Projection-Based Stopping Rules for Computerized Adaptive Testing in Licensure Testing

Applied Psychological Measurement

2018, Vol. 42(4) 275–290

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617726790

journals.sagepub.com/home/apm

Xiao Luo¹, Doyoung Kim¹, and Philip Dickison¹

Abstract

The confidence interval (CI) stopping rule is commonly used in licensure settings to make classification decisions with fewer items in computerized adaptive testing (CAT). However, it tends to be less efficient in the near-cut regions of the θ scale, as the CI often fails to be narrow enough for an early termination decision prior to reaching the maximum test length. To solve this problem, this study proposed the projection-based stopping rules that base the termination decisions on the algorithmically projected range of the final θ estimate at the hypothetical completion of the CAT. A simulation study and an empirical study were conducted to show the advantages of the projection-based rules over the CI rule, in which the projection-based rules reduced the test length without jeopardizing critical psychometric qualities of the test, such as the θ and classification precision. Operationally, these rules do not require additional regularization parameters, because the projection is simply a hypothetical extension of the current test within the existing CAT environment. Because these new rules are specifically designed to address the decreased efficiency in the near-cut regions as opposed to for the entire scale, the authors recommend using them in conjunction with the CI rule in practice.

Keywords

stopping rule, computerized adaptive testing, licensure testing, projection of future events

Introduction

Advancements in technology have given rise to the practical use of computerized adaptive testing (CAT), which takes advantage of powerful computers to select items and estimate parameters in real time to tailor the test for each test taker. Because CAT constantly customizes the test to match the test taker's ability during the test administration, it massively improves the testing efficiency by avoiding overly easy and hard items that provide little information to the measurement (McBride & Martin, 1983; Wainer, 1993; Weiss, 1982). Because of that, CAT has been used to administer various large-scale, high-stakes tests, including NCLEX, GRE, ASVAB, and GMAT to name a few. An operational CAT is typically implemented under the

¹National Council of State Boards of Nursing, Chicago, IL, USA

Corresponding Author:

Xiao Luo, Examinations Department, National Council of State Boards of Nursing, 111 East Wacker Drive, Suite 2900, Chicago, IL 60601-4277, USA.

Email: xluo@ncsbn.org

item response theory (IRT) framework (Lord, 1980; Lord & Novick, 1968), comprising of six major components: (a) an IRT model, (b) an item pool, (c) an entry point, (d) an item selection rule, (e) an ability estimation rule, and (f) a stopping rule (Weiss & Kingsbury, 1984). The present study is a discourse on the CAT stopping rule in licensure settings, where the purpose of the test is to make correct pass/fail classification decisions for test takers.

In essence, CAT is operated in the “item selection and ability estimation” cycles, and in each cycle, the item selection rule is triggered to select an item for administration most suited to the test taker’s latest θ estimate and the ability estimation rule is triggered to update the test taker’s θ estimate after the administration of that item. The stopping rule is essential because it determines when/how to exit the repetitive cycles. When the test is terminated too early, it could report an inaccurate measurement and lead to an erroneous classification decision. When terminated too late, the test could administer extra items without any substantial improvement on the measurement. This not only decreases the testing efficiency but also raises concerns over the elevated item exposure rates and potential test security breaches (e.g., item harvesting and brain dumping). That being said, the perfect exit point is difficult to identify in a variable-length CAT, and there are various methods in the literature.

The first method (the *SE* rule) is to stop the test when the standard error (*SE*) of the θ estimate falls below a threshold, which implies that the intended precision is attained and the measurement is deemed accurate enough for the testing purpose (Weiss, 1982). For instance, the probability of getting a correct response in the three-parameter-logistic (3PL) model is given by

$$P(\theta) = c + \frac{1 - c}{1 + \exp[-1.7a(\theta - b)]}, \quad (1)$$

where θ is the ability parameter of the test taker, and a , b , and c are the discrimination, difficulty and pseudoguessing parameters of the item. The Fisher’s information of item i is given by

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \quad (2)$$

where $Q_i = 1 - P_i$, and P'_i is the first derivative of the P_i . As a result, the *SE* of the θ estimate is given by

$$SE = \frac{1}{\sqrt{\sum_i I_i(\theta)}}. \quad (3)$$

The notion of test reliability in the classical test theory is replaced by the notion of information in the IRT, and the relationship between *SE* and test reliability is described as follows:

$$SE = S_\theta \sqrt{1 - \rho_{xx'}^2}, \quad (4)$$

where $\rho_{xx'}^2$ is the test reliability, and S_θ is the standard deviation of θ s of the population. This equation is particularly useful for choosing a threshold for the *SE* stopping rule. For example, assuming $S_\theta = 1.0$, a *SE* threshold of 0.32 roughly means a reliability of 0.90, and vice versa.

Another method (the *MI* rule) is to discontinue the test when none of the items in the pool exceeds the minimal information (*MI*) threshold (Dodd, Koch, & De Ayala, 1989; Gialluca & Weiss, 1979). Because of the reciprocal relationship between *SE* and information (see Equation 3), an item is not expected to improve the θ estimate in a meaningful way if the information is too small. The third method (the θ convergence rule) is to directly monitor the change in the θ estimate over the course of the test and terminate the test when the θ estimate converges—that

is, the change in θ drops below the convergence threshold (Weiss & Kingsbury, 1984). The fourth method (the PSER rule) is a more sophisticated one that utilizes the predicted standard error reduction (PSER) of the θ estimate to determine whether to stop the test or not (Choi, Grady, & Dodd, 2011). It is used in conjunction with the *SE* rule with two additional parameters—the hyper and hypo parameters. Specifically, when the *SE* threshold has been achieved, continue the test until the PSERs of remaining items in the pool are all below the hyper threshold parameter, meaning that no item could improve the θ estimate meaningfully any more. When the *SE* threshold has not been achieved, stop the test when the PSERs are all below the hypo threshold parameter.

Studies have been conducted to compare these stopping rules. Dodd et al. (1989) and Dodd, Koch, and De Ayala (1993) found that the *SE* rule was more efficient than the MI rule in terms of the test length and θ recovery quality. Babcock and Weiss (2012) confirmed that the *SE* rule outperformed the MI rule and the θ convergence rule in a simulation study; however, they recommended combining different stopping rules and adding the minimum test length constraint in practice. Choi et al. (2011) argued that the PSER rule was more efficient than the *SE* and MI rules by using a balanced evaluation of test length and measurement precision.

Because the purpose of the licensure test is to make correct classification decisions, the confidence interval (CI) stopping rule is commonly used in licensure settings (Weiss & Kingsbury, 1984). Under this rule, a 100 $(1 - \alpha)\%$ CI with respect to $\hat{\theta}$ is developed and compared with the cut score, where α is the size of Type I error. The test is terminated when the CI is not inclusive of the cut score, which suggests that a clear classification decision can be made with $\alpha\%$ chance of Type I error. This rule requires the “near-cut” test takers whose true abilities are close to the cut score to have more information and the “away-cut” test takers whose true abilities are distant from the cut score to have less information. In essence, this rule can be viewed as a special *SE* stopping rule in which the threshold is conditional on θ . In the near-cut regions, a lower threshold is required for making “confident” classification decisions, whereas in the away-cut regions, a higher threshold is required.

The conditional *SE* requirement can be transformed to the conditional information requirement using Equation 3. Figure 1 illustrates the information required for making classification decisions against three cut scores along the θ scale ($\alpha = .05$): -1 , 0 , and 1 . As expected, the information requirement increases with an accelerated velocity as θ edges toward the cut score and eventually grows to the infinity at the cut score. When $a = 1$ and $c = 0$, the maximum information of an item is 0.7225 in the 3PL model. Thus, to achieve a test information of 50 , 75 , and 100 , the test demands approximately 69 , 104 , and 138 items, respectively. This theoretical speculation is confirmed by an observation of a large-scale licensure examination, where most of the near-cut test takers exited the CAT by the “maximum length” rule instead of the more efficient “95% CI” rule (see Figure 2).

This kind of near-cut inefficiency in the CI rule leads to the proposition of a set of new stopping rules in this study. The new rules aim to use projected future events (i.e., unobserved items and responses) in the remaining part of the CAT to predict the range of the final θ estimate and use the θ range to terminate the CAT more efficiently. The future events are largely untapped resources in existing stopping rules, but this study demonstrates how to utilize them to facilitate the termination decisions. Take an extreme case for example. When 29 items have been administered in a maximally 30-item CAT, it is certain that at most one more item remains in CAT with two possible responses: correct vs. incorrect. With the item-balancing constraint, it can be inferred using the item selection rule that the item has to come from a specific content area. When additional constraints are imposed (e.g., speededness, item exposure control, answer key distribution, item type distribution, etc.), the scope of the next item may be further reduced to a small subset of the item pool, which allows one to explore all possible future events (perhaps

Figure 1. The minimum information requirements under the 95% confidence interval ($\alpha = .05$) stopping rule.

Figure 2. Stopping rules used in a large-scale licensure examination ($N = 1,000$): The cut score is denoted by the dashed line.

Note. CI = confidence interval.

with a brute force method) and find all possible final θ estimates. If the θ range excludes the cut score, the test can be stopped, because it means that the classification decision remains the same regardless of the last item and response; otherwise, continue the test to administer the last item.

This rationale can be generalized to the case with m remaining items in a maximally n -item CAT but requires more efficient methods for projecting the future events than the brute force method. When the projected θ range is narrower than the $(1 - \alpha)\%$ CI, it may allow for a more efficient early termination decision than the CI rule. Along this line, three projection-based stopping rules are introduced in this study. The first rule (R1) projects an m -item subtest (using unused items in the pool) that maximizes the information at the current θ estimate while satisfying all item selection constraints. These items are most likely to be observed in the unobserved remainder of the test, given observed past events. The lower- and upper-bound of the final θ estimate are subsequently developed by projecting all incorrect and correct responses to these

items, respectively, because these two response patterns result in the lowest and highest θ estimates among all 2^m possible response patterns.

Because the past events may be an inaccurate indicator of the future events, especially in the early stage of the CAT, the second rule (R2) proposes a more conservative projection method that uses two m -item subtests with maximum and minimum difficulty (sum of b parameters) to approximate the boundaries of the future event space in terms of their influence on the θ estimate. Similar to R1, it develops the lower- and upper-bound of the θ estimate by projecting all incorrect responses to the least difficult subtest and correct responses to the most difficult subtest. While R2 may be too conservative to have any practical utility, the third rule (R3) approximates the boundaries with two fairly, instead of the most, difficult/easy subtests, which are operationally defined as having the average difficulty equal to $\hat{\theta} \pm 2 \cdot SE$. The lower- and upper-bound of the θ estimate are developed in the same way as in R2.

The remainder of this article describes a simulation study and an empirical study analyzing the data of a large-scale CAT licensure examination that investigate the characteristics and performance of these new projection-based stopping rules in licensure settings. Specifically, this study intends to address three research questions as follows:

Research Question 1: Would the projection-based stopping rules show any advantage over the CI rule in licensure settings?

Research Question 2: Would the projection-based stopping rules sacrifice the measurement and classification precision while striving to shorten the test length?

Research Question 3: For which subgroups of the testing population are the projection-based stopping rules more useful than the CI rule?

The Simulation Study

Method

Study design. Because the test length determines how far the projection is, two conditions of varying maximum test length were included in the simulation study: 60 vs. 80 items. In each condition, R1 to R3 described above were compared with the 95% ($\alpha = .05$) and 99% ($\alpha = .01$) CI rules. The more lenient 95% CI rule is expected to administer fewer items yet be more error-prone than the 99% CI rule. All stopping rules only became effective after the half of the total items had been administered to ensure that a minimum measurement precision was attained. In other words, the minimum test length was 30 and 40 items under the two conditions. This kind of minimum length criterion is not uncommon in research and practice (e.g., Babcock & Weiss, 2012; National Council of State Boards of Nursing, 2016, pp. 15). For convenience, these two conditions were denoted as the 30/60 and 40/80 conditions in this study.

Simulation overview. The cut score was arbitrarily set to $\theta = 0$. Because the near-cut regions are where the testing efficiency decreases in the CI rule, the simulations were conducted only for the central region of the θ scale, from -1.0 to 1.0 in increments of 0.1 . A hundred replications were conducted at each θ point to temper the sampling error occurring in simulations. The simulation was implemented in R (R Core Team, 2016) using the *xxIRT* package (Luo, 2016).

Item pool. A pool of 600 items was generated with a , b , and c parameters, and each item was randomly assigned to one of the eight content areas. The size of the pool was 7.5 (in the 40/80 condition) or 10 (in the 30/60 condition) times as large as the maximum test length to provide sufficient items to simulations. Table 1 summarizes the characteristics of the generated item pool. Figure 3 shows the average information functions of the item pool and each content area.

Table 1. Descriptive Statistics of the Item Parameters in the Pool.

Parameters	M	SD	Minimum	Maximum
<i>a</i> (discrimination)	1.01	0.14	0.66	1.44
<i>b</i> (difficulty)	0.05	1.00	−3.18	3.11
<i>c</i> (pseudoguessing)	0.11	0.05	0.02	0.37

Figure 3. The average information functions of the item pool and content areas.

Item selection. A modified constrained-CAT (C-CAT) algorithm (Kingsbury & Zara, 1989, 1991) was used to select items under the content-balancing constraint, which selects a content area first and an item in that content area afterward. The first 10 content areas were randomly selected to randomize the initial sequence of content areas, and the rest were the content areas with largest discrepancies from the targets (in percentage). The randomness in the 10 initial items helps avoid the same sequence of content area selections. After each selection of the content area, the algorithm selects an item with maximum $I(\hat{\theta})$ in that content area. Although the content area had a nearly uniform distribution in the pool, the targets were set unevenly to slightly elevate the constraint severity (see Table 2).

Response generation. For the administration of the selected item, the probability of getting a correct response based on the true θ was first calculated using the 3PL model (see Equation 1). The response, u , was then generated by comparing the model-based probability, p , with a random number, x , from the uniform distribution, $U(0, 1)$, as follows:

$$u = \begin{cases} 1, & \text{when } p \geq x \\ 0, & \text{when } p < x \end{cases} \quad (5)$$

Ability estimation. The maximum likelihood estimator (MLE) was the primary θ estimator. However, when the observed responses are all correct or incorrect, MLE is inappropriate because the θ estimate would be positive or negative infinity. Instead, the expected a priori (EAP) estimator was applied for those response patterns with an uninformative prior distribution: $\mu = 0$ and $\sigma = 2$. Because the expected probability of getting a correct response is .5 for each item in CAT, the probability of k consecutive correct or incorrect responses is $.5^k$. Therefore, it is extremely unlikely to observe a large number of consecutive same responses, and EAP was only used for first few items early in the test.

Table 2. Content Area Distribution.

	Content areas							
	1	2	3	4	5	6	7	8
Pool								
%	0.12	0.10	0.12	0.14	0.13	0.12	0.12	0.15
<i>n</i>	69	62	73	85	80	71	70	90
Targets								
%	0.25	0.20	0.15	0.10	0.10	0.10	0.05	0.05
<i>n</i> (60 items)	15	12	9	6	6	6	3	3
<i>n</i> (80 items)	20	16	12	8	8	8	4	4

Stopping rule. Three projection-based stopping rules (R1 to R3) were compared against the 95% ($\alpha = .05$) and 99% ($\alpha = .01$) CI rules. For the interest of this study, they were calculated in a retrospective manner. That is, the simulation first proceeded to the maximum test length without applying the stopping rules. The stopping rules were applied afterward to each position of the CAT between the minimum and maximum length to “retrofit” the projected θ range (under R1 to R3) or the CI (under the 95% and 99% CI rules), which were used to find the exit points and final θ estimates of the test. For R1 to R3, the projected m -item subtests were assembled from the unused items in the pool using the mixed integer linear programming (MILP) automated test assembly (ATA) algorithm provided in the *xxIRT* package. In particular, the MILP ATA algorithm maximizes $\sum_i I_i(\hat{\theta})$ under R1, maximizes (for the most difficult subtest) and minimizes

(for the least difficult subtest) $\sum_i b_i$ under R2, and minimizes $\left| \sum_i b_i - (\hat{\theta} \pm 2 \cdot SE) \right|$ under R3, while satisfying all test length and content-balancing constraints. The projected items and responses were appended to administered items and responses to develop the lower- and upper-bound of the θ estimate using the same MLE/EAP estimation method described above.

Results

Figure 4 provides an example of how the projection-based rules differ from the CI rule using the simulation data randomly selected from the 95%-CI 40/80 condition (true $\theta = 0$). In the figure, circles represent the θ estimates (size proportional to SE), and dotted lines represent the CI (under the CI rule) or the projected θ range (under R1 to R3) at each position between the minimum and maximum test length. Because the item selections and responses are independent of the stopping rule in CAT, the circle sizes and locations were identical across four graphs. R1 to R3 resulted in wider θ ranges than the CI rule in the earlier stage of the CAT where the projection involved a relatively large number of future events, but narrower θ ranges in the later stage where the future events were well confined to a relatively small space.

Specifically, the width of the CI was reduced from .70 (40th position) to .51 (79th position) under the 95% CI rule, whereas the width of the projected θ range was reduced from 1.33 to 0.02 under R1, from 4.39 to 0.03 under R2, and from 1.79 to 0.01 under R3. R1 to R3 started to have a projected θ range narrower than the CI from 61st, 67th, and 60th position respectively, and an early termination decision could be made at 75th, 71st, 73rd position under R1, R2, and R3 for this simulee, but not even at the last position under the CI rule. These results supported the fundamental rationale behind the projection-based rules that the projected θ range could become narrower than the CI to facilitate the early termination decisions in CAT. A secondary

Figure 4. The CI (under the CI rule) and the projected θ range (under R1 to R3) in a simulation randomly selected from the 95%-CI 40/80 condition (true $\theta = 0$).
 Note. CI = confidence interval.

finding is that unlike the CI that is always symmetrical, the projected θ range developed under R1 to R3 is often asymmetrical. The difference between the upper interval (the distance from the θ estimate to the upper bound) and the lower interval (the distance from the θ estimate to the lower bound) ranged from 0.02 to 0.25 under R1, from 0.03 to 2.31 under R2, and from 0.01 to 0.83 under R3, whereas it remained 0 under the CI rule.

Figure 5 indicates the final test length under these stopping rules. Consistent with the hypothesis, R1 to R3 administered fewer items than the CI rules in the near-cut region of the scale. This advantage of the projection-based rules was more evident when R1 to R3 were compared with the 99% CI rule than the 95% CI rule. Also as expected, the projection-based rules did not necessarily discontinue the CAT earlier than the CI rules in the away-cut regions. In the upper-end of the scale, R1 to R3 resulted in longer tests than the 95% and 99% CI rules. The maximum test length did not affect results remarkably, as similar patterns were observed in the 30/60 and 40/80 conditions. A comparison of R1 to R3 revealed that R1 and R3 exhibited quite comparable performance, although they methodologically stem from distinct rationales. The more conservative R2, on the contrary, delivered much less efficient performance in the upper-end of the scale than R1 and R3.

How many items were saved under the projection-based rules? Due to the similarity between the 30/60 and 40/80 conditions, the authors focus their discussion on the 30/60 condition to avoid redundancy, and the discussion could be easily generalized to the 40/80 condition. Table 3 summarizes the test length differences between R1 to R3 and the CI rules, where a positive number means a longer test under R1/R2/R3 and a negative number a shorter test. The results implied substantial benefits under the projection-based rules. R1 to R3 could shorten the test

Figure 5. The number of items administered in the CAT under different stopping rules.
 Note. CAT = computerized adaptive testing; CI = confidence interval.

length over the region $[-0.7, 0.2]$, $[-0.7, 0.1]$, and $[-0.6, 0.2]$, respectively, when compared with the 95% CI rule, and over the region $[-1.0, 0.5]$, $[-1.0, 0.2]$, and $[-0.9, 0.4]$ when compared with the 99% CI rule. In the best case, R1 to R3 administered 6.45 ($\theta = -0.2$), 10.68 ($\theta = -0.2$), and 9.29 ($\theta = -0.2$) fewer items than the 95% CI rule, and 11.63 ($\theta = -0.3$), 15.36 ($\theta = -0.3$), and 13.93 ($\theta = -0.3$) fewer items than the 99% CI rule. Conversely, in the worst case, they administered 3.69 ($\theta = 0.6$), 16.32 ($\theta = 0.7$), and 6.63 ($\theta = 0.6$) more items than the 95% CI rule, and 1.70 ($\theta = 0.7$), 15.18 ($\theta = 0.8$), 4.52 ($\theta = 0.7$) more items than the 99% CI rule.

The testing efficiency should not only evaluate the test length but also the measurement quality. Therefore, the final θ estimates reported at the exit point under different stopping rules were averaged over the 100 replications of the simulations and compared with true θ s (see Figure 6). The estimated and true θ s exhibited a strong linear relationship across the θ scale for all stopping rules, even at points $\theta = -0.2$ and -0.3 where R1 to R3 substantially shortened the test. The results implied that the θ estimation under R1 to R3 was as precise as that under the 95% and 99% CI rules, and more importantly, that R1 to R3 improved the testing efficiency by administering fewer items but not sacrificing the θ recovery quality.

Similarly, the final classification decisions reported at the exit point were analyzed by the means of the passing rates. Figure 7 presents the observed passing rates, averaged over 100 replications of the simulations, under different stopping rules. If without the measurement error, the passing rates should be 0% when the true θ is less than or equal to cut score ($\theta = 0$) and 100% when greater than the cut score. Because of the measurement error, an S-shaped curve is expected in practice. It was found in Figure 7 that more classification errors were observed in the near-cut region (approximately from -0.3 to 0.3) than the away-cut regions and that the error rate reached the maximum at the cut score. More importantly, the passing rates were very comparable between R1 to R3 and the CI rules across the scale, which suggests that the

Table 3. Test Length Differences Between R1 to R3 and the CI Rules in the 30/60 Condition.

θ	95% CI			99% CI		
	R1	R2	R3	R1	R2	R3
-1.0	0.08	0.04	0.11	-0.03	-0.07	0.00
-0.9	0.15	0.11	0.21	-0.15	-0.19	-0.09
-0.8	0.19	0.15	0.34	-0.53	-0.57	-0.38
-0.7	-0.04	-0.25	0.04	-2.06	-2.27	-1.98
-0.6	0.13	-0.65	-0.14	-3.87	-4.65	-4.14
-0.5	-0.39	-2.54	-1.54	-5.89	-8.04	-7.04
-0.4	-2.92	-5.58	-4.35	-9.21	-11.87	-10.64
-0.3	-3.40	-7.13	-5.70	-11.63	-15.36	-13.93
-0.2	-6.45	-10.68	-9.29	-10.49	-14.72	-13.33
-0.1	-4.60	-8.37	-7.35	-7.91	-11.68	-10.66
0.0	-4.90	-6.09	-6.27	-7.12	-8.31	-8.49
0.1	-3.53	-1.37	-3.61	-7.61	-5.45	-7.69
0.2	-2.07	3.65	-0.64	-7.67	-1.95	-6.24
0.3	1.55	8.95	3.53	-7.05	0.35	-5.07
0.4	2.74	11.13	5.06	-6.21	2.18	-3.89
0.5	3.52	14.31	6.07	-1.28	9.51	1.27
0.6	3.69	15.86	6.63	1.52	13.69	4.46
0.7	3.09	16.32	5.91	1.70	14.93	4.52
0.8	1.76	15.61	4.37	1.33	15.18	3.94
0.9	1.21	15.03	3.44	0.74	14.56	2.97
1.0	0.56	14.60	2.67	0.43	14.47	2.54

Note. CI = confidence interval.

Figure 6. The relationship between true and estimated θ s under different stopping rules.

Note. CI = confidence interval.

Figure 7. The observed passing rates under different stopping rules.
 Note. CI = confidence interval.

projection-base rules would not deteriorate the precision of classification decisions after shortening the test length.

The Empirical Study

Method

Can the promise of the projection-based rules be realized in a real licensure examination as in the simulation study? To answer that question, 1,000 records that reached the maximum length were randomly retrieved from a large-scale CAT licensure examination. This licensure examination requires a minimum of 60 items and a maximum of 250 items. The passing standard is set to $\theta = 0$. The examination places three stopping rules in operation: 95% CI, maximum length, and run out of time (6-hr limit). The sample used in the analysis only included the tests exiting the CAT from the maximum length rule. Figure 8 presents the θ distribution in the sample of the empirical data, where the vast majority of θ s in the maximum-length tests lies between -0.2 and 0.2 . This agreed with the observation shown in Figure 2, where the maximum-length tests occurred primarily in the near-cut region.

The data involved four individual item pools, and items in the pools were calibrated using the Rasch model which is mathematically equivalent to the model in Equation 1 with $a = 0.58$ and $c = 0.00$. The same C-CAT item selection algorithm was used in operation to reach the balanced content distribution in eight areas as such 20%, 12%, 9%, 9%, 9%, 15%, 12%, and 14%. As in the simulation study, R1 to R3 were applied to these tests to “retrofit” the exit points and final θ estimates without modifying the actual item selections and responses.

Figure 8. The distribution of the final θ estimates in the sample of the empirical data.

Figure 9. The number of items saved under R1 to R3.

Results

Figure 9 presents the number of items that the tests were shortened by under R1 to R3, where the x axis is the original final θ estimate at the maximum length in the empirical data and the y axis is the number of items saved under R1 to R3. When applied to the real licensure examination, the projection-based rules were able to reduce the test length appreciably. Under the best circumstances, R1 administered 25 fewer items than the maximum length, R2 14 fewer items, and R3 24 fewer items. In general, a V-shape curve can be used to describe the relationship between the final θ estimates and the number of saved items, where the number of saved items increased monotonically along with the distance between θ and the cut score. The advantage of R1 to R3 gradually diminished as θ moved closer to the cut score. Consistent with the finding in the simulation study, R1 and R3 delivered quite similar results. Because R2 is by design the most conservative method among all, it unsurprisingly saved less items than R1 and R3. On average, R1 saved 9.7 items ($SD = 6.3$), R2 5.0 items ($SD = 3.4$), and R3 8.6 items ($SD = 5.6$).

Figure 10 presents the relationship between the original θ estimates reported at the maximum length in the empirical data (the x axis) and the new θ estimates reported under R1 to R3 (the y

Figure 10. The relationship between the original (in the empirical data) and new (obtained under R1 to R3) final θ estimates.

axis). The original and new θ estimates had a strong linear relationship, showing a Pearson's correlation of .96 (R1), .98 (R2), and .97 (R3). It suggested that the θ estimates reported—probably from shorter tests—under R1 to R3 were very similar to those originally reported at the maximum length. Because the cut score is $\theta = 0$ in this examination, a data point in the first and third quadrants in Figure 10 means that the classification decision resulted from R1 to R3 agrees with the original decision, whereas a data point in the second and fourth quadrants means a different decision (i.e., false positive and false negative, respectively). Except for one false positive case under R1, all data points were in the first and third quadrants in these graphs, indicating that the decisions under R1 to R3 well agreed with the original decisions. Put together, if the projection-based rules were implemented in operation to replace (or supplement) the CI rule, the same classification decisions could be made with short tests for most of these test takers.

Discussion and Conclusions

The CI stopping rule is commonly used in licensure settings to make classification decisions with fewer items in CAT. However, this rule tends to be less efficient in the near-cut regions of the scale, as the CI fails to be narrow enough for an early termination decision prior to reaching the maximum test length (see Figure 2). To solve this problem, this study proposed the projection-based stopping rules that base the termination decisions on the algorithmically projected range of the final θ estimate at the hypothetical completion of the CAT. Both the simulation and the empirical study described above evidenced the advantages of the projection-based rules over the CI rule in the near-cut regions by showing that they effectively reduced the test length without jeopardizing critical psychometric qualities of the test, such as the θ and classification precision (see Figures 5 to 7). When the projection-based rules were applied to a real licensure examination, the test length could be reduced by up to 25 items in a 250-item examination while obtaining the same classification decisions (see Figures 9 and 10).

It should be noted that the CI rule is still a reasonable and efficient stopping rule in general. The projection-based rules are nevertheless developed to compensate for its decreased efficiency in the near-cut regions, where the shrinkage of the CI usually fails to keep abreast with

the increment of the information required for classifications (see Figure 3). Instead of relying on the CI, the projection-based rules rely on the projected range of the final θ estimate developed from the algorithmically projected future events in the remainder of the test under all practical limitations imposed by the item selection rule, the ability estimation rule, and other operational CAT setup in the test. The utilization of the future events highlights an essential distinction of the projection-based rules from the existing rules. The existing rules (e.g., the *SE*, *MI*, and θ convergence rules) hardly take the future items and responses in the unobserved part of the test into account in their algorithms. Although the future items and responses are unobserved and difficult to pinpoint, this study demonstrated how to use them to curb the final θ estimate under operational limitations of the CAT. As the limitations become increasingly stringent in the later stage of the test, it could narrow the θ estimate down to a range that may allow for a better termination decision than the CI rule (see Figure 4).

Another distinguishing characteristic of the projection-based rules is the focus on the observed θ as opposed to the true θ . The rationale behind the *SE* and *MI* rules is to stop the test when the true θ is approximated closely enough (regulated by the *SE* threshold) or could not be approximated more closely (regulated by the *MI* threshold). Similarly, the essence of the CI rule is to develop an interval of the true θ and compare it against the cut score. However, the true θ is always an unobserved entity in reality, and the decision is eventually based on the observed θ . The projection-based rules attempt to directly predict the range of the observed θ at the end of the test under the existing CAT environment. The convergence rule is more relevant to these new rules than the *SE*, *MI*, and CI rules in the sense that it also directly monitors the observed θ and terminates the test when it converges. The difference is that the impact of future events on the observed θ is taken into account in the new rules but not in the θ convergence rule.

Given the common belief that more items lead to better measurement, why is it possible to reduce the test length without sacrificing the measurement quality? This is because that the purpose of the stopping rule is to prune redundant items as opposed to essential items. Items that add a sizable amount of information and bring forth meaningful changes in the θ estimate are essential to the test. Conversely, items that cause the random oscillation of the θ estimate are deemed redundant to the test. Taking the data in Figure 4 for example, the last 20 items reduced *SE* from 0.15 to 0.13 and the last 10 items from 0.14 to 0.13. The θ estimate was closer to the true value (0.00) at 56th position (0.00) than the 75th position (−0.07). Therefore, the last few items might be considered as redundant, as opposed to essential, items. Removing those redundant items would not necessarily damage the measurement quality of the test. We can also view different stopping rules through the lens of item redundancy. Items are redundant if they are administered after the required precision has been achieved in the *SE* rule, or would not yield significant changes in the θ estimate in the *MI* and θ convergence rules. In the projection-based rules, items are redundant if they could not possibly reverse the classification decisions. This explains why it is realistic for the projection-based rules to shorten the test length but not damage the measurement quality.

The critical underpinning of the projection-based rules is the algorithm used to project the future events and the final θ estimate. The three algorithms described in the study were shown to be fairly useful, at least as a fine complement to the CI rule. R2, which derived from a rather conservative rationale, unsurprisingly showed less bold reductions in the test length than R1 and R3. R1 and R3 delivered quite comparable performances but were derived from different methodological rationales—that is, R1 is the projection of the most likely events and R3 is a relaxation of R2. One caveat learned in developing these rules is that the projection should limit the involvement of the observed past events, because a careless use of them could turn the projection of the future events into the reinforcement of the past events. For instance, if the current θ

estimate was used to simulate future responses, the final θ estimate would likely be very similar to the current one.

To conclude, the projection-based stopping rules were introduced in this study for CAT in licensure settings to address the decreased efficiency in near-cut regions in the CI rule. Results from a simulation study and an empirical study indicated that these new rules could effectively reduce the test length while maintaining the comparable measurement quality, when compared with the CI rule. This study also demonstrated that the unobserved future events in CAT, which are normally overlooked in the existing stopping rules, could be useful for facilitating the testing efficiency. Operationally, these rules do not require additional regularization parameters (e.g., *SE*, *MI*, convergence threshold parameters), because the projection is simply a hypothetical extension of the current test within the existing CAT environment. Because these new rules are specifically designed to address the decreased efficiency in the near-cut regions as opposed to for the entire scale, the authors recommend using them in conjunction with the CI rule in practice.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

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, 1-18.
- Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, 71, 37-53.
- 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.
- Gialluca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement* (Research Report No. RR-79-6). Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359-375.
- Kingsbury, C. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Luo, X. (2016). *xxIRT: Practical item response theory and computer-based testing in R*. Retrieved from <https://cran.r-project.org/package=xxIRT>
- McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Eds.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). Cambridge, MA: Academic Press.

- National Council of State Boards of Nursing. (2016). *2016 NCLEX candidate bulletin*. Retrieved from https://www.ncsbn.org/089900_2016_Bulletin_Proof3.pdf
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from <https://www.R-project.org/>
- Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12, 15-20.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.
- Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361-375.