

IRT-based classification analysis of an English language reading proficiency subtest

Language Testing
2022, Vol. 39(4) 541–566
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322211068847
journals.sagepub.com/home/ltj



Elif Kaya 

Bilkent University, Turkey

Stefan O’Grady 

Bilkent University, Turkey

Ilker Kalender 

Bilkent University, Turkey

Abstract

Language proficiency testing serves an important function of classifying examinees into different categories of ability. However, misclassification is to some extent inevitable and may have important consequences for stakeholders. Recent research suggests that classification efficacy may be enhanced substantially using computerized adaptive testing (CAT). Using real data simulations, the current study investigated the classification performance of CAT on the reading section of an English language proficiency test and made comparisons with the paper-based version of the same test. Classification analysis was carried out to estimate classification accuracy (CA) and classification consistency (CC) by applying different locations and numbers of cutoff points. The results showed that classification was suitable when a single cutoff score was used, particularly for high- and low-ability test takers. Classification performance declined significantly when multiple cutoff points were simultaneously employed. Content analysis also raised important questions about construct coverage in CAT. The results highlight the potential for CAT to serve classification purposes and outline avenues for further research.

Keywords

Classification accuracy, classification consistency, computerized adaptive testing, language proficiency, Rudner approach

Corresponding author:

Elif Kaya, English Language Preparatory Programme, Bilkent University, East Campus N Building, 06800 Ankara, Turkey.

Email: elifskaya@gmail.com

An important role of educational measurement is to inform the classification of individuals for various purposes related to decisions on, for example, professional licensure, admissions, or placement into educational courses of study. Classification decisions are often high stakes as “errors in classification may lead individuals to be deprived of opportunities such as well-deserved educational or career development” (Zhang, 2010, p. 120). Despite these high stakes, misclassification of individuals is common and may result from measurement errors associated with sampling, equating, the assignment of cut scores, standard setting methods, and standard setting committees (Arce-Ferrer et al., 2002; Linn, 2003; Stone et al., 2005). Classification of individuals who obtained assessment scores close to decision cutoff points is particularly critical because misclassification of such individuals is highly likely (Eckes, 2017). Given the significant impact classification decisions may have on individuals’ lives, it is incumbent upon test developers to provide evidence of classification accuracy (CA) and classification consistency (CC; AERA et al., 2014; Lathrop, 2015).

In second language assessment, language proficiency testing is probably the area in which classification has the largest impact. Increasingly, language proficiency tests are used to determine prospective students’ ability to follow English medium instruction (EMI) at the undergraduate level. Turkish higher education is no exception, and the growing number of EMI universities in this context has led to increased scrutiny of the development of institutional English language proficiency tests (Selvi et al., 2021). In this context, proficiency testing is conducted using locally developed, high-stakes, paper-based tests (PBTs). Dimova et al. (2020) have argued that locally developed language proficiency tests reflect institutional learning objectives, and may therefore be particularly well suited to fulfill important ancillary placement and diagnostic functions in language programs. However, researchers have commented on the high probability of misclassification involved in traditional PBTs and suggested that computer adaptive testing (CAT) may reduce misclassification by identifying the most informative items in an item bank to increase discrimination around cutoff points, and hence enhance the validity of test-based classification decisions (Curi & Silva, 2019; Mizumoto et al., 2019; Rudner & Guo, 2011; Zhang, 2010). The purpose of the current study is to investigate the potential application of CAT in this context by comparing the classification performance of CAT and paper-based testing (PBT) versions of an English language proficiency reading subtest developed and administered at a Turkish university. This analysis was undertaken using a data set containing real item responses to conduct a series of post-hoc simulations, which reflect authentic test taker behavior (Wang et al., 1999).

Literature review

PBTs and CATs

Measurement of language proficiency for university admissions is commonly based on linear, fixed-form, PBTs. PBTs are characterized by a fixed set of items that are administered to every examinee and are traditionally developed using Classical Test Theory to maximize internal consistency (Cronbach, 1990; Gulliksen, 1950; Weiss, 2004). A major limitation of PBTs is that it is designed to measure a range of ability around the mean of the “anticipated trait distribution” and may measure inadequately when applied to

examinees whose ability levels deviate substantially from that range (Weiss, 2004, p. 71). High reliability of ability estimates around the mean ability range may come at the expense of reduced measurement precision at the distribution tails. This represents a limitation in the current context because the ability range in the test taking population is often wide, and test results are intended to be used for multiple-level placement decisions.

CAT resolves PBT limitations through the use of item selection algorithms (Dunkel, 1999). Measurement imprecision at the distribution tails may also be observed in CAT, though computer adaptive tests (CATs) often have the advantage of being based on large item pools covering a large ability continuum and are designed to identify the most appropriate items for each examinee, which increases precision across the ability continuum. CATs are typically based on item response theory (IRT) models which relate examinee ability to different item parameters as a probabilistic function. A typical CAT continues until the algorithm is able to make an estimate of examinee proficiency based on a predetermined test termination rule (Wainer et al., 2000). Each test is individually adapted to the examinee's level of ability with the effect that CATs are commonly shorter than PBTs and involve a higher degree of measurement precision (Davey & Pitoniak, 2006; Wainer et al., 2000; Way et al., 2006). On average, CATs may require about 60% of the number of items needed in a conventional PBT (Wainer et al., 2000). CATs that draw upon well-developed item banks assess test taker populations that contain large ability differences in a way that may be hard to achieve using PBTs. However, CAT also presents distinct challenges to test developers, such as the requirement for a large IRT-based calibration sample and the expectation of invariance of item and ability parameters (Hambleton et al., 1991). In addition, CAT may introduce obstacles related to test security, staff training, computer availability, computer literacy, and logistics (Larson, 1987; Liu et al., 2019; Wainer et al., 2000; Wise & Kingsbury, 2000). In contrast, PBT is a familiar test format that is simple to implement and score. Both CAT and PBT are widely used testing formats that may be considered based on the test taking population, the educational setting, and resources.

Research into CAT in language assessment has typically focused on the development of the measurement instrument and comparison with a conventional version of the same test. Mizumoto et al. (2019) reported on the development of the CAT version of the Word Part Levels test and concluded that CAT decreased the time required to complete the assessment, and led to higher measurement accuracy. Increased measurement accuracy has also been reported for CATs of vocabulary size (Tseng, 2016), and reading and listening proficiency (He & Min, 2017). The potential for CAT to increase measurement accuracy in situations where the test is designed to assess reliably at different levels of ability—for example, for purposes of placement testing—is clear, but relatively unexplored in actual language testing programs.

Classification

CA and CC are two important concepts that define the precision of classifications. CA refers to “the extent to which the true classifications of examinees agree with the observed classifications” (Diao & Sireci, 2018, p. 20). CC is “the rate at which the classification decision will be the same on two identical and independent administrations of the test” (Lathrop, 2015, p. 1). Both CA and CC indices evaluate the classification performance of test takers by calculating the measurement error associated with ability estimates (Lathrop, 2015).

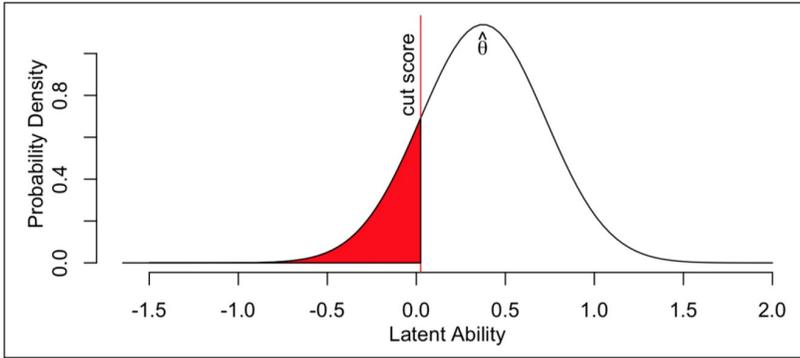


Figure 1. Illustration of CA and CC under Rudner's method.

Conditional probability distribution for a single examinee with ability estimate of 0.37 and total score of 28 from the Rudner approach. $\hat{\theta}$ denotes the examinee's latent ability estimate.

Reprinted from Q. N. Lathrop (2015, p. 3). Copyright 2020 by Quinn N. Lathrop.

To estimate CA and CC, there are several well-established methods rooted in IRT (for a review of these methods, see Diao & Sireci, 2018). The approach pioneered by Rudner (2001, 2005) is particularly appropriate for current purposes. In CAT, Rudner's classification method is recommended because different examinees are administered different versions of the same test and are assigned places on the ability continuum instead of total scores (Lathrop, 2015). For each examinee, a normal probability density as a function of ability is estimated using the item parameters (difficulty, item discrimination, and pseudo-guessing). Figure 1 (Lathrop, 2015, p. 3) illustrates this for a single examinee: the area above the cutoff scores (area not in red, or not shaded when viewing the figure in black and white) in the distribution represents the probability of being correctly classified as pass. The examinee's ability value exceeds the 0.0245 cut-score, and the examinee passes the test. The proportion of the unshaded (right) area squared (both pass) plus the proportion of the shaded or red (left) area squared (both fail) is the CC for this examinee. For two independent tests, the proportion of the area above the cutoff point squared (pass decision in both administrations) plus the proportion of the area below the cutoff point squared (pass decision in both administrations) gives the value of CC. Ability measures are calculated to estimate CA and CC values for each examinee. The individual CA and CC estimates are summed to obtain the CA and CC estimates for one test administration (Diao & Sireci, 2018).

CAT classification research has typically focused on the computerized adaptive classification test, which are CATs specifically designed for classifying individuals into different ability groups (Eggen & Straetmans, 2000; Gnamb & Batinic, 2011; Rudner & Guo, 2011). The effect of using multiple classification cutoff points in computerized adaptive classification tests has been investigated in several studies (Eggen, 2009; Spray, 1993). Cheng and Morgan (2012) examined classification performance using two to five cutoff points and found that classification performance was higher when fewer cutoff points were included. This result is expected because in the simplest scenario, classifying examinees into one of two decision categories (pass or fail) involves one cutoff point and

individuals may only be misclassified as either failing or passing. However, fine-grained categorization involves more cutoff points, which increases the potential sources of error. For example, with five cutoff points, misclassification may involve placing individuals into one of five categories. The literature suggests that complex classification is less accurate but a clearer picture of the relationship between the number of cutoff points and classification performance is needed.

There is a need for research to examine classification with multiple cutoff points on typical IRT-based CATs, because to date the main focus of the CAT literature has been estimating examinee ability as quickly and as precisely as possible rather than classifying examinees into different categories (Eggen, 2009; Wainer et al., 2000). For CATs to serve placement purposes, multiple classification categories must be available and the number of cutoff points should correspond to the number of ability levels targeted by the test.

The location of cutoff points on the ability continuum has an important impact on the classification performance of CAT. Classification performance is dependent on the cutoff location and increases when the cutoff point is located at the extremes of the ability continuum (Lee et al., 2002). Empirical research findings have demonstrated the important effect of the location of cutoff points on classification (Cheng & Morgan, 2012; Lathrop & Cheng, 2013), but these findings are based on generated data sets and have not been replicated with real data.

Whereas classification issues for linear PBTs have been extensively examined (Kim et al., 2006; Lee et al., 2002), examination of classification performance of CATs based on real data sets through post-hoc simulations is less common. This is an important focus because Monte Carlo studies, based on generated data sets, may not reflect test-taker guessing, speededness, and fatigue (Thompson & Weiss, 2011; Wang et al., 1999). By using real data involving examinee responses, CAT design can be enhanced to ensure high classification performance prior to the live CAT stage (Weiss, 2005).

The role of test termination in classification decisions

CATs finish once responses to a predetermined number of items have been provided or when a conventionally acceptable level of measurement precision (e.g., $\alpha = .70$ in Classical Test Theory) has been attained (Wainer et al., 2000). In the literature, fixed-length termination and standard error termination rules are commonly applied (Gushta, 2003). Fixed-length CATs terminate once a prespecified number of items has been administered. This termination rule has gained popularity in applied settings due to its simplicity and similarity to PBTs. However, a downside of implementing fixed-length termination, relating specifically to early stages of development when the item pool is limited, is variation in measurement precision between individuals. In addition, fixed-length termination may reduce efficiency by redundantly administering items that provide little new information about examinee ability. Alternatively, the test may be terminated before an estimate with the minimally acceptable degree of precision is obtained.

Standard error (SE)-based termination is another common termination rule, which stops the CAT once a predetermined standard error has been obtained (Boyd et al., 2010). An advantage of this rule is that examinees are tested with similar levels of measurement precision (Choi et al., 2010). Babcock and Weiss (2009) conducted a simulation study

and concluded that CATs terminated with an *SE* termination rule, as a variable-length method, discriminated highly between examinees and, contrary to claims made in the literature, did not perform any worse than fixed-length CATs (see also Chang & Ansley, 2003; Yi et al., 2001). The findings of the study suggested that CATs terminated with the *SE* termination rule resulted in equal measurement performance as the fixed-length versions given comparable average test lengths. However, it was argued that compared with fixed-length CATs, variable-length CATs operated with a standard error termination rule are more biased in estimates, suggesting a larger mean difference between the examinees' true ability scores and the estimated ability scores obtained by the CAT (Chang & Ansley, 2003; Yi et al., 2001). In the literature, many researchers use generated data to examine the classification performance of varying-length CATs and a probabilistic approach to make classifications (Spray & Reckase, 1994). However, in practice not every CAT is designed in a varying-length format. Examples of fixed-length approaches to classification include the Graduate Management Admission Test and ACCUPLACER (College Board, 2007). Overall, test termination represents a key consideration in CAT design that may influence the classification of examinees.

The present study

Despite the increasingly prevalent use of CATs to classify examinees, CA and CC have predominantly been discussed with reference to linear tests (Lee, 2010; Lee et al., 2002), and these indices are relatively unexplored in the CAT literature. In addition, CAT-based classification has not been investigated with data obtained from real examinees (Cheng & Morgan, 2012) and the effect of systematic manipulation to cutoff points on CA and CC in CAT is currently unclear.

CAT has potential to improve classification decisions in placement testing in the context of English medium university admissions. However, if CAT is to be considered a realistic alternative, it is important to compare the classification performance of different CAT formats with PBT because the findings of this comparison may provide guidance in terms of CAT design. Language programs face a decision about whether to use varying-length CATs or fixed-length CATs, and classification performance may differ substantially between the two approaches (Cheng & Morgan, 2012). Another decision language programs must make is whether to include single or multiple cut off points in the CAT. Binary decisions like mastery or nonmastery are common in testing (Eggen, 2009). However, classifying individuals into more than two groups for placement purposes is a commonly expected function of locally developed language tests (Dimova et al., 2020). The present study explores these gaps in the literature by investigating the classification performance of CAT using real data obtained from a locally developed and administered language test.

Research questions

It is clear that there is potential for the accuracy of university admissions and level placement decisions to be improved with the introduction of a CAT version of the English language proficiency test. However, at present questions remain regarding the optimal

test termination rule to apply in the CAT and the number of cutoff points the test can be expected to reliably support. In order to address these gaps, the following research questions were formulated:

1. To what extent do CA and CC values differ between PBT and CAT versions of an English reading subtest simulated with different test termination rules?
2. What is the effect of the location of the cutoff point on CA and CC values when a binary pass–fail decision is required?
3. How do CA and CC values vary when multiple cutoff points are used with respect to single cutoff scenarios?

Method

To answer Research Question 1, we calculated CA and CC values for different CAT scenarios (please see “Post-hoc Simulations” section) from the test data to compare classification performance between the CAT and PBT versions of the reading subtest. To answer Research Question 2, we identified relevant pass–fail cutoff points on the ability continuum, and we calculated CA and CC values to determine classification performance at these points. We conducted an analysis involving multiple cutoff points simultaneously to investigate the effect of adding more than one cutoff point on CA and CC to answer Research Question 3.

Participants

The study was conducted at a non-profit university in Ankara, Turkey, a research university with 12,000 students enrolled in different undergraduate and graduate programs. The study data were drawn from the students at the English language preparatory school within the university, in which 3000 students were enrolled. The English preparatory program provides courses at five different levels of ability: elementary, pre-intermediate, intermediate, upper-intermediate, and pre-faculty.

The number of examinees that took part in the current study was 1182, and their ages ranged from 17 to 21. The majority of the participants ($n=984$) were receiving English instruction in the program for 1 or 2 academic years and had finished the highest level of English language instruction offered in the university’s English language preparatory program. The remaining participants had external preparation status or amnesty student status, which indicates that they were not attending classes in the program but were eligible to sit the proficiency exam. Upon successful completion of the university English language proficiency test, students are able to start their graduate or undergraduate programs.

Instrument

The university English proficiency test is produced and administered in the English language preparatory school of the university. The test is designed to assess English proficiency at the B2 level on the Common European Framework of Reference for Languages

(Council of Europe, 2001). The B2 level is considered sufficient to follow EMI at the undergraduate level at the university (Kantarcioglu, 2012). Administered three times an academic year, the institution's proficiency test is taken by around 3000 students annually, in September, January, and June.

The data used in the study come from the PBT version of the university English proficiency reading subtest because examinee item responses were only made available for this particular subtest by the university. The reading subtest has three parts consisting of a total of six reading texts with 35 multiple-choice items that assess the students' ability to read for supporting details and specific information, follow textual coherence, make propositional and pragmatic inferences, and guess the meaning of vocabulary from context. An item bank of 35 items may be criticized as insufficient for CAT because the computer algorithm may be unable to identify appropriate items, resulting in lower CAT performance. However, this is an empirical study employing post-hoc simulations and was hence constrained by the research context; only items that were completed by real examinees were included. The results are expected to provide an insight into the potential administration of this test using CAT. Furthermore, CATs with a relatively smaller number of items in their item banks have been shown to function suitably. Sahin and Weiss (2015) have demonstrated that small item banks can still lead to accurate item parameter and ability estimates in CAT using simulated data with items having sufficient information. In another study conducted on a language test using real data, CATs with as few as 20 items were shown to produce promising results with *SE* values of 0.35 (Kaya & Kalender, 2018). These research findings indicate that CAT can certainly be administered with a very limited number of items. However, the corollary impact on construct coverage of this reduction is unclear and requires investigation.

Data set

We scored responses dichotomously (as correct or incorrect) using the answer key provided by the school's test development unit. The mean number of correct responses was 19.19 ($SD=6.51$, total available score was 35). The reliability of scores on the original paper-based reading subtest as shown by Cronbach's alpha was .84. Figure 2 shows the test information function and *SE* distribution of the PBT. As the figure shows, the reading subtest is more reliable between -1 and 0 on the ability continuum (this is the area associated with the lowest *SE* values). As an external validity criterion, we utilized students' scores on the listening subtest of the same exam (the examinees' listening data was calibrated by the university to estimate ability and item parameters). We calculated a Pearson's correlation coefficient between scores from the reading and listening subtests.

Study design

CAT is expected to use fewer items and estimate ability scores with higher reliability if it is to replace its PBT version. Thus, before examining classification performance, typical premises of the CAT (i.e., a reduction in the number of items administered and individual reliability estimates) were examined with different test termination rules. The first phase

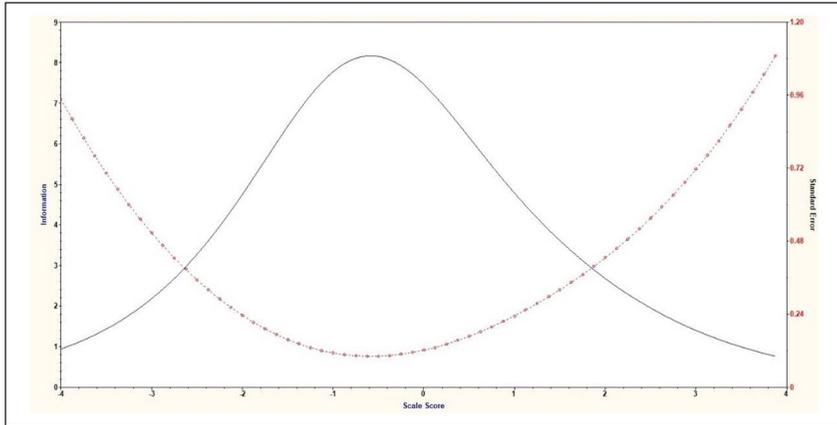


Figure 2. Item information and posterior standard error curves for PBT. Straight line shows item information and dotted line shows posterior standard deviation.

included post-hoc (real data) simulations, which is a common research strategy to investigate the feasibility of CAT in a given situation (Thompson & Weiss, 2011). In the literature, simulation studies are recommended to examine CAT design prior to a live CAT stage by investigating the optimum test termination rule and the extent to which test length can be reduced by readministering items adaptively (Weiss, 2005). Post-hoc simulations use real examinees' responses to a PBT to simulate testing behavior on a computer as if examinees were given a CAT; real responses reflect the psychometric characteristics of examinees better than generated data. In post-hoc simulations, items used in the PBT version constitute the item bank of the CAT. Prior to conducting simulations, the response data is used to estimate IRT parameters for the items in the test using software. In this study, BILOG-MG was used for IRT-based calibration (Zimowski et al., 1996). During the simulations, examinees' PBT responses were used to simulate a CAT. For this study, all CAT simulations were conducted using software developed by Kalender (2015).

The second stage involved classification analysis of simulation results. At this stage, different conditions based on two termination rules were tested (i.e., fixed-length termination and standard error termination; Gushta, 2003). For each examinee, final ability estimates and associated standard errors were calculated. To obtain a comparative analysis, the same quantities were also obtained for the real PBT.

Bayesian Expected A Posteriori (EAP) estimation was applied to generate ability estimates and item difficulty (Bock & Mislevy, 1982). An advantage of EAP is that EAP estimates ability with lower posterior standard deviation (Wang & Wang, 2001) and higher efficiency than the maximum likelihood estimation (MLE) method (Bock & Mislevy, 1982). Another advantage of EAP is that estimation of the latent trait proceeds when an examinee has only correct or incorrect responses (Desjardins & Bulut, 2018). An initial check of the data indicated the presence of all-incorrect response patterns ($n=45$, 3.8% of the data), and the MLE method fails to handle such response strings (Han, 2016). Because EAP can produce ability estimates with such response patterns, they were not removed from the data set.

Post-hoc simulations

Simulations were set to be terminated based on fixed-length and standard error termination rules. In this study, CATs terminated after 10, 15, 20, 25, and 30 items were simulated. Terminating a CAT after 10 or 15 items may raise questions about the construct coverage or the content validity of the test (Suvorov & Hegelheimer, 2013). However, these two test termination scenarios were also used in the study because it may prove informative for future research. Five *SE* rates were selected (i.e., below 0.5, 0.4, 0.3, 0.2, and 0.1) corresponding to different alpha values from .75 to .99 and labeled as SE05, SE04, SE03, SE02, and SE01 (Weiss, 2011). Thus, 10 different instances of CAT were simulated. Each simulation was replicated 100 times and the results were averaged. Fisher's maximum information was used to select items to minimize the *SE* associated with examinee ability estimation and maximize test information (Thissen & Mislevy, 2000; Veldkamp & Matteucci, 2013).

Classification

In this study, CA and CC values were estimated by the method proposed by Rudner (2001, 2005). The R package *cacIRT* developed by Lathrop (2015) was used to calculate CA and CC, both of which range from 0 to 1, with higher values indicating higher classification performance (e.g., a CA value of .76 indicates that there is a 76% chance that the individual has been appropriately classified and the same value for CC represents the probability of being classified into the same group across two administrations). For each examinee, a CAT was created using the responses they provided during the test, resulting in different combinations of items. To systematically examine the effect of different cut-off points on the ability continuum, nine cutoff values were set from the 10th to the 90th percentiles, increasing in increments of 10 and using percentile ranks both for CATs and PBT, which created 10 ability groups following a normal distribution. The cutoff points on the ability continuum corresponding to percentile scores of each decile were -1.35, -0.73, -0.51, -0.42, -0.17, 0.25, 0.51, 0.81, and 1.42. For classification analysis, the corresponding ability values at the same percentiles were used. An additional classification analysis was made using five cutoff points to investigate the potential for the CAT to classify individuals according to the six levels of the English preparatory program (see "Participants" section). Five cutoff points were set using 16.6, 33.3, 50, 66.6, and 83.3 as percentile ranks. These cutoff points were -0.92, -0.46, -0.12, 0.40, and 0.91. CA and CC values were calculated both for one cutoff score at a time as well as a simultaneous analysis of all cutoff scores. After classification analysis, the researchers completed a content analysis to examine the construct coverage of the CATs in the simulations.

Results

Preliminary analysis: Fit of model-data

Before answering the research questions, we present results of the IRT model-data-fit. A canonical factor analysis was carried out based on tetra-choric correlations using TESTFACT (Bock et al., 2003) to examine the factorial structure of the 35 items and

Table 1. Means (*SDs*) of the ability estimates by CATs averaged over 100 replications.

CAT	<i>M</i> (<i>SD</i>)	<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
FL10	-0.02 (0.88)	1.51	1181	.13	–
FL15	-0.02 (0.92)	2.44	1181	.01	0.06
FL20	-0.02 (0.94)	2.55	1181	.01	0.07
FL25	-0.02 (0.95)	3.20	1181	<.001	0.09
FL30	-0.02 (0.95)	5.08	1181	<.001	0.14
SE05	-0.02 (0.92)	1.95	1181	.05	–
SE04	-0.01 (0.94)	2.73	1181	<.001	0.07
SE03	-0.02 (0.95)	9.19	1181	<.001	0.22
SE02	-0.02 (0.95)	9.96	1181	<.001	0.30
SE01	-0.02 (0.95)	9.96	1181	<.001	0.30

CAT: computerized adaptive testing; *SD*: standard deviation; FL: fixed length; SE: standard error.

For each paired-samples *t*-test the mean and standard deviation of PBT was set to 0 and 1, respectively.

Alpha-level was set to .005 after Bonferroni correction.

establish the factorial structure of the trait measured. The results indicated that items measured a unidimensional latent trait as evidenced by a ratio of larger than 4 between the second eigenvalue to the first (5.6 and 1.3; Lord, 1980; Slocum-Gori, & Zumbo, 2011). The invariance of item parameters and ability estimates was also confirmed. Analysis of χ^2 values and plot of data fit suggested that the 2PL model showed the best fit for the data in hand. With data to model fit, ability estimates are more reliable and the accuracy and consistency of classifications are high (Lathrop & Cheng, 2013).

The mean (standard deviation) of item discrimination parameters was 0.90 (0.32), whereas item difficulty parameters had a mean of -0.25 (0.93). Minimum and maximum ability estimates were -2.57 and 2.81 ($M=0$, $SD=1$). *SE* of PBT had a mean (*SD*) of 0.34 (0.25) and 48.0% of the individuals had an *SE* of 0.4 or lower, whereas 90.4% had 0.5 or lower. Figure 2 presents the PBT item information curve.

Ability estimates by PBT and CAT

To establish the extent to which ability estimates vary according to the PBT, we calculated descriptive statistics. Table 1 shows averaged means and standard deviations of estimated ability values in the various CAT scenarios (PBT averaged means were set to 0.00) over 100 replications. The mean and standard deviation of PBT-based ability estimates were set to 0 and 1, respectively, for reference. The table denotes smaller standard deviations and negative means for all CATs, indicating that ability distributions with CATs are narrower than with the PBT, though not considerably. As Table 1 shows, the paired-sample *t*-test results indicated statistically significant mean differences after Bonferroni correction for 10 comparisons ($p (.05/10)=.005$) between all simulations and the PBT version. However, effect sizes of these differences were small with a minimum of 0.06 and a maximum of 0.30, as estimated by Cohen's (1988) *d*, indicating that ability levels estimated by CATs and PBT are similar.

Table 2. Descriptives of the number of items used in CATs with different *SE* thresholds.

CAT	<i>M</i> (<i>SD</i>)	Reduction in test length (%)	Minimum	Median	Maximum ^a
SE05	14.64 (7.27)	58.17	7	13.00	35 (5.25%)
SE04	26.63 (7.77)	23.91	14	27.00	35 (36.13%)
SE03	34.10 (2.30)	2.57	34	35.00	35 (85.44%)
SE02	35.00 (0.00)	0.00	35	35.00	35 (100.00%)
SE01	35.00 (0.00)	0.00	35	35.00	35 (100.00%)

CAT: computerized adaptive testing; *SE*: standard error.

^aThe values in parentheses show the percentage of examinees who were given all of the 35 items in the simulations.

Table 3. Descriptives of *SE*s in fixed-length (FL) CATs.

CAT	<i>M</i> (<i>SD</i>)	% of examinees with <i>SE</i> s below SE04	% of examinees with <i>SE</i> s below SE05
FL10	0.53 (0.06)	0	41.29
FL15	0.48 (0.06)	12.1	72.59
FL20	0.44 (0.06)	26.99	81.2
FL25	0.42 (0.07)	38.32	92.47
FL30	0.40 (0.07)	46.28	94.6

CAT: computerized adaptive testing; *SE*: standard error; FL: fixed length.

Test termination

Table 2 demonstrates that there was no reduction in the number of items using 0.2 and 0.1 *SE* thresholds. However, a relatively large reduction was achieved with 0.5 and 0.4 *SE* thresholds, which correspond to $\alpha \geq 0.75$ and 0.84, respectively (Cronbach, 1990). Only a small number of examinees were given all 35 items, 62 (5.25%), and 427 (36.13%) for CATs terminated with 0.5 and 0.4 *SE* thresholds, respectively. The number of examinees given the full item bank varies significantly across the two CAT scenarios terminating at different *SE* levels. Since the 0.4 *SE* level indicates a stricter reliability level associated with less error, more items were used to meet this level of precision.

Table 3 demonstrates that the mean *SE* values were around 0.40 for the CATs terminating with 20, 25 and 30 items. Increasing the number of test items caused *SE* estimations to decrease. Around 72%, 81%, and 92% of ability estimations were at or below 0.50 *SE* for the fixed-length CATs terminating with 15, 20, and 25 items, respectively. For the same fixed-length CATs, the percentages of ability estimations at or below 0.40 *SE* were around 12%, 27%, and 38%, respectively. Given that PBT produced a mean *SE* of 0.34, reliability of estimates by CAT were comparable.

As Table 4 demonstrates, with every termination criterion employed, CAT produced highly correlated and statistically significant Pearson correlations ($r = .93-1.00$, $p < .001$) with ability estimates obtained in the PBT. The table also shows the correlations between ability estimates from CAT and scores from the listening section (the external

Table 4. Pearson correlations between ability estimates by CAT and PBT.

CAT	r_{PBT}	df	r_{external}	df	CAT	r_{PBT}	df	r_{external}	df
FL10	.93	1180	.72	1180	SE05	1.00	1180	.72	1180
FL15	.96	1180	.74	1180	SE04	1.00	1180	.74	1180
FL20	.97	1180	.75	1180	SE03	.99	1180	.76	1180
FL25	.99	1180	.76	1180	SE02	.99	1180	.76	1180
FL30	.99	1180	.77	1180	SE01	.96	1180	.77	1180

CAT: computerized adaptive testing; FL: fixed length; SE: standard error.
All correlations were statistically significant at .001.

criterion) were higher than .70. In PBT, the correlation between reading and listening sections was .77 ($p < .001$).

Classification

To compare PBT and CAT classification and to explore the impact of adjusting cutoff points on CA and CC values in PBT and CAT, Figure 3(a)–(d) presents the CA and CC estimates with 9 cutoff points. Figure 3(a)–(d) shows CA and CC values for CATs with fixed-length and *SE*-thresholds. PBT values provide a reference point for comparison. The CA and CC values were obtained using only one cutoff score at a specific percentile. In other words, each dot in each line indicates a CA or CC value assuming there is only a specific cutoff score.

Figure 3(a)–(d) shows that both CA and CC estimates for all CATs with different termination rules were higher than 75%. In all CATs, CA estimates were slightly higher than CC estimates. CA and CC estimates showed an upward trend at the higher and lower tails of ability estimates. It seems that the higher the ability estimate, the higher the CA and CC estimates were for all CATs regardless of the test termination rule applied. The same trend can be observed for the lower ability groups. The high levels of classification at the tails was due to an absence of disproportionately difficult or easy items. CATs with fixed-items produced differing CA and CC values around the middle ability range (especially at the 50th and 60th percentiles), whereas *SE*-based CATs produced relatively more similar CA and CC estimates. At lower and higher ability levels, CAT CA and CC values were similar to the values in PBT; however, around the middle ability levels, the PBT recorded higher CA and CC values. CAT scenarios with lower numbers of items or higher *SE* values (FL10 and SE05) were associated with relatively lower CA and CC values around the middle ability ranges. The performance of CATs with 25 and 30 items at the average ability range were above 85% and the high-ability estimates were around 95%, which was very similar to the classification performance of the PBT version of the subtest. Similarly, CATs with SE04 had CA and CC values around 85% at the middle ability range and around 95% at the tails of the ability range.

With five cutoff points, a similar pattern in CA and CC values was observed around the middle ability range (see Figure 4(a)–(d)). To be specific, excluding FL10 and SE05, CATs produced CA estimates greater than 85% and CC values higher than 80% in the

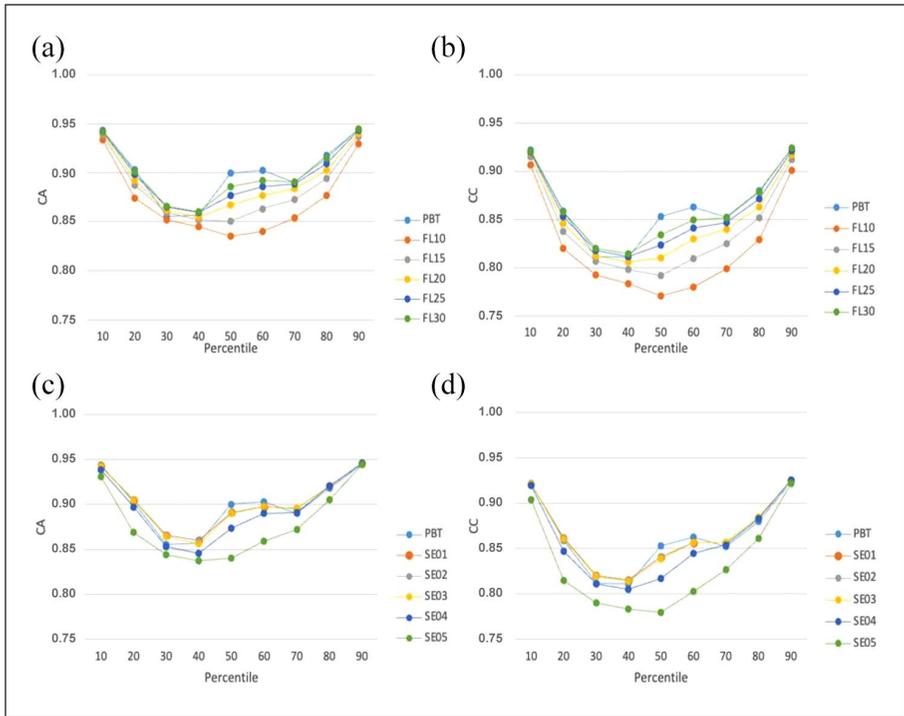


Figure 3. CA and CC values with 9 cutoff points: (a) CA at different percentiles for fixed-length CATs; (b) CC at different percentiles for fixed-length CATs; (c) CA at different percentiles for SE threshold CATs; and (d) CC at different percentiles for SE threshold CATs.

middle ability range. For fixed-length CATs, FL30 seemed to produce slightly higher CA and CC values and for CATs terminated after a prespecified SE threshold, CA and CC values slightly increased in SE01, SE02, and SE03 across all ability ranges. Another trend observed in Figure 4(a)–(d) is that like classification with nine cutoff points, CATs with fewer items or higher SE values showed relatively lower classification performance. This was expected because fewer items were available at the cutoff points, thus increasing misclassification rates. Similarly, applying high SE values resulted in CAT termination with fewer items, which may reduce the precision of classification. CATs with the highest CA and CC values were FL30 and SE01. The mean CA and CC values were around 90%. In contrast, FL10 and SE05 had a mean of 85% for CC and CA. However, unlike classification analysis with nine cutoff points, CA and CC values were different around the high and the low-ability levels. Generally, classification performance showed more variation in the tails. In all the CAT scenarios involving five cutoff points, CA values in the lower and upper percentiles fell within 85% and 95%. CC values, however, were slightly lower. CATs with 10 and 15 items or CATs with SE of 0.50 and 0.40 showed lower performance both in CA and CC. Both for nine and five cutoff points, CAT classification performed was equal to or slightly lower than the PBT version of the test.

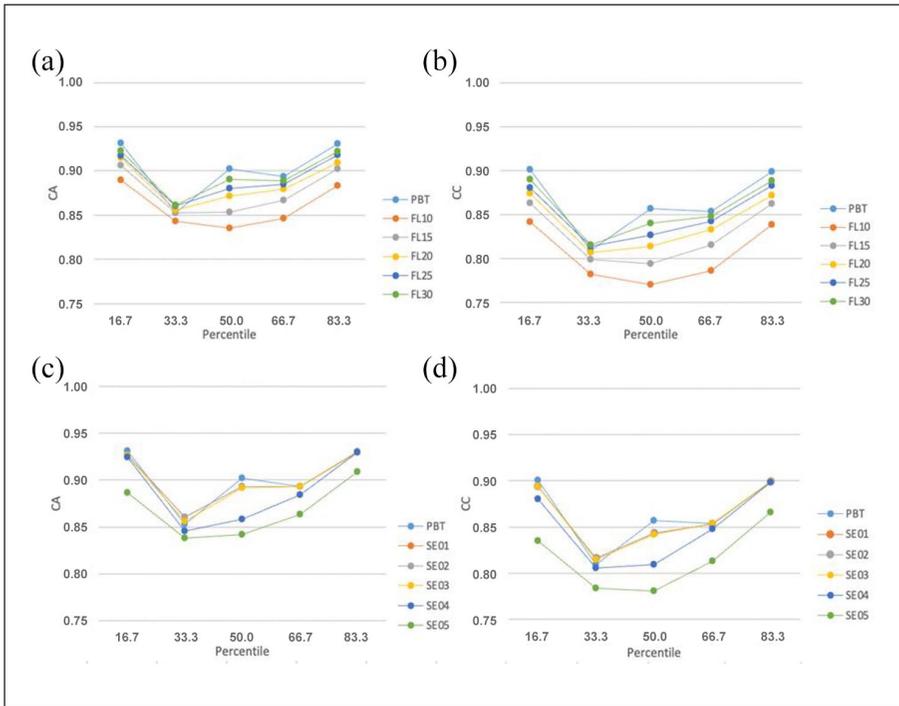


Figure 4. CA and CC values with 5 cutoff points: (a) CA at different percentiles for fixed-length CATs; (b) CC at different percentiles for fixed-length CATs; (c) CA at different percentiles for SE threshold CATs; and (d) CC at different percentiles for SE threshold CATs.

Figure 5 shows CA and CC values for each CAT scenario when nine and five cutoff points were applied simultaneously. CA and CC values significantly decreased compared with classifications with a single cutoff score. This is expected because increasing classification categories simultaneously increased the possibility of classification error. Classifications with a single cutoff score recorded CA and CC values ranging from 75% to 95%, whereas simultaneous classification produced CA and CC values between 25% and 60%. The results also showed that for all CAT scenarios, CA and CC values tended to rise significantly when the number of cutoff points decreased. As stated above, there was an inverse relationship between the number of decision categories and classification performance. Fewer items produced higher errors in ability, resulting in more misclassification. The classification with nine cutoff scores produced lower CA and CC values than with five cutoff scores. In general, different CAT scenarios produced similar CA and CC values for a given number of cutoff points. The figure also shows that with fixed-item CATs, higher numbers of items led to higher CA and CC values. This is due to the decreasing error rates in ability estimates associated with a greater number of items. With stricter ability estimates, more fine-grained classification became possible. With sufficient match between item difficulty and individual ability, this was true for

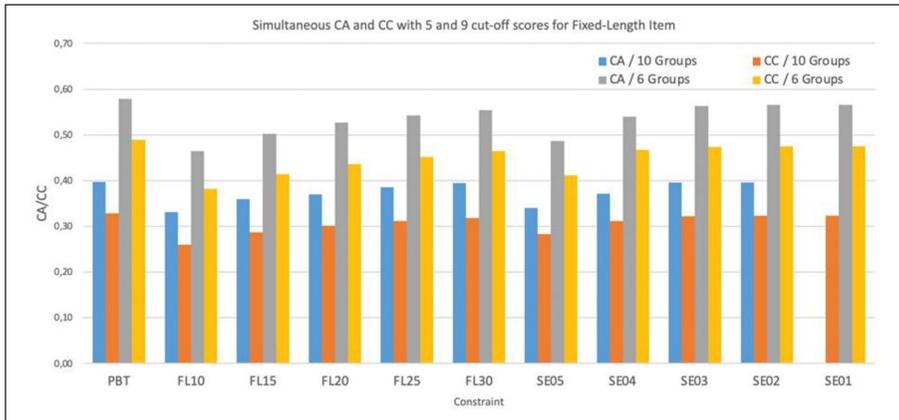


Figure 5. CA and CC values for simultaneous classifications with 5 and 9 cutoff points.

both CAT and PBT. In all cutoff point arrangements, PBT CA and CC values were higher than all CAT values although differences were minor. With the CATs terminating at SE01 and SE02, CA and CC values increased and were equal to the PBT.

Construct coverage

We investigated item distributions across the simulations, and the 25 most frequently used items were recorded (see Appendix 1). Items were placed into five categories (Chikalanga, 1992): propositional inference (14 items), paraphrase of supporting detail (14 items), pragmatic inference (4 items), meaning of unknown lexis in context (2 items), and textual cohesion (1 item) independently by two teachers in the English preparatory program with 100% agreement. As can be seen in Table 3, five of the simulations were involved in this analysis because there was no reduction in SE01, SE02, and the error of ability estimates were higher in FL10 (0.53), FL15 (0.48), and FL20 (0.44). The results show that the item categories were unequally represented across the simulations. For instance, the item assessing textual cohesion only featured in SE03. Although this finding may seem to indicate a significant reduction in construct coverage, it results from the disproportionate number of items associated with each category. That is, rather than evidence of systematic bias in item selection between the reading subdomains by the item selection algorithm, the unequal distribution was a result of the limited items identified as testing categories D and E.

To summarize the results in brief, with a single cutoff score, CAT-based classification performed very well but not for every ability level. As shown in Figures 3(a)–(d) and 4(a)–(d), PBT performs slightly better than CAT particularly around the middle ability group and a probable reason for this is that the PBT did not have the same termination restrictions as CAT. Classification performance dropped significantly when a simultaneous classification analysis was carried out with multiple cutoff points. Overall, CATs terminated with a fixed-length rule under 25 and 30 items (FL25 and FL30) and those

reaching a lower *SE* level (SE01 and SE02) showed better CA and CC performance. Finally, no significant pattern emerged in terms of the item focus distributed across various simulations.

Discussion

Results demonstrated that the PBT and CAT ability estimates were not substantially different. Correlations between the PBT and all CAT ability estimates were high (Pearson's $r = .93\text{--}1.00$). Overall, the simulation results clearly showed that a significantly lower number of items can be administered if the CAT version of the test is terminated at 0.50 and 0.40 *SE* threshold values or with 15, 20, and 25 items. Thus, premises of CATs (i.e., a reduction in the number of items administered, individual reliability estimates, and comparable results with PBT) were observed prior to classification analyses.

The first research question asked about the difference between CA and CC values on the PBT and CAT versions of the test with different test termination rules when one cut-off point was used at a time. As Figures 3(a)–(d) and 4(a)–(d) show, CA and CC values were lower than what would be expected in CAT regardless of the test termination method (i.e., CAT is expected to increase measurement precision over PBT; see Mizumoto et al., 2019 and Gyllstad et al., 2021 for a discussion of CAT expectations in language testing). CA values were around 80% while CC values decreased to 75% for the simulations. PBT-based CA and CC values were either higher or equal to the CA and CC values for both termination methods. Cheng and Morgan (2012) and Lathrop and Cheng (2013) reported higher levels of CA and CC values in CAT in their studies. However, these studies involved larger item banks and content balancing methods. Considering that no such method was employed in this study, these findings are still promising. The results indicate that there is potential to introduce CAT in this context with a larger item pool. As the analyses showed, not all scenarios provided equal CA and CC values. Classification was better for CAT scenarios with FL30 and SE01 around the middle ability range, whereas CATs with FL10 and SE05 provided lower classification performance by almost 10% at the same range. To answer Research Question 1, the PBT and CAT versions of this particular test proved similar in terms of classification performance.

Research Question 2 asked about the effect of the location of the cutoff point on CA and CC values. To answer this question, when only one cut-score was used for a binary decision such as pass or fail, classification performance of the test was 77% for all the ability groups using both test termination rules. For classifications with a single cutoff point, the graphs consisted of lines with a U-shape curve, suggesting that the classification performance for low and high-ability groups was higher than for mid-ability groups owing to a lack of disproportionately difficult or easy items. As stated earlier, at the tails, in every fixed-item and *SE*-based CAT, CA and CC estimations were above 75%, and reached around 90%. A U-shaped trend was observed in each analysis, meaning that CAT achieved better CA and CC for low and high-ability groups than for the middle-ability groups. In this sense, the results of this study are in parallel with those reported in the literature. Lathrop and Cheng (2013) found a similar U-shaped trend for different conditions (i.e., different test lengths, IRT models, and cut score locations) and CA performance significantly dropped around the center of the ability distribution.

In answer to Research Question 3, which asked about the effect of operating multiple cutoff points at the same time in all CAT scenarios, CATs with five cutoff points were better able to classify examinees than those with nine cutoff points. After reducing the number of cutoff points from nine to five, CA and CC values increased even at the tails of the ability range. The reason for this may be the location of cutoff points along the ability continuum. When the number of points decreased from nine to five, the locations and the distances between them also changed, which created more heterogeneous groups. This indicates that a lower number of cutoff points created groups with less homogeneity in the latent trait. Increasing the cutoff points and decision categories naturally yields a higher amount of classification error as there are multiple categorizations and comparisons made (Cheng & Morgan, 2012). The CATs with five cutoff points consistently performed almost 10% better than all the other CATs (see Figure 4(a)–(d)). CA estimations with five cutoff and nine cutoff points were around 40% and 60%, respectively. Cheng and Morgan (2012) reported similar findings. In their study, the number of decision categories was one of the factors that significantly affected classification performance of CAT using simulated item responses. The present study confirms their findings using real responses to an authentic test.

These findings indicate that caution should be taken when multiple cutoff points are used simultaneously because CA and CC decrease when more decision categories are used. Considering that the number of examinees in high and low-ability groups is generally much lower than the number falling into the middle ability group, as in a typical normal distribution, classification performance at the middle level actually carries more importance because it affects more examinees. Even when a binary decision (pass or fail) is going to be made about the examinees and the cutoff point is located around the middle ability group, the sensitivity of the score becomes more of an issue for all the stakeholders. Therefore, classifying individuals around the middle-ability range requires careful attention. Mismatch between the mean difficulty level of items and the ability levels of examinees should be considered while designing CATs for classification purposes. If item difficulty and ability levels are not aligned, examinees may exhibit aberrant response behavior such as blind guessing. This may, in turn, create problems with estimating ability levels with sufficient precision and classifying individuals correctly. This problem can be addressed by including items that provide more information around the cutoff points at the middle ability range in the item bank and ensuring a better match between item parameters and examinee ability (Berger et al., 2019).

The results of this study have important implications in the field of educational measurement. While most research concerning classification in computerized tests aims to classify examinees as “pass” or “fail,” the method and the findings of this study are relevant to situations where more than one cutoff score is used (Eggen & Straetmans, 2000; Spray, 1993; Weissman, 2004). The key finding of the study is that CAT classification functions differently relative to the number of decision categories and cutoff locations. The results clearly illustrate that CA and CC are higher when the number of decisions to be made on the basis of test results is limited, which supports the conclusions drawn by Cheng and Morgan (2012) that CA and CC drop if classification decisions are made based on more complex categorization. For this reason, it is advised that the only placement function this test is used for involves decisions located around the population tails

(i.e., to place at-risk test takers into elementary support levels and proficient test takers into advanced levels).

This study has implications specific to language testing. The absence of any substantial difference between test termination rules gives more freedom to language test developers while designing CATs. Instead of relying on fixed-length CATs, *SE*-based CATs may also be used for language assessment. This is critical information if CAT is introduced into the current context. There are social and even legal ramifications of test termination decisions, especially in Turkey where the public demand transparency in matters of educational assessment and decisions based on language tests are commonly contested in courtrooms. The finding that, statistically at least, the CAT termination method has little bearing on the classification of individuals has important implications for public debates surrounding assessment and is information that should be communicated to various stakeholder groups (Chalhoub-Deville & O'Sullivan, 2020). Considering the high-stakes involved in language proficiency exams at university preparatory schools and criticism that may come from students and families of CAT implementation (different items, different number of items, etc.), the results provide an opportunity for test developers to be flexible in their approach toward test termination.

The content analysis revealed that a variety of reading abilities was measured regardless of the test termination rule, although certain abilities were better represented than others. This implies that CAT would benefit from content balancing mechanisms to ensure the items from the different reading subdomains are administered. Using CAT in a language comprehension test featuring items that target different reading processes may result in a situation whereby the various item types are unevenly represented between the test takers. This level of variation is important; if two statistically equivalent tests vary substantially in terms of content, will stakeholders accept that the test assesses the same reading construct? As the literature review makes clear, CAT classification research primarily focuses on the potential to increase measurement precision. The impact of applying different termination rules on construct coverage has been neglected. Reading assessments include a range of items designed to test different cognitive operations (Khalifa & Weir, 2009). In tests of academic reading, the construct involves lower level processes such as scan and skim mechanisms and more complicated cognitive operations such as inference (Bax, 2013). Adequate construct coverage is crucial for test results to inform valid decisions about test candidates' academic reading ability (Weir et al., 2009). However, algorithmic item selection mechanisms maximize measurement precision potentially to the detriment of the test construct because certain cognitive operations may be underrepresented in the CAT version of the reading test. This represents a major drawback of CAT classification that should be underscored for language testers.

Future research may explore construct coverage in CAT to investigate bias in item selection algorithms. It may be the case that items representing particular subdomains, such as textual cohesion in reading, are not selected as frequently as others and this would have important ramifications for construct representation. In addition, a testlet format in which items based on a common stimulus are administered collectively to create a single partial credit item per reading passage may have been appropriate for this CAT. This multistage testing approach was considered but not administered due to constraints relating to the amount of obtainable data that would have decreased the number

of reading passages and items available in the simulation and hence restricted the analysis. For this reason, each item was considered independent rather than a member of a fixed set of items based on a common reading passage. Statistical evidence indicated that the 35 items were independent, however, items that share a common stimulus (e.g., the same reading passage) often exhibit local dependence because miscomprehension of the common stimulus is likely to impact upon responses to the interrelated items (Eckes, 2014). The testlet approach presents a solution to this problem but only if the number of items available to create the CAT is sufficient. Further studies are therefore recommended to examine the classification performance of CAT featuring multiple reading passages using the testlet approach.

It is important to acknowledge the limitations of the study. The first limitation relates to the size of the item bank. Using a limited number of items yields less test information and consequently, decreases classification performance. In addition, when interpreting the results of the study, it is important to recall that the item bank was not designed for typical CAT or computer adaptive classification tests and hence does not cater for a large range of ability. The study focused only on the B2 level of Common European Level of Reference for Languages, and this may limit the generalizability of the study. The test would be unable to make detailed distinctions between students that have substantially higher or lower levels of English.

Conclusion

We developed this study to determine the impact of introducing a non-classification CAT on the CA and CC of decisions relating to students' English language ability. To this end, the accuracy and consistency of classification decisions were investigated with single and multiple cutoff points in different CAT simulations. Overall, given that the test and item bank were designed for PBT with full fixed item administration, CAT yielded promising levels of CA and CC when a binary pass–fail decision was required. This was the case regardless of the test termination rule. CAT also performed well with simultaneous use of multiple cutoff points. The results of this preliminary study demonstrate that non-classification CAT has high potential for classification purposes in pre-sessional English language programs. It is hoped that the current study contributes to debates surrounding CAT in language testing and outlines a method of investigating the introduction of CAT in local contexts.

The findings obtained from the present study demonstrate the potentialities of using regular CATs to classify students in an educational program. This preliminary study shows that a regular CAT has the potential to give a similar level of classification performance to PBT with fewer items. We believe this is an important finding considering the limitations of administering PBT relating to time and the resources spent on such tests. With an initial investment into the item bank and the necessary facilities, using a regular CAT to classify learners can yield a similar classification performance to PBT. It also opens the door for more frequent assessment and early detection of misclassification, which will have a positive washback effect on teaching and learning practices in a language program.

Acknowledgements

We would like to thank the anonymous reviewers and the editors of the journal for their careful reading of our manuscript and for providing us with constructive feedback and insightful comments on the previous versions of it.

Declaration of conflicting interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Elif Kaya  <https://orcid.org/0000-0002-0482-1338>

Stefan O'Grady  <https://orcid.org/0000-0003-3810-713X>

İlker Kalender  <https://orcid.org/0000-0003-1282-4149>

References

- AERA, APA, & NCM. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Arce-Ferrer, A., Frisbie, D. A., & Kolen, M. J. (2002). Standard errors of proportions used in reporting changes in school performance with achievement levels. *Educational Assessment, 8*(1), 59–75. https://doi.org/10.1207/S15326977EA0801_04
- Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing* (pp. 1–21). <http://iacat.org/sites/default/files/biblio/cat-09babcock.pdf>
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441–465. <https://doi.org/10.1177/0265532212473244>
- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education, 4*(1). <https://doi.org/10.3389/feduc.2019.00001>
- Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TEST-FACT 4 user's guide*. Scientific Software International.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. <https://doi.org/10.1177/014662168200600405>
- Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229–255). Routledge.
- Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity theoretical developments and integrated arguments*. Equinox.
- Chang, S., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*(1), 71–103. <https://doi.org/10.1111/j.1745-3984.2003.tb01097.x>

- Cheng, Y., & Morgan, D. (2012). Classification accuracy and consistency of computerized adaptive testing. *Behavioral Research Methods*, 45(1), 132–142. <https://doi.org/10.3758/s13428-012-0237-6>
- Chikalanga, I. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in a Foreign Language*, 8(2), 697–709. <https://nflrc.hawaii.edu/rfl/item/506>
- Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, 19(1), 125–136. <https://doi.org/10.1007/s11136-009-9560-5>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- College Board. (2007). *ACCUPLACER online technical manual*. The College Board.
- Council of Europe. (2001). *Common European framework of reference for language learning and teaching*. Cambridge University Press.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Harper & Row.
- Curi, M., & Silva, V. (2019). Academic English proficiency assessment using a computerized adaptive test. *TEMA (São Carlos)*, 20(2), 381–401. <https://doi.org/10.5540/tema.2019.020.02.0381>
- Davey, T., & Pitoniak, M. J. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 543–573). Lawrence Erlbaum.
- Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press. <https://doi.org/10.1201/b20498>
- Diao, H., & Sireci, S. G. (2018). Item response theory-based methods for estimating classification accuracy and consistency. *Journal of Applied Testing Technology*, 19(1), 20–25. <http://www.jattjournal.com/index.php/atp/article/view/131016>
- Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing*. Routledge. <https://doi.org/10.4324/9780429492242>
- Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning and Technology*, 2(2), 77–93. <https://doi.org/10.125/25044>
- Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, 31(1), 39–61. <https://doi.org/10.1177/0265532213492969>
- Eckes, T. (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing*, 34(3), 383–411. <https://doi.org/10.1177/0265532216672703>
- Eggen, T. J. H. (2009). Three-category adaptive classification testing. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing. Statistics for social and behavioral sciences* (pp. 373–387). Springer. https://doi.org/10.1007/978-0-387-85461-8_19
- Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 60(5), 713–734. <https://doi.org/10.1177/00131640021970862>
- Gnamb, T., & Batinic, B. (2011). Polytomous adaptive classification testing: Effects of item pool size, test termination criterion, and number of cut-off points. *Educational and Psychological Measurement*, 71(6), 1006–1022. <https://doi.org/10.1177/0013164410393956>
- Gulliksen, H. (1950). *Theory of mental tests*. John Wiley. <https://doi.org/10.1037/13240-000>
- Gushta, M. M. (2003, May). *Standard-setting issues in computerized-adaptive testing* [Paper presentation]. Annual Conference of the Canadian Society for Studies in Education, Halifax, NS, Canada. <http://www.iacat.org/content/standard-setting-issues-computerized-adaptive-testing>

- Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, 38(4), 558–579. <https://doi.org/10.1177/0265532220979562>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.
- Han, K. C. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement*, 40(4), 289–301. <https://doi.org/10.1177/0146621616663137>
- He, L., & Min, S. (2017). Development and validation of a computer adaptive EFL test. *Language Assessment Quarterly*, 14(2), 160–176. <https://doi.org/10.1080/15434303.2016.1162793>
- Kalender, İ. (2015). Simulate_CAT: A computer program for post-hoc simulation for computerized adaptive testing. *Journal of Measurement and Evaluation in Education and Psychology*, 6(1), 173–176. <https://doi.org/10.21031/epod.15905>
- Kantarcioglu, E. (2012). *Relating an Institutional Proficiency Examination to the CEFR: A case study* [Unpublished doctoral dissertation]. University of Roehampton. https://pure.roehampton.ac.uk/ws/portalfiles/portal/443838/Elif_Kantarcioglu_PhD_2012.pdf
- Kaya, E., & Kalender, İ. (2018). Use of computerized adaptive testing in class-level language assessment. In T. Akşit, H. Mengü, & R. Turner (Eds.), *Classroom assessment: Bridging teaching, learning and assessment* (pp. 74–81). Cambridge Scholars.
- Khalifa, H., & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.
- Kim, D., Choi, S. W., Um, K. R., & Kim, J. (2006, April). *A comparison of methods for estimating classification consistency* [Paper presentation]. Annual Meeting of the National Council of Measurement in Education, San Francisco, CA, United States.
- Larson, J. W. (1987). Computer-assisted language testing: Is it portable? *ADFL Bulletin*, 18(2), 20–24.
- Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package caIRT. *Practical Assessment, Research & Evaluation*, 20(18), 1–5. <https://doi.org/10.7275/43vm-p442>
- Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, 37(3), 226–241. <https://doi.org/10.1177/0146621612471888>
- Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, 47(1), 1–17. <https://doi.org/10.1111/j.1745-3984.2009.00096.x>
- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26(4), 412–432. <https://doi.org/10.1177/014662102237797>
- Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, 11(31), 1–20. <https://doi.org/10.14507/epaa.v11n31.2003>
- Liu, C., Han, K. T., & Li, J. (2019). Compromised item detection for computerized adaptive testing. *Frontiers in Psychology*, 10(829). <https://doi.org/10.3389/fpsyg.2019.00829>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, 36(1), 101–123. <https://doi.org/10.1177/0265532217725776>

- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research and Evaluation*, 7(14), 1–5. <https://doi.org/10.7275/an9m-2035>
- Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research and Evaluation*, 10(13), 1–4. <https://doi.org/10.7275/56a5-6b14>
- Rudner, L. M., & Guo, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *Journal of Applied Testing Technology*, 12.
- Sahin, A., & Weiss, D. J. (2015). The effects of calibration sample and item bank size on ability estimation in Computerized Adaptive Testing. *Educational Sciences: Theory & Practice*, 15(6), 1585–1595. <https://doi.org/10.12738/estp.2015.6.0102>
- Selvi, A. F., Saracoğlu, E., & Çalışkan, E. (2021). When inclusivity means playing safe: Ideological discourses and representations in English testing materials. *RELC Journal*. Advance online publication. <https://doi.org/10.1177/00336882211008662>
- Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, 102(3), 443–461. <https://doi.org/10.1007/s11205-010-9682-8>
- Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Research report 93-7). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR93-07.pdf
- Spray, J. A., & Reckase, M. D. (1994, April). *The selection of test items for decision making with a computer adaptive test* [Paper presentation]. National Meeting of the National Council on Measurement in Education, New Orleans, LA, United States.
- Stone, C. A., Weissman, A., & Lane, S. (2005). Consistency of student proficiency classifications under competing IRT models for a state assessment program. *Educational Assessment*, 10(2), 125–146. https://doi.org/10.1207/s15326977ea1002_3
- Suvorov, R., & Hegelheimer, V. (2013). Computer-assisted language testing. In A. J. Kunnan (Ed.), *Companion to language assessment* (pp. 593–613). Wiley-Blackwell. <https://doi.org/10.1002/9781118411360.wbcla083>
- Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–133). Lawrence Erlbaum.
- Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, 16(1), 1–9. <https://doi.org/10.7275/wqzt-9427>
- Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, 97, 69–85. <https://doi.org/10.1016/j.compedu.2016.02.018>
- Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio*, 78(21), 57–82. <https://doi.org/10.1590/S0104-40362013005000001>
- Wainer, H., Dorans, N. J., Eignor, D., Flaughner, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum.
- Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, 25(4), 317–331. <https://doi.org/10.1177/01466210122032163>
- Wang, X. B., Pan, W., & Harris, V. (1999). *Computerized adaptive testing simulations using real test taker responses* (LSAC computerized testing report 96-06). Law School Admission Council. <https://files.eric.ed.gov/fulltext/ED467808.pdf>
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, United States. https://images.pearsonassessments.com/images/tmrs/Score_Comparability_of_Online_and_Paper_Administrations_of_TAKS_03_26_06_final.pdf

- Weir, C., Hawkey, R., Green, T., & Devi, S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS*. *British Council/IDP Australia IELTS Research Reports*, 9(4), 157–189. https://www.ielts.org/-/media/research-reports/ielts_rr_volume09_report4.ashx
- Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counselling and education. *Measurement and Evaluation in Counselling and Development*, 37(2), 70–84. <https://doi.org/10.1080/07481756.2004.11909751>
- Weiss, D. J. (2005). *Manual for POSTSIM: Post hoc simulation of computerized adaptive testing (Version 2.0) [Computer software]*. Assessment Systems Corporation.
- Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2(1), 1–27. <https://doi.org/10.2458/v2i1.12351>
- Weissman, A. (2004, April). *Mutual information item selection in multiple-category classification CAT* [Paper presentation]. Annual Meeting of the National Council for Measurement in Education, San Diego, CA, United States. <https://doi.org/10.1177/0013164406288164>
- Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, 21(1), 135–155. <https://www.uv.es/psicologica/articulos1y2.00/wise.pdf>
- 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. <https://doi.org/10.1111/j.1745-3984.2001.tb01127.x>
- Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, 27(1), 119–140. <https://doi.org/10.1177/0265532209347363>
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Scientific Software International.

Appendix I

Details of the twenty-five most frequently used items in CAT simulations

Table 5. Categories of the 25 most frequently used items in CAT simulations.

SE03		SE04		Test Termination Rule		SE05		FL30		FL25		FL20		
Item #	Domain	Count	Item #	Domain	Count	Item #	Domain	Count	Item #	Domain	Count	Item #	Domain	Count
33	A	1182	33	A	1182	29	B	1155	33	A	1182	33	A	1182
35	A	1182	35	A	1182	14	C	1097	35	A	1182	35	A	1182
12	A	1182	28	A	1182	28	A	1079	12	A	1182	12	A	1182
28	A	1182	11	B	1182	11	B	923	18	A	1182	4	B	1182
18	A	1182	14	C	1182	27	B	795	20	A	1182	23	B	1182
16	A	1182	27	B	1178	33	A	759	34	A	1182	15	B	1180
19	A	1182	29	B	1177	18	A	736	4	B	1182	21	B	1178
17	A	1182	12	A	1115	19	A	718	23	B	1182	29	B	1174
20	A	1182	17	A	1108	16	A	685	21	B	1182	9	C	1171
29	B	1182	23	B	1084	23	B	679	11	B	1182	14	C	1160
4	B	1182	20	A	1073	21	B	606	15	B	1182	20	A	1126
23	B	1182	3	C	1064	35	A	588	14	C	1182	6	D	1122
21	B	1182	19	A	1060	30	B	570	9	C	1182	18	A	1081
11	B	1182	4	B	991	9	C	523	7	C	1182	7	C	1070
15	B	1182	18	A	966	4	B	444	6	D	1182	28	A	1052
27	B	1182	15	B	964	32	B	442	29	B	1179	11	B	1002
30	B	1182	21	B	955	5	A	422	28	A	1168	16	A	960
13	B	1182	13	B	887	3	C	416	24	B	1156	27	B	889
14	C	1182	16	A	871	1	A	403	17	A	1150	30	B	843
9	C	1182	9	C	836	13	B	398	3	C	1109	3	C	778
3	C	1182	7	C	817	17	A	387	16	A	1105	19	A	768
7	C	1182	30	B	783	25	D	385	10	B	1099	17	A	742
6	D	1182	2	B	762	2	B	363	32	B	1079	34	A	682
31	E	1182	6	D	743	22	A	350	30	B	1004	32	B	660
34	A	1179	34	A	729	24	B	305	27	B	988	13	B	614
														488

SE: Standard error; FL: Fixed-Length; A: propositional inference; B: paraphrase of supporting detail; C: pragmatic inference; D: meaning of unknown lexis in context; E: textual cohesion. Count refers to the number of times that the item features in a test.