EQUIVALENCY EVIDENCE OF THE ENGLISH COMPETENCY TEST ACROSS DIFFERENT MODES: A RASCH ANALYSIS[1]

Muhammad Yoga Prabowo[a], **Sarah Rahmadian**[b]
([a]muhammadyoga@student.unimelb.edu.au; [b]sarah.r@kemenkeu.go.id)

[a]*University of Melbourne, Australia*
Grattan Street, Parkville, Victoria, 3010, Australia

[b]*Financial Education and Training Agency*
Ministry of Finance, Indonesia

Abstract: The outbreak of the COVID-19 pandemic has transformed the educational landscape in a way unseen before. Educational institutions are navigating between offline and online learning worldwide. Computer-based testing is rapidly taking over paper-and-pencil testing as the dominant mode of assessment. In some settings, computer-based and paper-and-pencil assessments can also be offered side-by-side, in which case test developers should ensure the evidence of equivalence between both versions. This study aims to establish the equivalency evidence of different delivery modes of the English Competency Test, an English language assessment for civil service officers developed and used by the Human Resources Development Education and Training Center, a civil service training institution under the Ministry of Finance of the Republic of Indonesia. Psychometric analyses were carried out with the Rasch model to measure the unidimensionality, reliability, separation, and standard error of measurement estimates. The findings demonstrate that the paper-and-pencil and computer-based versions of the language assessment exhibit comparatively equivalent psychometric properties. The computer-based version of the English Competency Test is proven to offer a reliable and comparable alternative to the paper-and-pencil version.

Keywords: computer-based testing, mode effects, paper-and-pencil testing, psychometric properties, Rasch model

DOI: http://dx.doi.org/10.15639/teflinjournal.v34i2/301-319

The paper-and-pencil testing has been in use for a long time in educational measurement and is regarded as the traditional assessment format. In this conventional method, students use a pencil or pen to write their answers or darken the circles on a scannable answer sheet (He & Lao, 2018). With the outbreak of COVID-19 in early 2020, the dominance of paper-and-pencil testing was eventually challenged by computer-based testing, which henceforth will be referred to as PPT and CBT, respectively. As a part of computer-assisted learning, CBT was introduced in the 1960s and has been used since then in language testing to make the process more efficient

[1] This article is based on a paper presentation at the 20th Asia TEFL – 68th TEFLIN – 5th iNELTAL International Conference at Universitas Negeri Malang, Indonesia, on 5 – 7 August 2022.

(Chapelle & Voss, 2016). After the pandemic, language assessment providers had to shift or adapt their tests to address various challenges arising from restrictions implemented by governments and institutions during the outbreak (Ockey, 2021). This has led to the proliferation of CBT at various educational levels and contexts.

While the majority of standardized language assessments are developed for educational contexts (e.g., TOEFL, IELTS), language assessment is widely applicable in other domains. For instance, the Australian Defence Force English Language Profiling Systems (ADFELPS) is developed specifically for military officers to assess their English language proficiency before deployment abroad (Yuzar & Rejeki, 2020). This reflects the fact that English language skills are increasingly required for professional and international communication between world governments. In Indonesia, several scholarships are offered to civil servants, military personnel, and police officers as targeted groups to develop human resources in the government sector (Indonesian Endowment Fund for Education Agency, 2022). Considering this fact, the Human Resources Education and Training Center, a civil service training institution under the Ministry of Finance of the Republic of Indonesia, developed the English Competency Test (ECT). The ECT is a standardized language test that is used for a range of purposes within the ministry, including competency mapping, training requirements, and scholarship selection (Prabowo & Rahmadian, 2022). The test comprises three sections: Listening, Structure, and Reading, each of which focuses on a different linguistic skill.

Up until early 2020, the ECT was administered as a paper-and-pencil test. In response to the COVID-19 outbreak, the ECT has been delivered as a computer-based test since 2021, which can be administered either at test centers or online (Prabowo & Rahmadian, 2022). The emerging trend of shifting to computer-based testing is also observed in other major standardized language assessment providers. The Test of English as a Foreign Language Internet-Based Test (TOEFL iBT) has been offered as TOEFL iBT Home Edition that can be taken from anywhere online since March 2020 (Papageorgiou & Manna, 2021). In the same year, the International English Language Testing System (IELTS) was relaunched as fully online IELTS Indicator (Isbell & Kremmel, 2020). Prior to this, a version of computer-delivered IELTS (CD IELTS) had been offered at testing centers as an alternative to the paper-and-pencil test version since late 2017 (Read, 2022). These tests are presented in similar content, structure, scoring, and timing to the tests administered in test centers.

Since there remain some circumstances where PPT and CBT versions of a test may be used together, the interpretation of the use of the test results is expected to be comparable across different modes. In broad terms, comparability has been used to refer to situations in which test users can be confident in making comparisons between results obtained at different times, places, or using variations in assessment content and procedures (Berman et al., 2020). The application of CBT raises one fundamental issue in how the computer assistance in language testing affects test takers' performance and whether the results of PPT and CBT versions of the same test are comparable (Stoynoff, 2012). Even when the PPT and CBT versions of a test are identical in content, the scores obtained from different modes could have differed. This is because the test-taking process involved in CBT differs from that of PPT; therefore, test performance may be affected by the convenience of the test-taking process and the test takers' computer experience (Wang et al., 2021).

The potential differences arising from different delivery modes are referred to as mode effect. The issue of mode effect has raised concerns for comparability between different test modes for decades. Early literature on computer-based testing demonstrated that computer familiarity, computer anxiety, and attitude toward computerized tests affect performance on the computer-based version of language assessment (Burke et al., 1987; Kernan & Howard, 1990; Powers & O'Neill, 1993). It is also indicated that the degree of complexity of test presentation, such as how the test information is displayed on the screen, could increase the possibility of mode effects (Pommerich, 2004). Such findings imply that test takers' performance in a computer-based test is not only related to the test contents but also distinct characteristics inherent to the delivery mode.

While later research suggests that mode effects are declining and becoming less of a concern with the increasing acceptance of computer-based testing technology (Stricker et al., 2004), it is understood that mode effects are very complex and likely depend on the particular assessment program (Kolen & Brennan, 2014). Various factors in terms of presentation and content can have different extents of influence on the equivalence and interchangeability of PPT and CBT (Wang et al., 2021). Since different studies documenting mode effects have shown mixed results, evaluation of test equivalence should be carried out when a paper-based test is adapted into a computer-based version. The equivalency evidence, or lack thereof, is crucial from the test fairness perspective. When two test forms are of different difficulty levels, examinees could be advantaged or disadvantaged from taking a particular form of a test. Maintaining fairness when assessments are not administered solely in one mode requires the assessments to be scrutinized for equivalency (Cizek & Earnest, 2015). The centrality of fairness in language assessment underlines the reason for which comparability studies should be a priority for research in the future (Stoynoff, 2012).

In response to the comparability issue, several international educational testing associations have developed relevant guidelines. In July 2005, the International Test Commission (ITC) devised the Guidelines on Computer-Based and Internet Delivered Testing, suggesting that test developers should consider the psychometric qualities of the test and ensure the evidence of equivalence between the PPT and its parallel CBT version (ITC, 2006). To put it another way, a test adapted from a PPT to a CBT version should have comparable validity and reliability estimates. This is consistent with the International Language Testing Association (ILTA) Guidelines for Practice which states "If the CBT or internet-delivered test is an adaptation of a pencil and paper test and the tests are used for the same purpose, developers must provide evidence of equivalence" (Berry et al., 2020, p. 3). In a similar vein, Standard 5.12 of the Standards for Educational and Psychological Testing also maintains that "A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably" (American Educational Research Association [AERA] et al., 2014, p. 105).

Following these guidelines, the appropriate statistical methodology should be applied to evaluate the equivalency between the different modes of a particular test. The statistical documentation needs to ensure that current psychometric standards (i.e., validity and reliability) still apply even when the delivery modes may differ (ITC, 2006). Adopting a similar position, Trisnawati (2015) argued that PPT and CBT test results could be equivalent as long as the test

design and algorithm were well-designed; therefore, psychometric issues should be a focus in establishing test equivalence as a part of the validity evidence as demonstrated by previous studies such as Retnawati (2015) and Papageorgiou and Manna (2021). Retnawati (2015) compared the classical reliability coefficients between the PPT and CBT versions of the Test of English Proficiency (TOEP) to evaluate the equivalency evidence of the test. Similarly, Papageorgiou and Manna (2021) also recommended that each test mode's standard error of measurement be compared to establish test comparability.

A number of statistical analysis methods can be used to analyze the psychometric properties of language assessment practices. Among these methods, the Rasch model is widely adopted and considered advantageous in language assessment construction and evaluation (Aryadoust et al., 2021). The model addresses the limitations of the earlier classical test theory (CTT), which is dependent on the particular sample and test examined (Erguven, 2013; Magno, 2009). This implies that item and person statistics computed from CTT analysis could vary significantly depending on the sample of respondents. On the contrary, Rasch measurement is entirely sample-free (Bailes & Nandakumar, 2020). As a result, item measurement is relatively stable regardless of the particular sample used. This characteristic is valuable, particularly when multiple analyses are carried out on data collected from different groups.

In addition to overcoming the practical limitations inherent in the classical test theory, the Rasch model offers improved precision and additional techniques to evaluate the quality of an assessment instrument (Boone, 2016). These additional techniques can be useful when traditional psychometric estimates only provide limited information. The Cronbach's alpha coefficients, traditionally used as the measure of reliability or internal consistency, only evaluate consistency across instrument items but cannot estimate the ability of the instrument to discriminate between people being tested (Merkin et al., 2020). It is also difficult to rely on this internal consistency alone because these estimates are sample-dependent (Magno, 2009). The Rasch measurement provides two additional reliability indices: person reliability and item reliability of instruments and respondents, which represent "an important additional tool to aid the development and use of measurement instruments in many fields" (Boone et al., 2014, p. 223).

Other important reliability metrics provided by the Rasch model are person and item separation indices, which indicate the statistically different levels of person ability or item difficulty measured by an instrument (Linacre, 2022). Person and item separation indices are useful for classifying examinees and confirming item difficulty hierarchy, respectively (Boone et al., 2014). Such indicators would be advantageous in evaluating high-stakes testing that requires more stringent standards. Another way of examining comparability is to measure the stringency of the assessments using the standard error of measurement (SEM). When the accuracy of the test result is the bottom line of the analysis, SEM would be the most appropriate metric of quality (Berman et al., 2020).

In recent years, the issue of PPT and CBT comparability has drawn the attention of international researchers who conducted comparability analysis with a variety of designs. Several studies in the domain of English language assessment have investigated the comparability of different modes of English language proficiency testing using the common person design (Ebrahimi et al., 2019; Hosseini et al., 2014; Khoshsima & Toroujeni, 2017). In

comparability studies with common person design, the data are collected from a singular group of examinees who are administered both PPT and CBT versions of the same test. Overall, these studies were mainly based on the comparison of test takers' scores between different modes, which were then analyzed with ANOVA to examine the significance of the differences. Among these studies, Ebrahimi et al. (2019) and Khoshsima and Toroujeni (2017) also reported the comparison of reliability indices using traditional Cronbach's alpha coefficients. Another type of comparability study collects data from two groups who each take a different version of the test. For example, one group is administered the PPT version and the other the CBT version. To date, there are fewer studies with this design, such as Retnawati's (2015), which compared the psychometric properties of English language assessment responses data from two different groups of examinees, each taking one version of the test. The study carried out psychometric analyses which focused on comparing each group's reliability indices and test information estimates. However, only the test information values were computed with the Rasch analysis, while the reliability indices were obtained from classical test theory computation. Consequently, the depth of information provided is more limited than what the Rasch model has to offer, and its values are subject to the sample-dependent effect. Examination of the Rasch model's key assumptions was also absent in the latter study despite the model's reliance on the strong assumption to produce valid statistics (Hambleton & Swaminathan, 1985).

The application of the Rasch model on the comparability study of the Ministry of Finance's ECT across different modes not only can provide evidence of equivalency between the PPT and CBT versions but also can expand the existing body of research in the field of language assessment by presenting a more comprehensive evaluation through the Rasch model. Moreover, the ECT is currently offered in both PPT and CBT delivery modes as a means of mapping employee's English language competency. Convincing evidence of equivalency across modes should be provided as a requirement for these different modes to be used interchangeably as per the International Guidelines on Computer-Based and Internet-Delivered Testing (ITC, 2006), the Standards for Educational and Psychological Testing (AERA et al., 2014), and the International Language Testing Association Guidelines for Practice (Berry et al., 2020). For that reason, analyzing the psychometric properties between the PPT and CBT versions of the test is necessary to warrant a fair and equitable comparison of examinees' results regardless of the mode administered to them. This is a step towards more accountable assessments, which have increased in demand in education as well as in bureaucracy.

Based on the background and the literature review, the following research question guided the study: Do the different delivery modes of the English Competency Test affect the test's psychometric properties? Accordingly, the primary purpose of this paper was to assess the evidence of equivalence between the paper-and-pencil and the computer-based versions of the English Competency Test. The secondary purpose was to thoroughly examine the psychometric quality of the test, as demonstrated by the Rasch analysis.

METHOD

Participants

The data for the present study were collected from two groups of participants who were employees of the Ministry of Finance of Indonesia. All of them were Indonesian native speakers who had completed undergraduate degree. The first group was administered the PPT version of the ECT, while the second group was administered the CBT version of the test. The data were obtained from test responses stored in the Human Resources Education and Training Center's database. Both groups took the test forms for the first time, reducing the possibility of pre-exposure to the items. The PPT test response data were obtained from several test batches from various Indonesian cities throughout 2017 to 2019. On the other hand, the CBT test response data were collected from one batch of a pilot test delivered in 2021 using the same test form. Therefore, the original data sources of PPT and CBT examinees were asymmetric in size. Table 1 breaks down the distribution of participants by mode and section.

Table 1. Distribution of English Competency Test Participants

Test Mode	Number of Participants		
	Listening	Structure	Reading
Group 1 (PPT)	309	353	189
Group 2 (CBT)	165	165	165

Instruments

The test was administered in the multiple-choice question format with four options. The audio and written stimuli were in North American English as specified by the test blueprint. The PPT version was administered with paper test booklets and scannable answer sheets utilizing digital marking recognition technology. On the CBT version, the test questions and answer choices were displayed on-screen, which could be operated with a computer mouse. Both versions' content, format, timing, and scoring were identical. This type of conventional to computerized test adaptation is referred to as linear or fixed form, which essentially replicates the conventional paper-and-pencil test administration model (Davey, 2011).

Despite the similarity, there were a number of differences between the two versions. The Listening section of the PPT version was administered through a central loudspeaker, while on the CBT version, it was administered through individual headsets. In both delivery modes, the audio was played only once. Another notable difference was in the item presentation. The PPT version presented multiple items on one page, whereas the CBT version displayed the items in a single-item presentation (one question per webpage). This difference is more noticeable in the Reading section. On the PPT version, each reading passage was printed once and followed by a set of ten questions, while on the CBT version, the reading passage was displayed on top of an individual question and reloaded when test takers proceeded to the next question.

In both delivery modes, general directions were displayed and read aloud to provide information on test policies, such as timing and test rules. Examinees were explicitly informed

that there was no penalty for incorrect answers. Additionally, the examinees on the CBT version were provided with a practice section to familiarize themselves with the CBT platform prior to the actual test (Prabowo & Rahmadian, 2022).

To enable a comparison between the PPT and CBT modes, a single form (question set) containing identical test items was used in both modes. This particular form of the ECT was chosen because it has been administered in both PPT and CBT modes, with more than 100 participants per mode in total. The test form is solely used for low-stakes purposes (i.e., tests with relatively low consequences), such as mapping employees' competency within the Ministry of Finance. The test comprises 140 items in total, as presented in Table 2.

Table 2. English Competency Test Composition and Number of Items

Order	Section	Number of Items
1	Listening Comprehension	50
2	Structure and Written Expression	40
3	Reading Comprehension	50

Data Analysis

It is essential that each group analyzed contain the same number of participants to prevent higher reliability estimates which come as an artefact of a larger sample size (Bond & Fox, 2015). Therefore, a sample of 120 examinees from each group was obtained using the simple random sampling method from the pool of test takers' response data. The particular sample size is considered feasible to draw from each group of the data pool after accounting for data with substantial missing responses (items unanswered by examinees). The rationale behind the random selection of the sample is to provide more accurate estimates, which could be influenced by sampling heterogeneity (Cappelleri et al., 2014).

The Rasch model analysis was chosen based on its particular suitability for analyzing data obtained from different groups due to its sample-independent characteristic. The Rasch model also requires fewer data than other item response theory (IRT) models to produce stable results, comparatively. IRT analysis, especially the 2-parameter and 3-parameter models, generally requires large samples ($n \geq 500$) to produce accurate and stable estimates (Cappelleri et al., 2014). In comparison, a dichotomous Rasch model analysis requires as few as 30 respondents to be performed (Linacre, 1994) or at least a sample size of 100 to produce adequately stable estimates (Chen et al., 2014). The sample size of 120 is adequate to provide item calibrations and person measures stable within ± ½ logit with 99% confidence (Linacre, 1994).

The responses and answer keys data were examined for miskeying (incorrect answer key) and missing responses before data processing. RStudio was then used to generate a sample size of 120 for each section and mode based on clean response data. Subsequently, the data were exported to Winsteps statistical software, where a dichotomous Rasch analysis was performed. In dichotomous analysis, the wrong responses were scored 0, and the correct responses were scored 1. The Rasch analysis produced the following metrics of psychometric validity and reliability: unidimensionality, reliability (person, item, Cronbach's alpha), separation (person

and item), and standard error of measurement (SEM) statistics. The resulting data were then compared to evaluate the equivalency evidence across different modes of the ECT.

FINDINGS AND DISCUSSION

Findings

Technical analyses carried out using Winsteps resulted in the estimation of psychometric properties for both versions of the English Competency Test. The results were reported and discussed from the following aspects of psychometric analysis: unidimensionality, reliability and separation, and standard error of measurement.

Unidimensionality

The unidimensionality assumption constitutes a crucial requirement of the Rasch model. This assumption demands that each attribute or 'dimension' be measured one at a time (Bond & Fox, 2015). In English language assessment, the dimensions comprise listening or reading ability, for instance. Unidimensionality suggests that the instrument measures a singular underlying latent construct (Fan & Bond, 2019). When an instrument is unidimensional, it signifies that there is no presence of other substantial dimensions in the Rasch model's residuals (Aryadoust et al., 2021). As a key assumption of the Rasch model, if unidimensionality is not met, the Rasch analysis would be meaningless (Boone et al., 2014). In other words, the inferences taken from the Rasch model would be of questionable value.

Principle component analysis of residuals (PCAR) is a commonly used method to evaluate test unidimensionality assumption via two indicators: the raw variance explained by measure and the raw unexplained variance in first contrast (Ishak et al., 2018). PCAR identifies the attenuation of unidimensionality by detecting the differences between the predicted and actual data, which constitute the components in Rasch residuals (Aryadoust et al., 2021). To indicate unidimensionality, the raw variance of a test instrument should be higher than 20% (Sumintono & Widiarso, 2014, as cited in Ahmad & Siew, 2021). Additionally, the unexplained variance in first contrast should be lower than 15% (Ishak et al., 2018).

The raw variance explained by measure statistics are visualized in Figure 1. It is apparent from the figure that the raw variance explained by measure values showed variations within an acceptable range (≥ 20%). The Listening and Structure sections in the CBT mode exhibited slightly higher raw variance explained by measure values compared to their PPT counterpart, which is more favorable. In contrast, the PPT version of the Reading section demonstrated a higher value of raw variance explained by measure than the CBT version.

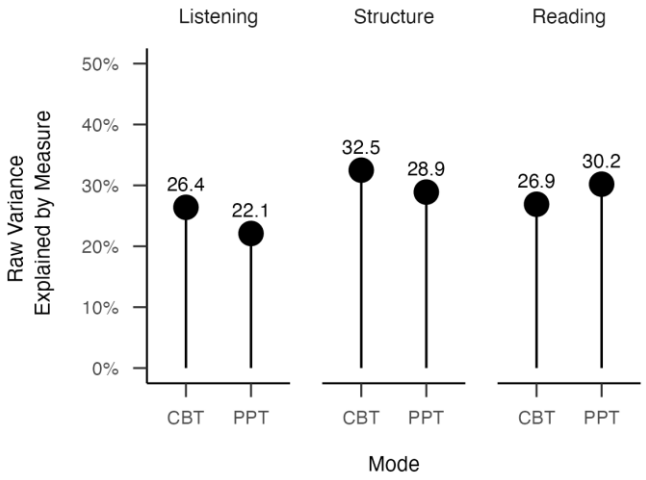

Figure 1. Raw Variance Explained by Measure Statistics of ECT

The raw unexplained variance values as displayed in Figure 2 were all significantly less than 15%. Similar patterns were observed, with the Reading section yielding different results than the other sections across all modes. Nonetheless, the overall results suggested that the instruments met the Rasch model's unidimensionality assumption and thus measured a single dimension of the construct. CBT version tended to show better unidimensionality in the Listening and Structure sections compared to the PPT version, as shown by slightly higher results of raw variance explained by measure estimates and lower raw unexplained variance estimates. On the contrary, the Reading section showed a different pattern, indicating relatively better indicators of the unidimensionality of the PPT version.

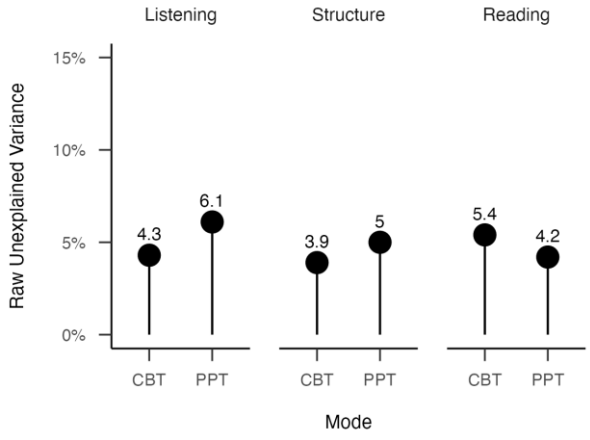

Figure 2. Raw Unexplained Variance Statistics of ECT

Reliability and Separation

Reliability statistics denote the item measures reproducibility when tested to another group or the consistency of person measures if they were retested (Bond & Fox, 2015). The reliability index ranges from 0 to 1, with higher values indicating higher reliability. The ECT's reliability estimates, as demonstrated by person reliability, item reliability, and Cronbach's alpha values generally showed satisfactory values that reflect the internal consistency of the items. Theoretically, a high-stakes assessment should have a reliability index equal to or greater than 0.8 (Carr, 2011). This criterion is met by both the PPT and CBT versions of the test, indicating that both versions are highly reliable and therefore would be likely to produce the same scores when the tests or performance tasks are repeated (Meyer, 2010).

Figure 3 shows that the three types of reliability statistics for both modes closely matched each other. The minimum observed values of Cronbach's alpha, person reliability, and item reliability were 0.86, 0.85, and 0.91, respectively. Across the different modes, the variations observed were marginal (ranging from 0.00 to 0.02). The reliability of item difficulty estimates was the highest among the three coefficients (ranging from 0.91 to 0.96), suggesting a strong likelihood that the test will be able to reproduce the item order hierarchy along the measured variable when it is given to other comparable groups of examinees.

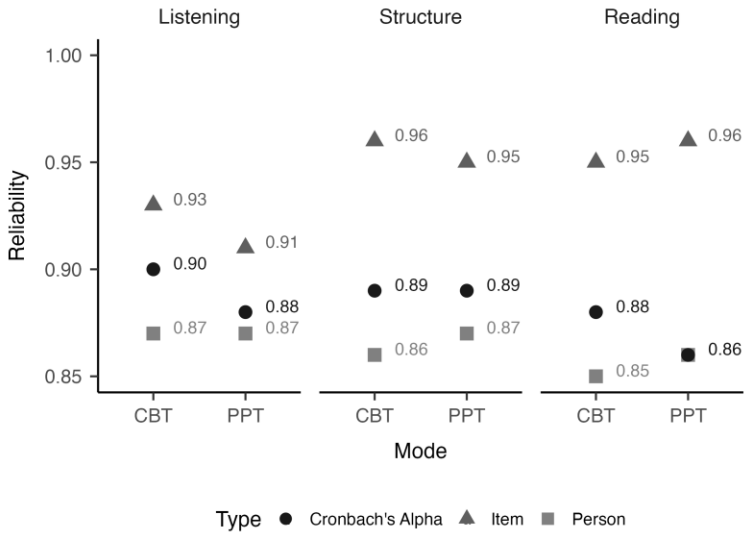

Figure 3. Comparison of Reliability Statistics of ECT

Another important indicator of reliability that the Rasch model provides is separation. The Rasch model analysis distinguishes separation statistics into two measures: person and item separation. Separation index ranges from 0 to infinity, unlike the reliability index, which has a ceiling of 1 (Boone et al., 2014). The person separation coefficient is an integral part of the Rasch measurement that reflects the accuracy and precision of the instrument in separating or

discriminating test takers based on their performances (Cappelleri et al., 2014). Person separation indices of the instrument were invariably higher than two (2.36–2.59), which means the instrument was adequately sensitive to differentiate the test takers into two or three levels. This is useful when we need to classify the test takers into different ability levels (e.g., low-performing, mid-performing, and high-performing groups). There were no stark differences observed between the PPT and CBT person separation estimates.

Furthermore, all item separation values were higher than three (3.16–5.07). It shows that the instrument could confirm the item difficulty hierarchy (Linacre, 2022). On the CBT mode, the item separation coefficients ranged from 3.66 to 4.62. In comparison, the same coefficients on the PPT mode ranged from 3.16 to 5.07. It implies that the instrument can differentiate the items into approximately three to five difficulty levels. As shown in Figure 4, the Listening and Structure sections on the CBT version had higher item separation values, in contrast to the Reading section, which showed higher value on the PPT version. The cross-mode differences were less than one (0.42–0.69). Overall, the Rasch analysis demonstrates that the ECT instrument was highly reliable and capable of differentiating the test takers and items into different levels of ability and hierarchy.

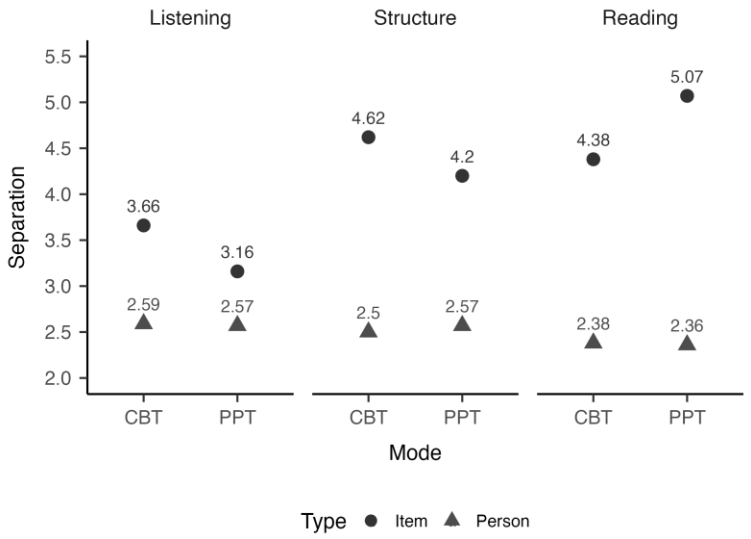

Figure 4. Comparison of Separation Statistics of ECT

Standard Error of Measurement of Items

Reporting the standard error of measurement alongside the reliability estimates is a useful practice in educational measurement (Meyer, 2010). SEM illustrates how precise or accurate an item is in measuring a person's ability, or in other words, "the impact of measurement error on the outcome of the measurement" (AERA et al., 2014, p. 222). Different from reliability and separation coefficients which indicate precision at sample-level, SEM suggests precision at item- or person-level (Aryadoust et al., 2021). It is worth pointing out that any measure is

imperfect and therefore subject to at best a small degree of error (Carr, 2011). A lower SEM is considered better, with a value of 0.5 logit still considered acceptable (Sumintono & Widiarso, 2015). Generally, values below the 0.5 logit threshold are observed on all items, except for two items, each in the CBT Structure and CBT Reading sections which slightly exceeded the threshold by 0.03 and 0.02 points, respectively. The lower mean values of each group also demonstrated that the PPT version of ECT had slightly better precision.

Nonetheless, the overall SEM values of both modes were still well within the accepted range, which is a good indication that the assessment items were quite precise. Table 3 provides the minimum, maximum, and average values of each group's SEM. On average, the ECT's mean SEM ranged from 0.21 to 0.25 for all sections. This finding strongly suggests that the ECT was generally quite accurate in estimating test takers' observed scores in comparison with their true scores.

Table 3. A Summary of Standard Error of Measurement Estimates

Section	Mode	Standard Error of Measurement (SEM)		
		Min	Max	Mean
Listening	CBT	0.20	0.34	0.23
	PPT	0.20	0.23	0.21
Structure	CBT	0.21	0.53	0.25
	PPT	0.21	0.38	0.22
Reading	CBT	0.20	0.52	0.24
	PPT	0.20	0.43	0.23

Discussion

Adopting the Rasch model analysis, the present study aimed at providing equivalency evidence for the paper-and-pencil and computer-based versions of the English Competency Test developed and used by the Ministry of Finance of Indonesia. The discussion of the investigation and findings of this study followed Meyer's (2010) four-pronged outline of reliability documentation, which includes (a) the description of the examinee population, (b) the description of the measurement procedure and research design, (c) the presentation of the assumptions examined, and (d) the reliability estimates and the standard error of measurement (SEM). At the same time, this study made use of the Rasch analysis to provide more comprehensive metrics of validity and reliability.

The data for this research is collected from the Ministry of Finance's employees who participated as ECT examinees. All participants were native speakers of Indonesian who worked for the Ministry of Finance with undergraduate degree qualifications. In that respect, this study is one of the first to examine test modes equivalency in an English language assessment specifically developed to be used in the civil service offices.

As there is no single agreed-upon method for examining test modes equivalency, this study attempted to investigate the subject with the Rasch model. Taking advantage of the Rasch model's sample-independency, this prospective study was designed to use data collected from

two different groups, each taking a different version of the test (PPT or CBT) with test forms that were identical in content, structure, timing, and scoring. This is similar to Retnawati's (2015) study, which collected data from two different groups and also made use of the Rasch model, in contrast to the majority of studies in the field that used ANOVA analysis on data collected from one group of test takers subjected to both PPT and CBT delivery modes (Ebrahimi et al., 2019; Khoshsima & Toroujeni, 2017). However, unlike the aforementioned studies that only reported classical reliability estimates (Cronbach's alpha), this study expanded the range of reliability measurement by including the estimates of person/item reliability and separation, which have hitherto been unreported in similar studies.

As the first critical step, this research tested the unidimensionality assumption of the ECT. Despite its importance as an underlying assumption of the Rasch measurement, unidimensionality is largely unreported in language assessment publications or claimed through unsuitable measures (Aryadoust et al., 2021; Taber, 2018). In addressing the gap, this study faithfully examined the unidimensionality of ECT through PCAR analysis. The findings suggest that the unidimensionality assumption on both versions of the test was not violated, and therefore the comparison derived from the Rasch analysis could be meaningfully interpreted and used.

The instrument's reliability was thoroughly examined using a wide range of indices, including the traditional Cronbach's alpha reliability coefficient, person and item reliability coefficients, as well as the person and item separation indices. The results revealed not only the test's replicability across different modes but also the instrument's sensitivity in categorizing examinees and items into different levels or strata based on the person ability and item difficulty. In general, both versions of the test were proven to have fairly high reliability as reflected by the estimates of Cronbach's alpha, person reliability, and item reliability (≥ 0.85). Person separation coefficients were adequate (≥ 2), indicating the instrument was quite sensitive in classifying the examinees into different levels of ability. Furthermore, the high item separation (≥ 3) and item reliability observed across all modes suggest that the sample size was sufficient to confirm item hierarchy, providing support for the instrument's construct validity (Boone et al., 2014). At item level, SEM estimates proved that all items were quite precise, except for two items on the CBT version with negligible differences from the threshold. This indicates a lesser amount of test score discrepancy when the test is repeated (Meyer, 2010). Together, these indicators increase the breadth and depth of the reliability measurement (e.g., sensitivity in classifying person ability and item difficulty) and ensure that measurement errors do not jeopardize measurement validity.

With that being said, the findings reveal reasonable differences in a few respects when examined carefully. These are shown by unidimensionality statistics and item separation coefficients which demonstrated slightly higher variances across different modes, while other estimates remained essentially equal. To begin with, the CBT version of the Listening and Structure sections had comparatively better indicators of unidimensionality. In contrast, the PPT version of the Reading section was better, as demonstrated by relatively higher results of raw variance explained by measure estimates and lower raw unexplained variance estimates. A similar pattern was observed in item separation analysis, in which the CBT version had higher values on the Listening and Structure sections as opposed to the PPT version, which performed better in the Reading section.

However, it is safe to say that those differences do not adversely affect mode equivalency for several reasons. First of all, while unidimensionality is the key assumption of the Rasch model, it is not a criterion for the consistency of test scores when the test is replicated. Unidimensionality merely suggests that the instrument measures a single trait or ability (Erguven, 2013). Secondly, the cross-mode item separation differences, of which the highest was observed in the Reading section at 0.69, illustrated that the spread of item difficulty varied by less than one level across different modes. Furthermore, since the test form in this analysis is particularly used for low-stakes purposes (e.g., competency mapping, training requirements), the reliability of person ability estimates and the instrument's sensitivity to classify examinees into different groups of ability levels (i.e., person reliability and person separation coefficients) are of relatively greater importance than item separation. In this context, person reliability and person separation have provided supporting equivalency evidence for the particular test's intended use. This interpretation, however, may not necessarily be generalizable to high-stakes purposes, such as selecting scholarship awardees. When rank is used to determine eligibility for a limited number of scholarships, the comparability study should consider more evidence at test score level (Berman et al., 2020). In doing so, replicating the study on a larger sample size ($n \geq 250$) would be preferable to ensure the definitive parameter estimation stability required for high-stakes tests (Linacre, 1994).

Compared to other sections, the contrastive results from the Reading section are in line with other studies, such as Choi et al. (2003) who found that the reading comprehension of the Seoul National University's Test of English Proficiency exhibited the largest cross-mode discrepancy. It is suggested that the discrepancy in the Reading section might be caused by a different test layout rather than the test content (Choi et al., 2003; Ebrahimi, 2019; Pommerich, 2004). In terms of presentation, the CBT adaptation of ECT Reading section is arguably more distinct from its PPT counterpart than the adaptation of other sections. The CBT version of the Reading section displayed the reading passage in repetition alongside an individual question, affecting the visual process when the examinees tried to answer the question. Another factor to consider is screen size, which may influence examinees' performance (Wang et al., 2021). Prabowo & Rahmadian (2022) reported that some ECT examinees indeed expressed that the font size on the CBT test was too small on a 22-inch screen. This suggests an area where the presentation of the computer-based adaptation could be improved to decrease the examinees' performance gap.

In terms of accuracy, this study demonstrates partial similarity to Retnawati (2015), which showed a tendency that the PPT version of TOEP was more accurate in certain conditions (i.e., test takers with low and high ability levels), while the CBT version showed higher accuracy for examinees with moderate ability levels. Through the examination of SEM, the PPT version of ECT was shown to have relatively lower measurement errors on average and no item exceeding the precision threshold, indicating a higher level of accuracy in all sections. The average SEM for the CBT version of the ECT was slightly lower but nonetheless still comparable. However, it should be noted that the former study's measure of accuracy was computed through test item function at sample-level, while this study's SEM estimates were measuring precision at item-level. At a higher level, the ECT's precision could be inferred through various reliability coefficients, as discussed earlier. The mixed findings relative to previous studies seem to be in

accordance with Kolen and Brennan's (2014) notion that the presence and significance of mode effects would likely be distinctive for each test.

Taken together, the above findings reveal that the effects of different delivery modes on the psychometric properties of ECT were fairly inconsequential. This is shown by the high degree of equivalency between the psychometric properties of the PPT and CBT versions of the test, which is essential for establishing the validity of comparative inferences. This provides supporting evidence for the interchangeability of test scores, from which we could expect consistency. This is very useful when the test is offered in two different modes. In addition, the study indicates the satisfactory quality of the psychometric properties of the ECT in both PPT and CBT modes, suggesting that the ECT is a valid and reliable English language assessment for the specific purpose for which it is used.

CONCLUSIONS

This research set out to provide supportive evidence of comparability between the PPT and CBT versions of the English Competency Test, an English language test that is developed and used by the Ministry of Finance of Indonesia. By means of performing the Rasch model analysis on the low-stakes form of the test, the investigation revealed the equivalency evidence as shown by the similar psychometric analysis results on both modes. With respect to the research question of this study, the different delivery modes were proven to have no substantial effects on the test's psychometric properties. The evidence presented thus far supports the idea that the computer-based version of the ECT could be used as a reliable and comparable alternative to the paper-based version.

This research contributes to the field of language assessment by providing empirical evidence of the equivalency across different delivery modes of the ECT through a comprehensive psychometric analysis. This represents a further step towards a fair and accountable assessment that is based on sound psychometric principles. Another important implication of this study is that it is documenting the use of computer based ECT within the context of language competency development in civil service offices, which remains a largely uncharted area in language assessment research. By doing so, it also supports the implementation of digitalization in human resource management which constitutes an important pillar of Indonesian bureaucratic reform.

Although the study has successfully demonstrated the equivalency evidence of the paper-based and the computer-based modes of ECT, this study’s limitation was on the use of response data from a test form exclusively used for low-stakes purposes. It is important to acknowledge that this limitation means the study findings need to be interpreted cautiously when the test is specifically used as a high-stakes test (e.g., in scholarship selection). While the equivalency evidence remains valid for the test's use in low-stakes competency mapping, it is recommended that ECT for scholarship selection be administered uniformly in one mode until the cross-mode equivalency evidence for high-stakes use is ascertained. Future research may aim to fill the gap in the limitation of this study to provide definitive evidence for intended use in high-stakes settings. In addition, future research should pay special attention to other sources of comparability threat, such as the differential item functioning (DIF).

ACKNOWLEDGMENTS

We would like to thank our colleagues in the Division of Integrated Test Management and Quality Assurance, Leadership and Managerial Education and Training Center, who provided the assessment data and made this study possible. We would also like to acknowledge the Financial Education and Training Agency of the Ministry of Finance of Indonesia for granting financial support to present this study at the 20th Asia TEFL – 68th TEFLIN – 5th INELTAL International Conference taking place in Malang, Indonesia. Finally, we wish to extend our appreciation to Amalia Nindy Astuti for her diligent proofreading of our manuscript.

REFERENCES

Ahmad, J., & Siew, N. M. (2021). Curiosity towards STEM education: A questionnaire for primary school students. *Journal of Baltic Science Education, 20*(2), 289–304. https://doi.org/10.33225/jbse/21.20.289

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Bailes, L. P., & Nandakumar, R. (2020). Get the most from your survey: An application of Rasch analysis for education leaders. *International Journal of Education Policy and Leadership, 16*(2), 1–19. https://doi.org/10.22230/ijepl.2020v16n2a857

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). *Comparability of large-scale educational assessments: Issues and recommendations.* National Academy of Education.

Berry, V., Kremmel, B., & Plough, I. (2020). *International Language Testing Association guidelines for practice*. International Language Testing Association

Bond, T., & Fox, C. (2015). *Applying the Rasch model,* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education, 15*(4). https://doi.org/10.1187/cbe.16-04-0148

Burke, M. J., Normand, J., & Raju, N. S. (1987). Examinee attitudes toward computer-administered ability tests. *Computers in Human Behavior, 3*, 95–107.

Cappelleri, J. C., Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36(*5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. *Language Learning and Technology, 20*(2), 116–128. http://llt.msu.edu/issues/june2016/chapellevoss.pdf

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research, 23*(2), 485–493. https://doi.org/10.1007/s11136-013-0487-5

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295–320. https://doi.org/10.1191/0265532203lt258oa

Cizek, G. J., & Earnest, D. S. (2015). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 212–237). Taylor & Francis Group. https://doi.org/10.4324/9780203102961

Davey, T. (2011). *Practical considerations in computer-based testing*. Educational Testing Service.

Ebrahimi, M. R., Toroujeni, S. M. H., & Shahbazi, V. (2019). Score equivalence, gender difference, and testing mode preference in a comparative study between computer-based testing and paper-based testing. *International Journal of Emerging Technologies in Learning (IJET), 14*(07), 128–143. https://doi.org/10.3991/ijet.v14i07.10175

Erguven, M. (2013). Two approaches in psychometric process: Classical test theory & item response theory. *Journal of Education, 2*(2), 23–30. https://jebs.ibsu.edu.ge/jms/index.php/je/article/view/84

Fan, J., & Bond T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment, Vol. I: Fundamental techniques* (pp. 83–102). Routledge. https://doi.org/10.4324/9781315187815

Hambleton, R.K., Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. https://doi.org/10.1007/978-94-017-1988-9_2

He, D., & Lao, H. (2018). Paper-and-pencil assessment. In B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1199–1200). SAGE Publications, Inc. https://www.doi.org/10.4135/9781506326139.n496

Hosseini, M., Abidin, M. J. Z., & Baghdarnia, M. (2014). Comparability of test results of computer-based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Procedia - Social and Behavioral Sciences, 98*, 659–667. https://doi.org/10.1016/j.sbspro.2014.03.465

Indonesian Endowment Fund for Education Agency. (2022). *Annual report 2021*. Retrieved from https://lpdp.kemenkeu.go.id/storage/information/report/file/yearly/yearly_report_1662003384.pdf

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*(2), 143–171. https://doi.org/10.1207/s15327574ijt0602_4

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing, 37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Ishak, A. H., Osman, M. R., Mahaiyadin, M. H., Tumiran, M. A., & Anas, N. (2018). Examining unidimensionality of psychometric properties via Rasch model. *International Journal of Civil Engineering and Technology, 9*(9), 1462–1467. http://iaeme.com/Home/issue/IJCIET?Volume=9&Issue=9

Kernan, M. C., & Howard, G. S. (1990). Computer anxiety and computer attitudes: An investigation of construct and predictive validity issues. *Educational and Psychological Measurement, 50*(3), 681–690. https://doi.org/10.1177/0013164490503026

Khoshsima, H., & Toroujeni, S. M. H. (2017). Transitioning to an alternative assessment: Computer-based testing and key factors related to testing mode. *European Journal of English Language Teaching, 2*(1), 54-74. https://doi.org/10.5281/ZENODO.268576

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions 7*(4), 328.

Linacre, J. M. (2022). *A user's guide to Winsteps Ministeps Rasch-model computer programs: Program manual 5.2.2*. https://www.winsteps.com/a/Winsteps-Manual.pdf

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1–11. https://files.eric.ed.gov/fulltext/ED506058.pdf

Merkin, A. G., Medvedev, O. N., Sachdev, P. S., Tippett, L., Krishnamurthi, R., Mahon, S., Kasabov, N., Parmar, P., Crawford, J., Doborjeh, Z. G., Doborjeh, M. G., Kang, K., Kochan, N. A., Bahrami, H., Brodaty, H., & Feigin, V. L. (2020). New avenue for the geriatric depression scale: Rasch transformation enhances reliability of assessment. *Journal of Affective Disorders, 264*, 7–14. https://doi.org/10.1016/j.jad.2019.11.100

Meyer, P. (2010). *Reliability*. Oxford University Press.

Ockey, G. J. (2021). An overview of COVID-19's impact on English language university admissions and placement tests. *Language Assessment Quarterly, 18*(1), 1–5. https://doi.org/10.1080/15434303.2020.1866576

Papageorgiou, S., & Manna, V. F. (2021). Maintaining access to a large-scale test of academic language proficiency during the pandemic: The launch of TOEFL iBT Home Edition. *Language Assessment Quarterly, 18*(1), 36–41. https://doi.org/10.1080/15434303.2020.1864376

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6), 1–45. https://files.eric.ed.gov/fulltext/EJ905028.pdf

Powers, D. E., & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1*, 153–173. https://doi.org/10.1002/j.2333-8504.1992.tb01506.x

Prabowo, M. Y., & Rahmadian, S. (2022). Computer-based English competency assessment for scholarship selection: Challenges, strategies, and implementation in the Ministry of

Finance. *Jurnal Sosioteknologi, 21*(1), 84–96. https://doi.org/10.5614/sostek.itbj.2022.21.1.9

Read, J. (2022). Test review: The International English Language Testing System (IELTS). *Language Testing, 39*(4), 679–694. https://doi.org/10.1177/02655322221086211

Retnawati, H. (2015). The comparison of accuracy scores on the paper and pencil testing vs. computer-based testing. *TOJET: The Turkish Online Journal of Educational Technology, 14*(4), 135–142. http://www.tojet.net/articles/v14i4/14413.pdf

Stoynoff, S. (2012). Research agenda: Priorities for future research in second language assessment. *Language Teaching, 45*(2), 234–249. https://doi.org/10.1017/S026144481100053X

Stricker, L. J., Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based Test of English as a Foreign Language. *Computers in Human Behavior, 20*(1), 37–54. https://doi.org/10.1016/S0747-5632(03)00046-3

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan* [Application of Rasch modelling in educational measurement]. Trim Komunikata Publishing House.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Trisnawati, I. K. (2015). Validity in computer-based testing: A literature review of comparability issues and examinee perspectives. *Englisia Journal, 2*(2), 86–94. https://doi.org/10.22373/ej.v2i2.345

Wang, T. H., Kao, C. H., & Chen, H. C. (2021). Factors associated with the equivalence of the scores of computer-based test and paper-and-pencil test: Presentation type, item difficulty and administration order. *Sustainability, 13*(17), 1–14. https://doi.org/10.3390/su13179548

Yuzar, E., & Rejeki, S. (2020). The correlation between productive and receptive language skills: An examination on ADFELPS test scores. *SALEE: Study of Applied Linguistics and English Education, 1*(02), 99–113. https://doi.org/10.35961/salee.v1i02.111