the test length and SEE, SEE can indirectly impact CSEMs and Cbias even when EAP is employed. Consequently, if the test length is designed or expected to be short, perhaps with a large SEE termination criterion (e.g., > 0.5), it might not be advisable to use EAP with a strong prior when consistent score precision is required. Instead, one should consider using MLE or Bayesian methods with a noninformative prior in such cases. Moreover, for shorter tests, alternative score methods, such as weighted likelihood estimation (WLE; Warm, 1989), merit consideration. This method has exhibited smaller biases and CMAE values than MLE and MAP, although it tended to overestimate at lower ability levels and underestimate at higher ability levels for the 3PL model (Warm, 1989). Further investigation into additional score methods will be conducted in future studies.

Furthermore, the simulation study results highlight the impact of the choice of item selection methods on the intricate relationship among SEE, empirical CSEM, and test length, as evidenced in Figures 4 and 5. In general, the association between SEE and CSEM appears to remain stable and consistent across the studied combinations of item selection criteria, provided that SEE is effectively controlled. This observation suggests potential generalizability, assuming the presence of a sufficiently large and diverse item bank to facilitate CAT item selection, aligning with the principles applied in this study. It is worth noting, however, that the capacity of the item bank itself consistently represents one of the most important factors in CAT administration when assessing its performance and behavior. If the item bank is constrained in size or not effectively aligned with the ability distribution, it can have adverse effects on CAT's optimality and adaptability, potentially resulting in extended test lengths. In such scenarios, the relationship between SEE and empirical CSEM might be affected due to the uneven test length experience by different examinees, as consistently revealed in this study's findings. In light of these considerations, it is advisable to conduct simulations as a means to attain a deeper understanding of CAT performance and behavior, particularly when employing a specific item selection algorithm and dealing with diverse item banks.

This study aimed to thoroughly investigate the SEE-CSEM relationship in variable-length CAT applications, including short tests for both low-stakes and high-stakes assessments. The findings have broader implications for longer CATs common in high-stakes contexts. Practical guidance for test developers and users emphasizes the effective use of SEE to enhance test score precision, especially in short CATs. Additionally, the results of this study illuminate interactions among critical CAT factors affecting SEE, SEM, and test lengths, advancing the understanding of measurement precision across diverse testing scenarios.

References

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), 1–18. CrossRef

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In Lord, F. & Novick, M. (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. CrossRef

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3)*,* 213–229. CrossRef

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37–53. CrossRef

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*(2), 129–143. CrossRef

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*(1), 61–77. CrossRef

Eggen, T. H. J. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249–261. CrossRef

Feldt, L. and Brennan, R. L. (1989). Reliability. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). The American Council on Education, MacMillan.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65-110). Rowman & Littlefield.

Han, K.T. (2012a), An efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement*, 49 (3), 225–246. CrossRef

Han, K. T. (2012b). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement, 36*(1), 64–66. CrossRef

Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement, 40*(4), 289–301. CrossRef

Harwell, M. (2018). A strategy for using bias and RMSE as outcomes in Monte Carlo studies in statistics. *Journal of Modern Applied Statistical Methods, 17*(2), eP2938. WebLink

Kolen, M. J. (2006). Scaling and Norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Rowman & Littlefield.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

MBA.com (n.d.). *Assessment Structure of the Executive Assessment.* WebLink

NCSBN: National Council of State Boards of Nursing (2019) NCLEX-RN examination: Test plan for the national licensure examination for registered nurses. WebLink

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100. CrossRef

Sulak, S., & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 315–326. CrossRef

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Lawrence Erlbaum Associates.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25,* 317–331. *CrossRef*

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(2), 109–135. *CrossRef*

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492. *CrossRef*

Yen, W. M., & Fitzpatrick A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-153). Rowman & Littlefield.

Yi, Q., Wang, T., & Ban, J. C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, *38*(3), 267–292. *CrossRef*

Author Address

Chansoon (Danielle) Lee *Email*: DLee@abim.org

Citation

Lee, C. & Han, K. T. (2024). Evaluating the effectiveness of the standard error of score estimation as a CAT termination criterion. *Journal of Computerized Adaptive Testing, 11*(2), 13-29.