Comparison of real data and simulated data analysis of a stopping rule based on the standard error of measurement in computerized adaptive testing for medical examinations in Korea: a psychometric study

Dong Gi Seo¹,²*, Jeongwook Choi,¹,² Jinha Kim¹

¹Department of Psychology, Hallym Applied Psychology Institute, College of Social Science, Hallym University, Chuncheon, Korea

²The CAT Korea Company, Chuncheon, Korea

Abbreviated title: Stopping algorithm in a CAT in medical examination

*Corresponding email: dgseo@hallym.ac.kr

Editor: Sun Huh, Hallym University, Korea

Received: June 13, 2024; Accepted: June 27, 2024; Published: July 9, 2024

This article is available from: http://jeehp.org

© 2024 Korea Health Personnel Licensing Examination Institute

(cc) This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the
Abstract

Purpose: This study aimed to compare and evaluate the efficiency and accuracy of computerized adaptive testing (CAT) under two stopping rules (SEM 0.3 and 0.25) using both real and simulated data in medical examinations in Korea.

Methods: This study employed post-hoc simulation and real data analysis to explore the optimal stopping rule for CAT in medical examinations. The real data were obtained from the responses of 3rd-year medical students during examinations in 2020 at Hallym University College of Medicine. Simulated data were generated using estimated parameters from a real item bank in R. Outcome variables included the number of examinees' passing or failing with SEM values of 0.25 and 0.30, the number of items administered, and the correlation. The consistency of real CAT result was evaluated by examining consistency of pass or fail based on a cut score of 0.0. The efficiency of all CAT designs was assessed by comparing the average number of items administered under both stopping rules.

Results: Both SEM 0.25 and SEM 0.30 provided a good balance between accuracy and efficiency in CAT. The real data showed minimal differences in pass/fail outcomes between the 2 SEM conditions, with a high correlation ($r = 0.99$) between ability estimates. The simulation results confirmed these findings, indicating similar average item numbers between real and simulated data.

Conclusion: The findings suggest that both SEM 0.25 and 0.30 are effective termination criteria in the context of the Rasch model, balancing accuracy and efficiency in CAT.

Keywords: Computerized adaptive testing; Medical student; Republic of Korea; Psychometrics; Statistical model
Introduction

Background

Computerized adaptive testing (CAT) was first implemented in high-stakes exams such as the National Council Licensure Examination for Registered Nurses (NCLEX-RN) in the United States in 1994. Subsequently, the National Registry of Emergency Medical Technicians (NREMT) in the U.S. adopted CAT for the Emergency Medical Technicians licensing examination in 2007. Additionally, CAT has been adopted in other assessments, such as the Medical Council of Canada Qualifying Examination. The implementation of CAT is based on item response theory (IRT), which operates under the assumption that both the examinee ability parameters and item parameters remain invariant across different testing situations.

In the CAT process, after each interim ability estimate is calculated, the standard error of the current ability estimate is evaluated. CAT is stopped if this standard error falls below a predetermined criterion; otherwise, it continues until the standard error associated with an interim ability estimate meets the criterion. The final ability estimate is then determined by the most recent interim ability estimate once item selection has finished. CAT stopping rules are designed to ensure equivalent measurement precision for all examinees by terminating item administration once a predetermined measurement standard error is reached. These tests vary in length because the number of items required to meet the termination criterion can differ among examinees.

Researchers have introduced multiple variable-length stopping rules, all aimed at administering the minimum number of items necessary to achieve a reliable estimate of an examinee's ability, though they vary in the criteria used to determine when an examinee's ability has been adequately assessed. One such rule is the standard error of measurement (SEM) stopping rule, which ceases item administration when the standard error of the current ability estimate reaches a pre-specified level [1]. In a CAT simulation study, the efficacy of three variable-length stopping rules—standard error (SE), minimum information, and change in $\theta$—was
evaluated both individually and in combination with constraints on the minimum and maximum number of items [2].

Objectives

The purpose of this study was to evaluate the SEM-based termination criteria for CAT in medical examinations. The study aimed to assess the efficiency and accuracy of CAT using the Rasch model. Extensive research on CAT has employed SEM termination criteria to increase measurement efficiency and accuracy. This study, which used both simulation and real data, investigated the SEM termination criteria for implementing CAT in medical examinations.

Methods

Ethics statement: This study utilized students' responses in CAT and simulated data; therefore, neither approval by the institutional review board nor the obtainment of informed consent was required.

Study design: Both post-hoc simulation and real data studies were designed to determine the termination rule for computerized testing.

Data sources: Data were obtained from the responses of third-year medical students during examinations in medical courses, the topics of which included all clinical areas, in 2020 at Hallym University College of Medicine. Simulated data were generated using estimated parameters in R. The real data were obtained from real CAT examinations conducted using the adaptive testing platform, LIVECAT (The CAT Korea, https://www.thecatkorea.com/). The data included information calculated each time an examinee responded to individual items, such as the examinee's ability estimate and the number of items administered. CAT was implemented with a termination criterion of SEM 0.25, with data for each examinee being recorded from the first item until the SEM for the examinee's ability reached 0.25. To compare the examinee’s ability and the number of items administered based on different termination criteria, data from the original dataset were
extracted up to the point where the SEM for the examinee's ability reached 0.30. This approach enabled the creation of data for the same individuals under termination criteria of 0.25 and 0.30.

**Setting:** In a Monte Carlo simulation study, 1,012 real item parameters from an actual item bank were used, and the abilities (θ) of 1,000 students were generated from a distribution with a mean of 0 and a standard deviation of 1 to mimic real conditions. This study utilized 1,012 item parameters from real medical examination data at Hallym University College of Medicine, collected in 2016 and 2017. In the actual CAT, 83 students from Hallym University College of Medicine participated in a CAT examination using the LIVECAT platform in 2020 [3].

**Variables:** The outcome variables included the number of examinees who passed or failed according to the cut score with SEM values of 0.25 and 0.30 in the real CAT study, and the correlation and number of items administered with respect to the various SEM criteria in the post-hoc simulation study.

**Measurement:** This simulation study of CAT employed a post-hoc simulation design. In this approach, item responses were generated from 1,000 candidates responding to real items. A conventional test was previously administered to measure the candidates and to create a full data matrix that was later utilized in the simulated CAT. Because the true θ is unknown, a post-hoc simulation typically assesses the impact of varying CAT termination criteria on test efficiency. The CAT simulation was conducted using the “catR” packages in the R program [4, 5]. The real CAT study evaluated the consistency of pass or fail based on a cut score of 0, using SEM values of 0.25 and 0.30.

**Bias:** All participants were included in this study, and simulation data were generated using R packages. Therefore, there was no participant bias in this study.

**Study size:** Sample size estimation was not required for this study.

**Statistical methods:** Measurement results were presented using descriptive statistics. The SEM of examinees in CAT can be computed by observed standard error (OSE) of ability estimate.[6] SEM can be computed by
taking the inverse of the square root of the second derivative of the likelihood function when \( \theta \) is estimated by the maximum likelihood estimation (MLE) and expected a posteriori (EAP) methods. The SEM is described as:

\[
\sigma_\theta = \frac{1}{\sqrt{\left( \frac{\partial^2 \ln L(u | \theta_j)}{\partial \theta^2_j} \right)}}. \tag{1}
\]

where,

\[
\frac{\partial^2 \ln L(u | \theta_j)}{\partial \theta^2_j} = -\sum_{i=1}^{n} P_{ij} Q_{ij} \tag{2}
\]

and \( Q_{ij} = 1 - P_{ij} \). \( \ln L(u | \theta_j) \) is the log-likelihood function of examinee \( j \) directly from the Rasch model, following the assumption of local independence. The Rasch model is described as

\[
P_i(\theta_j) = \frac{\exp[(\theta_j - b_i)]}{1 + \exp[(\theta_j - b_i)]} \tag{3}
\]

The second derivative of the log-likelihood function containing observed values \( (u_i) \) can be called the SEM. The SEM value in Equation 1 was used to terminate CAT for individual examinees.

Equation 2 is equivalent to the test information function \( I(\hat{\theta}_j) \). Therefore, SEM can be expressed by the test information function, \( I(\hat{\theta}_j) \), as follows:

\[
\sigma_\theta = \frac{1}{\sqrt{I(\hat{\theta}_j)}}. \tag{4}
\]

The variable defined in Equation 4 can be called the theoretical standard error, which is distinct from the OSE.
Results

Real data results

The Hallym University medical exam was administered as variable-length CAT. In 2021, a total of 83 candidates were used in this analysis. In real data, the average number of items based on SEM=0.25 was 71.53 and the average number of items based on SEM=0.30 was 50.38. The correlation between the two abilities using SEM=0.25 and SEM=0.30 was 0.99 (P<0.001). The real data results were compared with the results of post-hoc simulation study.

The number of candidates who passed or failed based on SEM=0.25 and 0.3 is presented in Table 1. The cut-score was set to 0.0 based on the mean of the standard normal distribution. The classification consistency of pass or fail between the two SEM criteria was 0.927, indicating that there were no major differences between the two SEM criteria.

Table 1. Number (percent) of candidates who passed or failed (cut-score=0) based on standard error of measurement (SEM)=0.25 and 0.30

<table>
<thead>
<tr>
<th>Decision</th>
<th>SEM=0.25</th>
<th>SEM=0.30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pass</td>
<td>Fail</td>
</tr>
<tr>
<td>Pass</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Fail</td>
<td>4</td>
<td>47</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>49</td>
</tr>
</tbody>
</table>

Table 2. The number of items and proportion administered and exposed to examinees with respect to item difficulty ranges (SEM=0.25).

<table>
<thead>
<tr>
<th>Items administered</th>
<th>Items exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty ranges</td>
<td>Number</td>
</tr>
<tr>
<td>------------------</td>
<td>--------</td>
</tr>
<tr>
<td>$\theta &lt; -5$</td>
<td>6</td>
</tr>
<tr>
<td>$-5 \leq \theta &lt; -4$</td>
<td>28</td>
</tr>
<tr>
<td>$-4 \leq \theta &lt; -3$</td>
<td>48</td>
</tr>
<tr>
<td>$-3 \leq \theta &lt; -2$</td>
<td>61</td>
</tr>
<tr>
<td>$-2 \leq \theta &lt; -1$</td>
<td>160</td>
</tr>
<tr>
<td>$-1 \leq \theta &lt; 0$</td>
<td>252</td>
</tr>
<tr>
<td>$0 \leq \theta &lt; 1$</td>
<td>224</td>
</tr>
<tr>
<td>$1 \leq \theta &lt; 2$</td>
<td>131</td>
</tr>
<tr>
<td>$2 \leq \theta &lt; 3$</td>
<td>57</td>
</tr>
<tr>
<td>$3 \leq \theta &lt; 4$</td>
<td>24</td>
</tr>
<tr>
<td>$4 \leq \theta &lt; 5$</td>
<td>12</td>
</tr>
<tr>
<td>$5 &lt; \theta$</td>
<td>9</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>1012</strong></td>
</tr>
</tbody>
</table>

There were 1012 operational items in the item pool in CAT. Table 2 shows the number of items and proportion administered and exposed to examinees with respect to item difficulty ranges under $SEM=0.25$. Figure 1 shows number of items administered and exposed in CAT under $SEM=0.25$. 
The number of items administered and exposed in computerized adaptive testing (SEM=0.25)

Table 3. Correlation and number of items with respect to the SEM
<table>
<thead>
<tr>
<th>SEM</th>
<th>Correlation between $\theta$, $\hat{\theta}$</th>
<th>Mean number of items administered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>EAP</td>
</tr>
<tr>
<td>0.10</td>
<td>0.995</td>
<td>0.992</td>
</tr>
<tr>
<td>0.15</td>
<td>0.990</td>
<td>0.989</td>
</tr>
<tr>
<td>0.20</td>
<td>0.979</td>
<td>0.974</td>
</tr>
<tr>
<td>0.25</td>
<td>0.979</td>
<td>0.967</td>
</tr>
<tr>
<td>0.30</td>
<td>0.957</td>
<td>0.950</td>
</tr>
</tbody>
</table>

**Discussion**

**Interpretation**

This CAT study, employing both real and simulation data, investigated the impact of the SEM on the accuracy (correlation between true $\theta$ and CAT estimates) and efficiency (mean number of items administered in CAT) of CAT. The Rasch model was used as the underlying model for data generation because many assessment corporations (e.g., NCLEX-RN and NREMT) have implemented CAT using the Rasch model [7]. Two SEM criteria were employed to explore the accuracy and efficiency of CAT for both simulation and real data. The results using simulation data provided similar accuracy and efficiency to those of the real data, which can help generalize the findings from this study.

**Comparison with previous studies**

CAT for real data has a variable length for individual examinees. All examinees must respond to at least 60 test items, and the successive estimates are checked to determine whether the confidence interval (CI) around the estimate contains the passing score for the examination [8]. The CI stopping rule is commonly used in licensure settings to make classification of pass or fail decisions with fewer items in CAT. However, it tends to be less efficient in the near-cut regions of the ability scale, as the CI often fails to be narrow enough for an
early termination decision prior to reaching the maximum test length [9, 10]. Thus, combining SEM with a fixed-length component resulted in an efficient test where only examinees with relatively high or low abilities exist [2]. In a previous CAT simulation study, the efficacy of three variable-length stopping rules—standard error, minimum information, and change in \( \theta \)—was evaluated both individually and in combination with constraints on the minimum and maximum number of items. These rules were also compared to a fixed-length stopping rule. Each rule was assessed using two different termination criteria (SEM = 0.35 vs. SEM = 0.30) within the framework of a polytomous IRT model. The termination criterion of SEM = 0.30 performed better than SEM = 0.35 in terms of balancing accuracy with efficiency [2].

**Limitation**

Only the Rasch model was used in this study. Future research should apply other models, such as the 2-parameter logistic model (2PLM) and the 3-parameter logistic model (3PLM), to extend the findings.

**Generalizability**

Both SEM criteria values performed well; the 0.25 criterion yielded more precise estimates but slightly increased the test length. A researcher’s choice between a higher or lower SEM termination criterion should depend on whether efficiency or precision is prioritized. If reducing the testing burden is paramount, a higher criterion should be adopted to shorten the average test length. Conversely, a lower criterion should be used if optimal measurement precision and accuracy are more important. However, the differences between these criteria were minimal, and both provided an excellent balance between efficiency and measurement quality.

**Conclusion**

This study demonstrates that two termination criteria in CAT achieved a good balance between accuracy and efficiency. They performed effectively within the Rasch model, with support from both real data and post-hoc simulation results. These consistent findings across various data sources underscore the effectiveness of
an ideal termination criterion. However, future research should examine this criterion with other models and populations to confirm its broader applicability.

**ORCID**

Dong Gi Seo: [https://orcid.org/0000-0002-3049-5556](https://orcid.org/0000-0002-3049-5556)

Jeongwook Choi: [https://orcid.org/0000-0003-4611-8904](https://orcid.org/0000-0003-4611-8904)

Jinha Kim: [https://orcid.org/0009-0000-1440-9155](https://orcid.org/0009-0000-1440-9155)

**Authors’ contributions**

Conceptualization: DGS

Data curation: JWC, JHK

Methodology: JWC, DGS

Formal analysis/validation: DGS, JWC

Project administration: DGS

Funding acquisition: DGS

Writing-original draft: DGS

Writing-review & editing: JWC, JHK

**Conflict of interest**
Dong Gi Seo has been the CEO of CAT Korea since 2019 and Jeongwook Choi has worked at CAT Korea since 2020. Otherwise, no potential conflicts of interest relevant to this article were reported.

**Funding**

This study was supported by a research grant from Hallym University (HRF-202404-009).

**Data availability**

Dataset 1. Estimated item parameters for the computerized adaptive testing in R, which were used for generating simulated data.

**Acknowledgments**

None.

**Supplementary materials**

Supplement 1. A sample of the R code to generate the responses for computerized adaptive testing.

Supplement 2. Item response process for each person in computerized adaptive testing when the standard error of the estimate was set to 0.25, including 5,937 total responses for all students.

Supplement 3. Item response process for each person in computerized adaptive testing when the standard error of estimate was set to 0.3, including 4,182 total responses for all students.

Supplement 4. Audio recording of the abstract.

**References**


