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

Elif Kaya, Stefan O'Grady and İlker Kalendar

IRT-based Classification Analysis of an English Language Reading Proficiency Subtest

An important role educational measurements are expected to fulfil is to inform the classification of individuals for various purposes related to professional licensure and admissions to 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 that record scores on assessments close to decision cut-off points is particularly critical because misclassification 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 and consistency (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. Dimova et al. (2020) have recently 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 paper-based testing (PBT) and suggested that computer adaptive testing (CAT) may reduce misclassification by identifying the most informative items in an item bank to increase discrimination around cut-off 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 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 behaviour (Wang et al., 1999).

Literature Review

Paper-based and Computerized Adaptive Tests

Measurement of language proficiency for university admissions is commonly based on linear, fixed-form paper-based tests. PBT is 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 PBT 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 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) report on the development of the CAT version of the Word Part Levels test and conclude that CAT decreases the time required to complete the assessment and leads 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

Classification Accuracy (CA) and Classification Consistency (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).

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 cut-off scores (not in red) 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 red (left) area squared (both fail) is the CC for this examinee. For two independent tests, the proportion of the area above the cut-off point squared (pass decision in both administrations) plus the proportion of the area below the cut-off 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 up to obtain the CA and CC estimates for one test administration (Diao & Sireci, 2018).

INSERT FIGURE 1 HERE

CAT classification research has typically focused on the computerized adaptive classification test (CACT), which are CATs specifically designed for classifying individuals into different ability groups (Eggen & Straetmans, 2000; Gnambs & Batinic, 2011; Rudner & Guo, 2011). The effect of using multiple classification cut-off points in CACTs has been investigated in several studies (Eggen, 2009; Spray, 1993). Cheng and Morgan (2012) examined classification performance using two to five cut-off points and found that classification performance was higher when fewer cut-off 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 cut-off point and individuals may only be misclassified as either failing or passing. However, fine-grained categorization involves more cut-off points, which increases the potential sources of error. For example, with five cut-off 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 cut-off points and classification performance is needed.

There is a need for research to examine classification with multiple cut-off 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 cut-off points should correspond to the number of ability levels targeted by the test.

The location of cut-off points on the ability continuum has an important impact on the classification performance of CAT. Classification performance is dependent on the cut-off location and increases when the cut-off 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 cut-off 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 (Lee et al., 2002; Kim et al., 2006), 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 pre-specified 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-based termination (SE) 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 a 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 suggest 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 is argued that compared to 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 studies 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 cut-off 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 cut-off 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 cut-off point on CA and CC values when a binary pass-fail decision is required?
3. How do CA and CC values vary when multiple cut-off points are used with respect to single cut-off scenarios?

Method

To answer research question one, CA and CC values were calculated for different CAT scenarios (please see Post-hoc Simulations) from the test data to compare classification performance between the CAT and PBT versions of the reading subtest. To answer research

question two, relevant pass-fail cut-off points were identified on the ability continuum and CA and CC values were calculated to determine classification performance at these points. An analysis involving multiple cut-off points simultaneously was conducted to investigate the effect of adding more than one cut-off point on CA and CC to answer research question three.

Participants

The study was conducted in a non-profit university in Ankara, Turkey. It is 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, in which 3,000 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 one or two academic years and had finished the highest level of English language instruction offered in the university 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 English medium instruction at the

undergraduate level at the university (Kantarcıoğlu, 2012). Administered three times an academic year, the institution's proficiency test is taken by around 3,000 students annually, in September, January, and June.

The data used in the study comes 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. Results are expected to provide an insight into the potential administration of this test using CAT. Furthermore, CATs with 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

The responses were dichotomously scored (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, students' scores on the listening subtest of the same exam were utilized (the examinees' listening data was calibrated by the university to estimate ability and item parameters). A Pearson correlation coefficient between scores from the reading and listening subtests was calculated.

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 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 re-administering items adaptively (Weiss, 2005). Post-hoc simulations use real examinees' responses to a PBT to simulate testing behavior on a computer as if examinees are 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) was applied to generate ability estimations 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). Since EAP can produce ability estimates with such response patterns, they were not removed from the dataset.

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 standard error rates were selected (i.e., below 0.5, 0.4, 0.3, 0.2, 0.1) corresponding to different alpha values from 0.75 to 0.99 and labelled as SE05, SE04, SE03, SE02, and SE01 (Weiss, 2011). Thus, 10 different CATs 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 standard error 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 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 cut-off values were set from the 10^{th} to the 90^{th} 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 cut-off 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, 1.42. For classification analysis, the corresponding ability values at the same percentiles were used. An additional classification analysis was made using five cut-off 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 cut-off points were set using 16.6, 33.3, 50, 66.6 and 83.3 as percentile ranks. These cut-off points were -0.92, -0.46, -0.12, 0.40 and 0.91. CA and CC values were calculated both for one cut-off score at a time as well as a simultaneous analysis of all cut-off 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, results of the IRT model-data-fit are presented. 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 establish the factorial structure of the trait measured. 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). Standard error of PBT had a mean (standard deviation) of 0.34 (0.25). 48.0% of the individuals had a standard error of 0.4 or lower, whereas 90.4% had 0.5 or lower. Figure 2 presents the PBT item information curve.

INSERT FIGURE 2 HERE

Ability Estimates by PBT and CAT

To establish the extent to which ability estimates vary according to the PBT, descriptive statistics were calculated. 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 the PBT, though not considerably. As Table 1 shows, the paired-samples t-test results indicated statistically significant mean differences after Bonferroni correction for ten 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.

INSERT TABLE 1 HERE

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 terminated 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.

INSERT TABLE 2 HERE

Table 3 demonstrates that the mean SE values are around 0.40 for the CATs terminated 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 are at or below 0.50 SE for the fixed-length CATs terminated with 15, 20, and 25 items respectively. For the same fixed-length CATs, the percentages of ability estimations at or below 0.40 SE are around 12%, 27%, and 38%, respectively. Given that PBT produced a mean SE of 0.34, reliability of estimates by CAT were comparable.

INSERT TABLE 3 HERE

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

INSERT TABLE 4 HERE

Classification

To compare PBT and CAT classification and to explore the impact of adjusting cut-off points on CA and CC values in PBT and CAT, Figures 3a-d present the CA and CC estimates with 9 cut-off points. Figures 3a-d show 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 cut-off 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 cut-off score.

INSERT FIGURE 3a-d

Figures 3a-d show that both CA and CC estimates for all CATs with different termination rules are higher than 75%. In all CATs, CA estimates are slightly higher than CC estimates. CA and CC estimates show an upward trend at the higher and lower tails of ability estimates. It seems that the higher the ability estimate is, the higher CA and CC estimates are 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 is due to an absence of disproportionately difficult or easy items. CATs with fixed-items produce differing CA and CC values around the middle ability range (especially at the 50^{th} and 60^{th} percentiles), whereas SE-based CATs produce relatively more similar CA and CC estimates. At lower and higher ability levels, CAT CA and CC values are similar to the values in PBT; however, around the middle ability levels, the PBT records higher CA and CC values. CAT scenarios with lower numbers of items or higher SE values (FL10 and SE05) are 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 are above 85% and the high ability estimates are around 95%, which is 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 cut-off points, a similar pattern in CA and CC values was observed around the middle ability range (see Figures 4a-d). To be specific, excluding FL10 and SE05, CATs produced CA estimates greater than 85% and CC values higher than 80% in the middle ability range. For fixed-length CATs, FL30 seems to produce slightly higher CA and CC values and for CATs terminated after a pre-specified SE threshold, CA and CC values slightly increase in SE01, SE02, and SE03 across all ability ranges. Another trend observed in Figures 4a-d is that like classification with nine cut-off points, CATs with fewer items or higher SE values showed relatively lower classification performance. This was expected because fewer items are available at the cut-off points, thus increasing misclassification rates. Similarly, applying high SE values results 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 cut-off points, CA and CC values are 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 cut-off points, CA values in the lower and upper percentiles fall within 85% and 95%. CC values, on the other hand, are 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 cut-off points, CAT classification performed was equal to or slightly lower than the PBT version of the test.

INSERT FIGURE 4a-d

Figure 5 shows CA and CC values for each CAT scenario when nine and five cut-off points were applied simultaneously. CA and CC values significantly decreased compared to classifications with a single cut-off score. This is expected because increasing classification categories simultaneously increases the possibility of classification error. Classifications with a single cut-off 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 show that for all CAT scenarios, CA and CC values tend to rise significantly when the number of cut-off points decreases. As stated above, there is an inverse relationship between the number of decision categories and classification performance. Fewer items produce higher errors in ability, resulting in more misclassification. The classification with nine cut-off scores produced lower CA and CC values than with five cut-off scores. In general, different CAT scenarios produced similar CA and CC values for a given number of cut-off points. The figure also shows that with fixed-item CATs, higher numbers of items leads 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 becomes possible. With sufficient match between item difficulty and individual ability, this is both true for CAT and PBT. In all cut-off point arrangements, PBT CA and CC values are higher than all CAT values although differences are minor. With the CATs terminated at SE01 and SE02, CA and CC values increase and are equal to the PBT.

INSERT FIGURE 5 HERE

Construct Coverage

Item distributions across the simulations were investigated and the 25 most frequently used items were recorded (see Appendix). 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 features 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 is a result of the limited items identified as testing categories D and E.

To summarize the results in brief, with a single cut-off score, CAT-based classification performed very well but not for every ability level. As shown by Figures 3a-d and 4a-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 cut-off 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 r= .93-1.00). Overall, the simulation results clearly show 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 3a-d and 4a-d show, CA and CC values are 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 are around 80% while CC values go down to 75% for the simulations. PBT-based CA and CC values are 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 one, the PBT and CAT versions of this particular test proved similar in terms of classification performance.

Research question two asked about the effect of the location of the cut-off point on CA and CC values. To answer this question, when only one cut-score is used for a binary decision such as pass or fail, classification performance of the test is 77% for all the ability groups using both test termination rules. For classifications with a single cut-off point, the graphs consist of lines with a U-shape curve, suggesting that the classification performance for low and high ability groups is 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 are above 75%, and reach around 90%. A U-shaped trend was observed in each analysis, meaning that CAT achieves 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 three, which asked about the effect of operating multiple cut-off points at the same time in all CAT scenarios, CATs with five cut-off points were better able to classify examinees than those with nine cut-off points. After reducing the number of cut-off 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 cut-off points along the ability continuum. When the

number of points decrease from nine to five, the locations and the distances between them also changes, which creates more heterogeneous groups. This indicates that a lower number of cut-off points creates groups with less homogeneity in the latent trait. Increasing the cut-off 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 cut-off points consistently perform almost 10% better than all the other CATs (see Figures 4a-d). CA estimations with five cut-off and nine cut-off points are 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 cut-off 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 cut-off 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 cut-off 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 cut-off 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 cut-off 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 CATs featuring multiple reading passages using the testlet approach.

It is important to acknowledge 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 a typical CAT or CACT 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

The study was developed to determine the impact of introducing a non-classification CAT on the classification accuracy and consistency of decisions relating to students' English language ability. To this end, the accuracy and consistency of classification decisions were investigated with single and multiple cut-off points in different CAT simulations. Overall, given that the test and item bank was designed for PBT with full fixed item administration, CAT yielded promising levels of classification accuracy and consistency 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 cut-off 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.

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/cat09babcock.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., & 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

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.

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 Associates.

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/10125/25044

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

Gnambs, 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 & Sons Inc. https://doi.org/10.1037/13240-000

Gushta, M. M. (2003). *Standard-setting issues in computerized-adaptive testing*. Paper presented at the Annual Conference of the Canadian Society for Studies in Education, Halifax, Nova Scotia. 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 Publications.

Han, K. C. T. (2016). Maximum likelihood score estimation ~~m~~ethod with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement. 40*(4), 289-301. https://doi.org/10.1177/0146621616631317

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. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi,* 6(1), 173-176. https://doi.org/10.21031/epod.15905

Kantarcıoğlu, 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 Publishing.

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). *A Comparison of Methods for Estimating Classification Consistency.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Larson, J. W. (1987). Computer-assisted language testing: is it portable? *ADFL Bulletin, 18*(2), 20-24.

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

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package cacIRT. *Practical Assessment, Research & Evaluation, 20*(18), 1-5. https://doi.org/10.7275/43vm-p442

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*. https://doi.org/10.3389/fpsyg.2019.00829

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

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.* http://www.jattjournal.com/index.php/atp/article/view/48363/0

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.* 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, Inc. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR93-07.pdf

Spray J. A., & Reckase M. D. (1994). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the national meeting of the National Council on Measurement in Education, New Orleans, LA.

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). Wileyu 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 Associates.

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., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer (2nd edition)*. Lawrence Erlbaum Associates.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(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). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA. 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]*. St. Paul, MN: 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*). Mutual information item selection in multiple-category classification CAT*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego, CA. 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

Details of the Twenty-Five Most Frequently Used Items in CAT Simulations

Table A

Categories of the Twenty-Five Most Frequently Used Items in CAT Simulations

Test Termination Rule																	
SE03			SE04			SE05			FL30			FL25			FL20		
Item #	Domain	Count	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	33	A	1182
35	A	1182	35	A	1182	14	C	1097	35	A	1182	35	A	1182	29	B	1182
12	A	1182	28	A	1182	28	A	1079	12	A	1182	12	A	1182	4	B	1161
28	A	1182	11	B	1182	11	B	923	18	A	1182	4	B	1182	35	A	1155
18	A	1182	14	C	1182	27	B	795	20	A	1182	23	B	1182	14	C	1126
16	A	1182	27	B	1178	33	A	759	34	A	1182	15	B	1180	23	B	1078
19	A	1182	29	B	1177	18	A	736	4	B	1182	21	B	1178	12	A	1045
17	A	1182	12	A	1115	19	A	718	23	B	1182	29	B	1174	21	B	1014
20	A	1182	17	A	1108	16	A	685	21	B	1182	9	C	1171	28	A	977
29	B	1182	23	B	1084	23	B	679	11	B	1182	14	C	1160	18	A	921
4	B	1182	20	A	1073	21	B	606	15	B	1182	20	A	1126	9	C	921
23	B	1182	3	C	1064	35	A	588	14	C	1182	6	D	1122	11	B	913

21	B	1182	19	A	1060	30	B	570	9	C	1182	18	A	1081	15	B	898
11	B	1182	4	B	991	9	C	523	7	C	1182	7	C	1070	16	A	826
15	B	1182	18	A	966	4	B	444	6	D	1182	28	A	1052	27	B	792
27	B	1182	15	B	964	32	B	442	29	B	1179	11	B	1002	19	A	741
30	B	1182	21	B	955	5	A	422	28	A	1168	16	A	960	17	A	683
13	B	1182	13	B	887	3	C	416	24	B	1156	27	B	889	30	B	679
14	C	1182	16	A	871	1	A	403	17	A	1150	30	B	843	20	A	653
9	C	1182	9	C	836	13	B	398	3	C	1109	3	C	778	3	C	632
3	C	1182	7	C	817	17	A	387	16	A	1105	19	A	768	32	B	559
7	C	1182	30	B	783	25	D	385	10	B	1099	17	A	742	13	B	546
6	D	1182	2	B	762	2	B	363	32	B	1079	34	A	682	25	D	520
31	E	1182	6	D	743	22	A	350	30	B	1004	32	B	660	6	D	496
34	A	1179	34	A	729	24	B	305	27	B	988	13	B	614	22	A	488

Note. 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.

Table 1

Means (SDs) of the Ability Estimates by CATs Averaged over 100 Replications

CAT	*M (SD)*	*t*	*df*	*P*	*d*
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

Note. 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.

Table 2

Descriptives of the Number of Items Used in CATs with Different SE Thresholds

CAT	*M* (*SD*)	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%)

[a]The values in the parentheses show the percentage of examinees who were given all of the 35 items in the simulations.

Table 3

Descriptives of SEs in Fixed-Length (FL) CATs

CAT	*M* (*SD*)	% of examinees with SEs below SE04	% of examinees with SEs 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

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

Note. All correlations are statistically significant at .001.