rEFLections
Vol 32, No 3, September - December 2025

A Rasch-based Validation of the Reading Section of Test of English for Thai Engineers and Technologists (TETET)

NATJIREE JATURAPITAKKUL[1]
PAUL HORNESS[2]
DUANGJAICHANOK PANSA[1*]
[1]*School of Liberal Arts, King Mongkut's University of Technology Thonburi, Thailand*
[2]*The World Language Centre, Soka University, Japan*
***Corresponding author email: duangjaichanok.pan@kmutt.ac.th**

Article information
Article history:
Received: 8 Apr 2025
Accepted: 3 Dec 2025
Available online: 16 Dec 2025

Keywords:
ESP
Rasch model
Reading
Test of English for Thai Engineers and Technologists (TETET)
Validation

Abstract

This study examined the validity of the reading section of the English for Specific Purposes (ESP) test, the Test of English for Thai Engineers and Technologists (TETET), using the Rasch model in relation to Messick's (1989, 1995) six aspects of validity. Data were collected from 179 participants at a science and technology university in Bangkok, Thailand, and analyzed with WINSTEPS Rasch software version 5.4.1. The results indicated that the reading section generally met Messick's criteria, with strong evidence for content, substantive, and structural validity, and preliminary evidence for external, generalizability, and consequential validity. To distinguish it from general English tests, some items, particularly in the survival and internet reading sections, required revision to enhance content (representativeness and technical quality), and substantive aspects. These findings not only provide practical guidance for refining TETET and similar ESP assessments, but also highlight both the strengths and the limits of Rasch analysis in addressing a comprehensive validity framework. Future research should therefore complement psychometric findings with qualitative evidence to build a more complete validity argument.

INTRODUCTION

In recent years, test developers and psychometricians have focused more on improving the quality of language tests, recognizing the challenges of measuring language skills more accurately. Among the psychometric qualities, validity is regarded as the most critical factor (Messick, 1989). It is seen as a validation process that requires a combination of empirical evidence and rational judgment (Engelhard & Wind, 2017; Messick, 1995). The concept of validity in language tests can have significant effects on test-takers, whether they are high-stakes or low-stakes. In both cases, the way tests are designed and administered can influence a test-taker's perception of their abilities, their motivation, and their future opportunities (Shohamy et al., 2017). Therefore, the psychometric quality of tests especially the high-stakes ones are extremely crucial to both test developers and test users. It needs to

be demonstrated to be both scientifically sound and fair to all test-takers (Bachman & Palmer, 2010; Chapelle et al., 2010).

There are several approaches or frameworks to explore validation of language tests (Bachman & Palmer, 1996; Chappelle, 2012; Kane, 2006; Messick, 1989; Weir, 2005). However, Messick's framework for test validation is foundational and highly regarded in the field of psychometrics and educational measurement due to its comprehensive and integrative approach. To elaborate, Messick's framework integrates various forms of validity—content, criterion, and construct—into a unified concept. This comprehensive approach ensures that all relevant aspects of a test's validity are considered simultaneously, providing a more complete assessment of a test's effectiveness (Messick, 1989). In addition, Messick (1994) proposed that construct validity is at the core of this framework, which involves ensuring that the test measures the theoretical construct it is intended to measure. This focus is essential for developing tests that are both meaningful and scientifically sound. By selecting Messick's framework, test developers and researchers can ensure a rigorous and thorough validation process that addresses all critical aspects of test quality and their implications, making it a robust choice for validation in various testing scenarios.

Different aspects of Messick's framework can be investigated in various ways. One of them is to implement the use of the Rasch model to examine whether aspects of validity hold or not. Messick's framework and the Rasch model or the Rasch-Messick link (Bond & Fox, 2015) are complementary in the validation process. The Rasch model provides the empirical foundation to support the validity evidence required by Messick's framework, particularly in terms of construct validity, measurement precision, and fairness. Together, they offer a robust approach to ensuring that tests are both scientifically sound and practically useful.

In the context of Thailand, TOEIC is one of the most extensively required English tests among graduate students for both academic and career focuses. With approximately 3 million test-takers every year (ETS, 2023), it is widely recognized by corporations and institutions worldwide for measuring test takers' English level including Thailand. While it is recognized worldwide, TOEIC mainly benefits students in business, liberal arts and social studies or non-technical professionals (ETS, 2019) as it focuses on specific vocabulary and grammar for everyday work situations. This means it may not fully meet the needs of students in other fields namely, engineering, science, and technology. To fill this gap, the Test of English for Thai Engineers and Technologists (TETET) has been developed specifically for those fields. This computer-based test assesses all four language skills using various item test types and is mandatorily used at a science and technology university in Bangkok to evaluate students' readiness for workplace communication. Even though TETET has gone through standard setting procedures (Jaturapitakkul & Watson Todd, 2018), it needs more evidence to reveal its test quality in regards to different aspects of validity.

This study examines the validity of the reading section of TETET through the Rasch model in relation to Messick's (1989, 1995) six aspects of validity: content, substantive, structural, external, generalizability, and consequential. The reading section was chosen because it reflects the technical demands that engineers encounter in the workplace, and prior research has

identified reading as the most essential skill for engineering instruction and learning (Jaturapitakkul, 2007). The guiding research question is: What is the validity evidence of the reading section of TETET using the Rasch model in relation to six aspects of Messick's validity framework?

The study provides evidence on how well the test items measure intended reading abilities and whether results align with test objectives and item difficulty. Rasch analysis, supported by expert reviews, offers robust findings for content, substantive, and structural validity, while insights into external, generalizability, and consequential validity remain partial. Presenting results across all six aspects highlights both the strengths and limitations of Rasch-based validation and possibly contributes to cumulative validity arguments. At the same time, it points to the need for complementary qualitative evidence—such as expert judgments and stakeholder perspectives—in future research. In addition, the results can provide fruitful feedback to improve the quality of test items and develop more items to be capable of assessing person's reading ability in specific fields.

LITERATURE REVIEW

Validity in language tests

Validity refers to the extent to which a test measures what it claims to measure. In the context of language tests, this means assessing how accurately a test evaluates a test-taker's language ability and whether the results can be interpreted meaningfully (Alderson & Wall, 1993; Bachman & Palmer, 1996). Validity is substantially crucial as it determines the effectiveness and credibility of the test in different contexts such as academic admissions, job placements, and language proficiency assessments.

Validity is extremely essential for high-stakes tests because these assessments significantly influence individuals' educational and career opportunities, requiring accurate measurement to ensure fair and equitable outcomes. Many scholars (i.e., Kane, 2013; Lane et al., 2016; Messick, 1989) acclaim valid tests build trust and credibility with stakeholders, hold institutions accountable, and provide valuable feedback for improving educational programs. Moreover, Kane (2006) stated that validity helps safeguard against bias, ensuring fairness across diverse populations, and is essential for addressing legal and ethical considerations, particularly when the test's use is scrutinized or challenged. By ensuring decisions based on these tests are sound, validity supports the integrity and effectiveness of high-stakes assessments.

Rasch model and Messick's validation framework

Messick's framework and the Rasch model are pivotal in psychometrics, offering complementary approaches to ensuring the validity and precision of measurements. Messick (1989) defined validity as a unified concept that combines different types of evidence to confirm the accuracy and usefulness of test score interpretations. Similarly, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) emphasize validity as a single, unified

construct, supported by multiple sources of evidence including content, response processes, internal structure, relations to other variables, and consequences. Referencing these Standards positions this study within internationally recognized guidelines for test validation and highlights the broader applicability of Messick's framework. The Rasch model aligns with these perspectives by providing a mathematical foundation to ensure that test items measure a single construct, supporting construct validity, measurement precision, and fairness (Bond & Fox, 2015). Both Messick's framework and the Rasch model emphasize connecting theoretical constructs to actual test data, with the Rasch model providing detailed information to improve tests (Kane, 2006). Furthermore, the Rasch model ensures fairness by maintaining consistent measurement across different groups, which supports Messick's emphasis on equity (Wright & Masters, 1982). By combining Messick's comprehensive validity framework with the Rasch model's precision and empirical rigor, they help researchers build strong validation arguments to make tests reliable, fair, and practical. The concepts of Messick's framework and the Rasch model are outlined sequentially.

According to Messick (1989, 1995), construct validity encompasses six distinct aspects that serve as standard criteria for evaluating measurements in education and psychology. These aspects include content, substantive, structural, external, generalizability, and consequential validity. This study examines all six aspects, with each was explained individually.

First, the content aspect of validity in Messick's framework (1995) refers to the extent to which a test or assessment accurately represents the construct it aims to measure. Content aspect deals with three elements: relevance, representativeness, and technical quality. For content relevance, it is about whether the test items are appropriate and relevant to the construct being measured. In other words, it checks if the items truly reflect the intended content area. In terms of representativeness, it involves covering the full scope of the construct without including irrelevant elements, while the technical quality ensures that items are clear, well-designed, and free from flaws that could introduce measurement errors. The Rasch model aids in evaluating these aspects through item separation strata which assess how well the items on a test are spread out along the measured construct (Wright & Masters, 1982), and fit statistics which determine whether individual items and person responses fit the Rasch model's assumptions or not (Bond & Fox, 2015; Kane, 2006). Additionally, wright map which visually represents how well the items cover the range of abilities or traits within the target population (Linacre, 1994) can help identify gaps or overlaps in the construct representation and assure the test comprehensively measures the intended domain.

Second, the substantive aspect emphasizes the theoretical and empirical evidence supporting the processes by which test responses are generated and aligned with the intended construct (Messick, 1989). This aspect examines whether the cognitive and behavioral mechanisms elicited by the test items reflect the construct being measured (Wolfe & Smith, 2007). Person fit statistics (infit and outfit) are used to detect misfitting items or responses. In addition, item-person maps or wright maps verify whether the test items are appropriately targeted to the test-takers' abilities, ensuring alignment between cognitive demands and respondent performance (Bond & Fox, 2015; Wright & Masters, 1982).

Third, the structural aspect ensures that a test's internal structure aligns with the theoretical construct it aims to measure (Messick, 1995). This aspect examines whether the relationships among test items, scoring, and measurement models accurately represent the construct's dimensionality. The Rasch model's unidimensionality testing through Principal Component Analysis (PCA) of residuals can help identify whether items measure a single construct. Furthermore, item fit statistics (infit and outfit) are employed to reveal whether individual items align with the overall test structure. High reliability indices further support the structural integrity of the test (Bond & Fox, 2015; Wright & Masters, 1982).

Fourth, the external validity involves assessing how well test scores correlate with external criteria, such as real-world outcomes and behaviors, to ensure the test's practical utility. This aspect of validity as suggested by Wolfe and Smith (2007) emphasizes how the difficulty estimates are distributed in relation to the spread of ability estimates. This can be examined by the person separation index and wright maps (Bond & Fox, 2015; Wolfe & Smith, 2007).

Fifth, the generalizability emphasizes the extent to which test results are consistent and applicable across different groups, times, and contexts. This aspect of validity can be examined by assessing the invariance of item and person parameters across diverse groups and contexts, detecting Differential Item Functioning (DIF), and cross-validating test results across populations and time. These methods help ensure that a test is valid not only in a specific setting but also in a broader range of contexts, supporting the test's applicability and fairness (Bond & Fox, 2015; Messick, 1989).

Last but not least, the consequential validity refers to the intended and unintended consequences of test use, including the impact of test results on decision-making, behavior, and societal outcomes. The Rasch model can be employed to assess this aspect by detecting Differential Item Functioning (DIF), and ensuring that the test aligns with ethical and educational goals. Item fit statistics and Wright maps can also help identify unintended consequences, such as bias or discrimination, that may arise from the test's use, supporting the ethical and fair application of assessments (Bond & Fox, 2015; Messick, 1989).

Previous studies using Rasch model and Messick's validation framework

Recently, many research studies have been carried out on language test validation using the Rasch model in relation to the six aspects of Messick's framework. For example, Baghaei and Amrahi (2011) investigated a multiple-choice vocabulary test employing Messick's framework. Their results indicated that some test items did not align with the Rasch model. Distractor analysis revealed that a few options were not working as intended and should have been removed. The Wright map showed that the test was well-suited to the target group and covered a broad range of abilities. Additionally, performance on two subsets sections showed participants had consistent scores, supporting the test's unidimensionality. Similarly, Ravand and Fioozi (2016) examined the construct validity framework of the Iranian University Entrance Exam (UEE) for applicants into MA program. Their findings showed that while some indices supported aspects of the validity framework, others did not. Their findings indicated issues such as lack of unidimensionality, misfit among examinees, and low reliability in various

sections of the test, raising concerns about the test's validity. More recently, Noroozi and Karami (2024) explored the validity of the University of Tehran English Language Proficiency test (UTEPT). Their results revealed that the structural, generalizability, and substantive aspects held. However, the evidence for the context, external, and consequential aspects of validity was less definitive or conclusive. Their findings also provided insights into the test's measurement properties, including item difficulty and person ability estimates, contributing to the understanding of the test's validity and reliability.

Despite the growing body of research on test validation using the Rasch model and Messick's six aspects of validity, significant gaps still remain. Challenges such as lack of definitive evidence, misfitting items, multidimensionality, and poor targeting of test takers continue to warrant further investigation. Moreover, although Rasch analysis has been applied in Thailand across fields such as health assessments and performance evaluations (Dhakal et al., 2023; Kangwanrattanakul, 2025; Patarapichayatham et al., 2009; Thepsathit & Tangdhanakanond, 2024), few studies have explicitly combined Rasch validation with Messick's framework. This highlights the need for research that applies both approaches in the Thai language testing context. In particular, no study has yet examined the validity of an ESP test such as TETET. Addressing these gaps, the present study aims to provide comprehensive validity evidence that can enhance the reliability, fairness, and effectiveness of high-stakes language assessments.

METHODOLOGY

Participants

The participants of this study involved 179 fourth-year undergraduate students who took the 2021 version of TETET. They were 80 male and 99 female students from different faculties at a science and technology university in Bangkok, Thailand. They consisted of 51 students from faculty of Science, 42 students from faculty of Industrial Education and Technology, 36 students from Engineering faculty, 15 students from school of Information Technology, and 35 students from Institute of Field Robotics and School of Architecture and Design. There is no information regarding the participants' personal information or their scores in this research study.

Instrument

TETET is the key instrument of this study. TETET is a computer-based test which was developed to assess participants' four English communicative skills (i.e., reading, listening, writing and speaking) in the workplace context, especially in engineering and technological sectors. Test takers need to perform various test tasks such as matching, multiple choice, drag and drop, dictation, email composition, or answer questions verbally as voice-recordings, with 79 items in 12 sections within 2.5 hours. For this study, only the reading section was examined for validation. There are 32 items across four subsections in this section which include Survival Reading, Reading from the Internet, Reading Technical Manuals, and Email Reading as described in Table 1 below (see Appendix A for examples), following the design by Jaturapitakkul and Watson Todd (2018). All items are rewarded with one mark each.

Table 1

TETET reading section information

Section	Item test type	Aim	Description and objectives
Survival Reading (items 1–5)	Multiple-choice	To measure basic reading abilities	With five items in five minutes, the candidate needs to identify the right information due to the situation in each item based on a set of drop-down information given (e.g., from the list of telephone directory, the candidate needs to choose who to call in a particular situation.)
Reading from the Internet (items 6–15)	Copy and paste information	To measure the abilities of finding specific information and evaluating the information	There are two situations: five items for each. From the information and situation given, the candidate needs to read and comprehend the situation and then find the corresponding answers by going through offline webpages and evaluating the information. Copying and pasting information are what each candidate is required to perform. Candidate has 15 minutes to finish this section.
Reading Technical Manuals (items 16–30)	Part 1: Matching pictures with instructions Part 2: Matching the objects with the instructions	To measure the understanding and interacting with the instructions provided	It consists of two parts. For the first part with six items, the candidates need to read the instructions of setting up or installing some equipment and relate it with the given pictures representing each instruction. In the second part, there are 9 items. The candidate needs to read and correctly follow the safety regulations by clicking on the objects against the rules stated in the test for three items. For those objects which correspond to the regulations stated or not related to the instructions, the candidate must leave them as they are for 6 items. This section needs to be done within 15 minutes.
Reading Email (items 31–35)	Multiple-choice	To measure reading for main ideas, details, and inference	The candidate is required to read and comprehend the email and answer five multiple-choice questions regarding finding the main idea, details, and inferences within 10 minutes.

Within these four reading sections in this test, each section was designed to assess test-takers' reading abilities through a variety of test tasks. Each section imposed different cognitive demands on test-takers, depending on the nature of the tasks. In each section, a range of difficulty levels was included; however, there was no fixed distribution of these levels. Instead, the expectation was that the items ranged from easy to difficult along the Wright map. In terms of dimensionality, the test was designed to distinguish between ESP reading items and general reading items. There was an expectation that the ESP items would be harder overall, but no hypothesis was formed based on cognitive difficulty.

Data collection

After the participants had taken this computer-based test at the university's test center, the data from all 179 participants were prepared and organized to be analyzed in WINSTEPS computer program version 5.4.1 (Linacre, 2022). Each section required a particular coding schema as shown and clarified in Appendix B.

Data analysis

To validate the test and address different aspects of validity covering test score meaning, score relevance and utility, value implications, and social consequences (Shaw & Crisp, 2011), Messick's

(1995) six-aspect framework was adopted as a guiding model. Evidence was examined for content, substantive, structural, external, generalizability, and consequential validity. Rasch analysis was the primary method, supplemented by expert reviews for content and substantive aspects. While the Rasch model offers robust psychometric evidence (Baghaei & Amrahi, 2011), it cannot by itself fully capture all six aspects. Therefore, results for external, generalizability, and consequential validity are reported as preliminary, with future work recommended to incorporate qualitative data for stronger claims. Data analysis was conducted using the WINSTEPS Rasch Software version 5.4.1. The Rasch model features and considerations applied in this study are illustrated in Table 2.

Table 2

Rasch analysis for each aspect of validity (Chong et al., 2022)

Validity aspects	Rasch analysis
Content validity	Point-measure correlations (PT-measure), Item fit, Wright map, Item separation index
Substantive	Person fit, Wright map
Structural	Dimensionality analysis
External	Wright map, Person separation index
Generalizability	Differential item functioning (DIF)
Consequential	Item fit, Differential item functioning (DIF), Wright map

RESULTS

In some studies, each section of a test was analyzed separately due to their differing levels of difficulty and specific assumptions associated with each section. However, in this study, all four reading sections were analyzed as a whole set. This was executed to disclose how individual items across sections function compared to other sections. Results are presented aspect by aspect according to Messick's validity framework (1995) while the Rasch analysis, complemented by expert reviews, provides strong evidence for content, substantive, and structural validity. However, findings related to external, generalizability, and consequential validity are reported as preliminary, since these aspects cannot be fully addressed through Rasch statistics alone.

1. Content aspect

The content aspect refers to the specification of the test in terms of structure of the construct domain included in the test (Messick, 1995; Noroozi & Karami, 2024). The content aspect of validity requires evidence for content relevance, representativeness, and technical quality (Messick, 1995).

1.1 Content relevance

The aim of this analysis was to examine whether the test content was relevant with the test specifications. Based on the results of the panel discussions of ten experts who were stakeholders of English tests and TETET, the experts had no objection on the conflict of the test content. Moreover, the correlations between the reading test and other skill sections were

high (Jaturapitakkul & Watson Todd, 2018; Maneekhao et al., 2006). Consequently, it can be summarized that the content of the reading test fits the designers' expectations.

1.2 Representativeness

Representativeness was in regards to the coverage of the items in the test if they were covering all domain content as expected. The anticipation of this test was that the items should vary sufficiently in difficulty to represent test-takers' abilities across a range of proficiency levels, from low to high. This was examined visually through the Wright map, and by examining the item strata and separation scores. In order to present representativeness of the test, the test-takers' responses and items' difficulty must be placed on the same scale, which is called the Wright map (Boone, 2016). A Wright map can be used for checking whether (1) the test set contains an adequate number of items, (2) the empirical item hierarchy shows sufficient spread, and (3) there are gaps in the item difficulty hierarchy and spread (Ha, 2021).

Wright map

In Figure 1, the Wright map illustrates a two-sided diagram in which the items were arranged by their difficulties on the right side whereas the test-takers' abilities were displayed on the left side. The top area of the map represents higher level of item difficulty and test-taker ability. The average of item difficulty and test-takers' performance are centered at 0 on the scale. The majority of the takers' performances are clustered in the middle of the map. The distribution of the test items across all four sections ranged appropriately from easy to difficult even though some gaps among groups of items were observed.

With regard to the number of items and the intention of the test design mentioned above, each section was expected to contain items in a range of difficulty levels. However, the internet reading – situation 1 section was functioning differently. All five items in this section were bunched at the same difficulty level which contrasted with the test design's expectation. This issue was further investigated to determine whether the items should be revised or eliminated. The remaining reading sections performed as planned and aligned with the test specification.

Item strata and separation index

The indices of both strata and separation should be greater than 2 so as to illustrate different levels of item difficulties and person abilities (Linacre, nd; Linacre, 2005). Item separation index can be inspected as indicator to distinguish groups of items and to illustrate representativeness of the test (Noroozi & Karami, 2024). In the current study, the item separation index was 5.79, indicating around six groups of items could be distinguished through this analysis. This is good because there were four reading subset sections and two levels of cognitive processes, so the item separation matches the designers' intention. The person separation index was 1.74 with a strata value of 2.65 which indicates the range of learners' abilities were large enough to cover the item difficulty range. However, precise could be improved as person reliability was 0.77, which indicated that the range of students' ability was a little narrow. As seen in Figure 1 above, the range of item difficulty and person ability were close to +/-3 logits from the mean, so the test the item difficulties and person abilities are evidence of the test's representativeness.

1.3 Technical quality

Technical quality concerns the test quality based on the statistical results of fit statistics and Point-measure correlations (Pt-measure). The fit value for both infit and outfit criteria was with the range of 0.5 to 1.5 and the Pt-measure should be positive. For those items that did not fit the criteria, they were interpreted unusual items (Linacre, 2002; Noroozi & Karami, 2024). In this study, there was only one unusual item—the survival item no. 4 whose outfit (2.18) exceeded the acceptable range for fit statistics and Pt-measure was negative. High value of outfit indicated underfit, meaning that some test-takers might have employed special knowledge, guessing, or other techniques to correctly complete the item while their ability indicated otherwise. As the item did not work as intended, test designers need to revise or eliminate the item. The rest of the items were found to function appropriately in terms of representing takers' real performance and abilities.

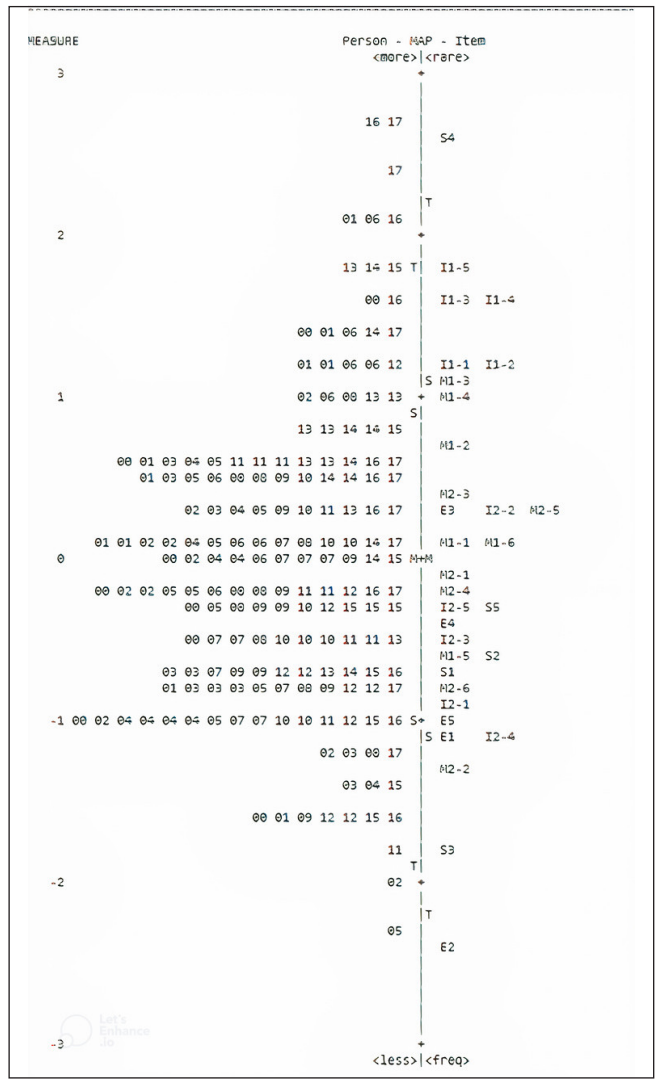


Figure 1 Wright map of item difficulty and person ability

Note. I1 = Internet reading section 1; I2 = Internet reading section 2; E1 = Email reading section 1; E2 = Email reading section 2; S1 = Survival reading section 1; S2 = Survival reading section 2; M1 = Manual reading section 1; M2 = Manual reading section 2.

TABLE 26.1 Reading_test2.xlsx ZOU171WS.TXT Aug 15 2024 15:35
INPUT: 179 Person 32 Item REPORTED: 179 Person 32 Item 2 CATS WINSTEPS 5.4.1.0

Person: REAL SEP.: 1.83 REL.: .77 ... Item: REAL SEP.: 5.79 REL.: .97

Item STATISTICS: CORRELATION ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
4	17	179	2.55	.27	1.23	1.16	2.18	2.94	-.03	.28	89.4	90.6	S4
31	102	179	-.38	.16	1.28	4.53	1.39	4.12	.08	.38	53.6	66.8	E4
26	77	179	.28	.16	1.23	3.65	1.26	2.95	.16	.39	53.6	67.2	M2-5
5	99	179	-.30	.16	1.18	3.13	1.34	3.71	.17	.38	58.7	66.4	S5
23	133	179	-1.28	.18	1.15	1.66	1.17	1.12	.17	.33	73.7	75.5	M2-2
28	128	179	-1.12	.18	1.14	1.71	1.26	1.79	.18	.34	69.8	73.3	E1
24	74	179	.36	.16	1.16	2.53	1.27	2.99	.21	.39	59.8	67.8	M2-3
27	119	179	-.85	.17	1.13	1.84	1.16	1.38	.22	.35	66.5	70.1	M2-6
25	97	179	-.25	.16	1.13	2.27	1.16	1.85	.25	.38	60.9	66.3	M2-4
17	61	179	.73	.17	1.09	1.25	1.19	1.83	.28	.38	67.6	71.2	M1-2
22	90	179	-.06	.16	1.09	1.55	1.18	2.18	.28	.38	63.7	66.1	M2-1
30	78	179	.25	.16	1.08	1.31	1.14	1.71	.30	.39	64.2	67.1	E3
3	146	179	-1.75	.20	.98	-.12	.85	-.70	.33	.29	81.0	81.8	S3
29	159	179	-2.39	.25	.91	-.50	.63	-1.35	.37	.24	88.8	88.8	E2
12	75	179	.34	.16	1.00	.07	1.04	.53	.38	.39	65.9	67.6	I2-2
18	50	179	1.06	.18	.96	-.42	1.00	.00	.40	.37	77.1	75.3	M1-3
20	111	179	-.62	.17	.95	-.73	.99	-.12	.40	.37	70.9	68.1	M1-5
21	83	179	.12	.16	.95	-.83	1.01	.15	.42	.39	72.6	66.4	M1-6
32	125	179	-1.02	.17	.92	-1.09	.84	-1.30	.44	.34	72.6	72.1	E5
19	52	179	1.00	.18	.93	-.78	.90	-.78	.44	.38	76.0	74.5	M1-4
7	47	179	1.16	.18	.92	-.86	.85	-1.10	.46	.37	78.2	76.4	I1-2
10	30	179	1.82	.21	.84	-1.20	.79	-1.06	.48	.34	86.6	84.0	I1-5
16	85	179	.07	.16	.91	-1.72	.87	-1.76	.49	.39	73.7	66.3	M1-1
11	122	179	-.93	.17	.87	-1.95	.80	-1.80	.49	.35	78.2	71.0	I2-1
9	34	179	1.65	.20	.86	-1.18	.74	-1.50	.49	.35	84.9	82.2	I1-4
6	46	179	1.20	.18	.88	-1.33	.82	-1.31	.49	.37	78.8	76.9	I1-1
14	128	179	-1.12	.18	.84	-2.17	.75	-2.02	.51	.34	77.7	73.3	I2-4
8	35	179	1.61	.20	.83	-1.47	.72	-1.68	.52	.35	85.5	81.7	I1-3
15	100	179	-.33	.16	.86	-2.66	.80	-2.52	.53	.38	73.7	66.5	I2-5
1	114	179	-.71	.17	.84	-2.72	.79	-2.20	.53	.36	75.4	68.8	S1
2	110	179	-.59	.17	.83	-3.06	.80	-2.21	.54	.37	77.1	68.0	S2
13	107	179	-.51	.16	.83	-3.13	.77	-2.67	.55	.37	77.7	67.4	I2-3
MEAN	88.6	179.0	.00	.18	.99	-.04	1.01	.10			72.9	72.7	
P.SD	36.0	.0	1.10	.02	.14	1.99	.29	1.98			9.4	6.8	

Figure 2 Fit statistic and Pt-measure results, arranged by correlation order

2. Substantive aspect

The substantive aspect of validity evaluates whether the test items perform as expected based on theoretical rationales in terms of response patterns and item hierarchy (Wolfe & Smith, 2007). In this study, these four reading sections are distinctive in terms of the test tasks and reading skills which require different cognitive levels to complete each test task. As shown in Figure 1, the distribution of items in the Wright Map looks satisfactory; there are items to assess the takers' reading abilities ranged from easy to difficult. Even though there was no explicit cut-off point, the continuum of the test difficulty could be observed. Items within the same section were generally functioning the same, except for the item no. 4 in the reading survival part. As mentioned previously, it suggests that test-takers might have needed to apply different skills to complete this item, so item revision or deletion needs to be considered.

3. Structural aspect

Structural aspect refers to the investigation of dimensionality of the test. It can be studied through the unidimensionality analysis by examining the principal components analysis of the residuals (PCAR). As shown in Figure 4, the raw variance explained by measures, the percentage of the functioning items observed, was 35.5%. To determine whether the test includes multiple

dimensions, in other words, testing different skills, the unexpected variance in first contrast should exceed 2 eigenvalues (Raiche, 2005). From the analysis in this study, the first contrast was 3.95 eigenvalues or 11.1% of the total, indicating that approximately four items were functioning distinctively compared to the rest of the items.

For cluster determination, three clusters were observed as presented in Figure 4. After examining the clusters and disattenuated correlation closely, Figure 4 shows that the content in clusters 2 and 3 was likely part of the same category as its correlation was 0.77, whereas the relationship between clusters 1 and 3 was weak, 0.30. The items in cluster 1 were from the section of internet reading situation 1. This finding can be inferred that this test section requires further investigation as suggested previously. For ESP test designers, this point is critical to distinguish reading content unique to the topic, but not so unique as to encourage other compensatory skills such as guessing to respond correctly.

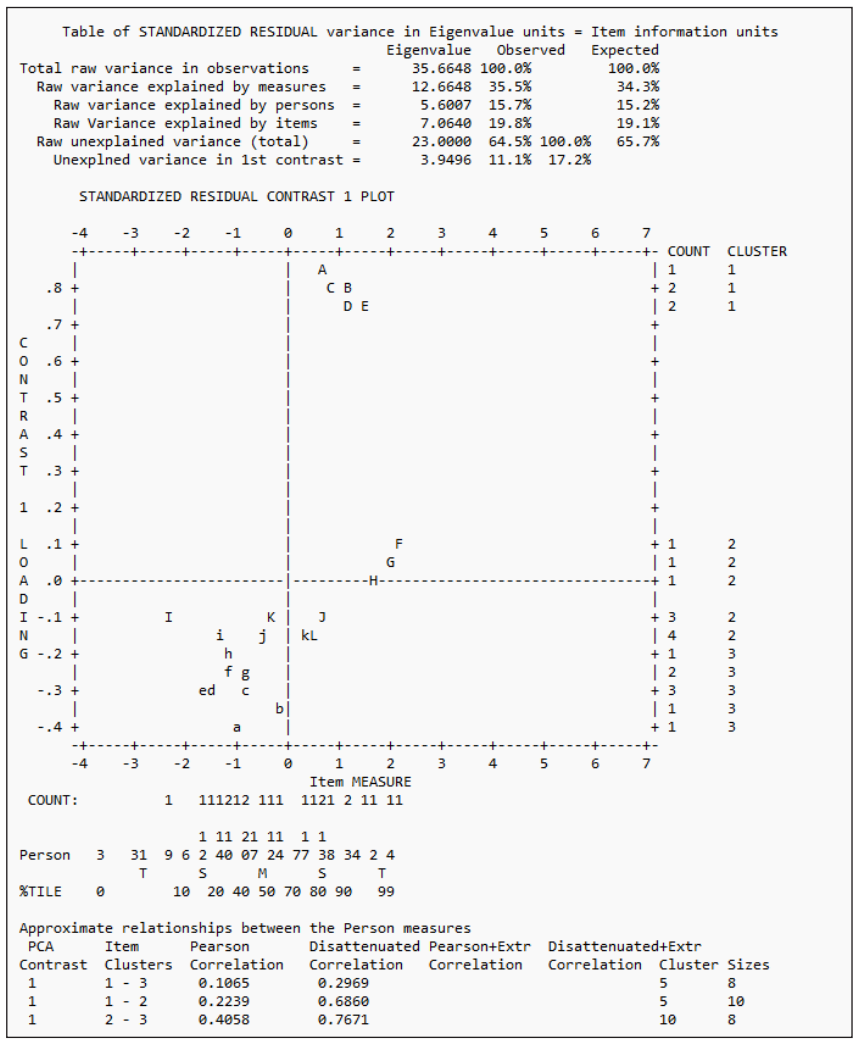


Figure 3 Dimensionality analysis

Note. A, B, C… and a, b, c … = Individual Winsteps coded test items.

4. External validity

The external aspect of construct validity examines whether the test effectively differentiates individual takers with varying ability levels. The observation of the spread of items in relation to person measure dispersion on the wright map and the analysis of person separation index can be done to address this aspect of construct validity (Chong et al., 2022). Considering the Wright map in Figure 1, there is no critical gap in the continuum, indicating that majority of the items were appropriately targeted for participants across all ability levels.

According to Bond and Fox (2015), the person strata index should be utilized when the distribution is ranged from very high to very low. On the other hand, the person separation index can be observed when the sample is large and normally distributed. For this study, the test takers' ability was ranged from high to low; thus, person strata index was selected. The person strata index was 2.65, indicating there were three main groups of test takers found based on the performance. Therefore, the external aspect of validity of this test is able to distinguish groups of test takers.

5. Generalizability aspect

The generalizability aspect regards to the consistency of item invariance calibrations across contexts and groups of test-takers (Wolfe & Smith, 2007). This aspect can be addressed by differential item functioning (DIF). DIF index was determined based on reference criteria suggested by Noroozi and Karami (2024): the value of 0.41 or lower means slightly different; the value from 0.42 to 0.63 means slightly to moderately different; the value exceeding 0.63 means moderate to large difference among those two groups. Additionally, Rasch-Welch probability ($p < 0.05$) was examined to determine whether the observed differences were statistically significant.

In this study, DIF was analyzed across genders and females were coded as 1 and males as 0. From the DIF index as illustrated in Table 2, nine items showed slight to moderate differences, but none of these were statistically significant. Whereas considering the *p*-value of the email reading item 4 whose DIF contrast was - 0.72 representing moderate to large difference, and its *p*-value of 0.03 indicating statistically significant difference in performance between males and females test-takers. The finding indicated that this item favored males answering correctly. This finding is a bit odd as the item focused on the vocabulary term *bid*. Nonetheless, consideration must be noted with further test-takers to determine whether term *bid* is indeed systematically biased in favor of males. In general, the analysis of the DIF contrast indicated that there was no significant difference, outside the usage of *bid*, in test fairness among test-takers' gender. The analysis supports the generalizability of the test.

Table 3

Differential item functioning (DIF) results

No	Item	DIF contrast	*t*	Rasch-Welch prob
1	Survival item 1	.15	.43	.66
2	Survival item 3	-.46	-1.12	.26
3	Internet reading – situation 1 – item 4	-.59	-1.43	.15
4	Internet reading – situation 2 – item 1	.37	1.03	.30
5	Technical manual – part 1 – item 1	.62	1.86	.06
6	Technical manual – part 1 – item 3	.35	.97	.33
7	Technical manual – part 2 – item 2	.20	.54	.58
8	Technical manual – part 2 – item 6	.51	1.44	.15
9	Email reading – item 3	.09	.27	.78
10	Email reading – item 4*	- .72	-2.19	.03

*$p < 0.05$

6. Consequential validity

The consequential aspect of validity refers to the intended and unintended consequences of test use, including its impact on test-takers, institutions, and other stakeholders. In the present study, this aspect can only be addressed indirectly. While Rasch-based analyses such as DIF and Wright maps provide preliminary evidence related to fairness and score distribution, they cannot fully capture the broader social and educational consequences of test use. A more comprehensive evaluation of consequential validity would require qualitative data, such as interviews or surveys with test-takers and decision-makers, to examine how test results are interpreted and what outcomes they produce in real contexts (Chapelle & Voss, 2021). This process requires a more longitudinal approach. For this study, however, a more limited approach was taken. The authors were key participants in gathering insights from the different stakeholders. Through the evaluative process, the authors had direct communication with the test item writers, university executives promoting the implementation of TETET, and test proctors. Although anecdotal, these brief encounters allowed the authors insight on how the test was viewed. In addition, general feedback from students also allowed the authors to feel confident the implementation of TETET was proceeding smoothly and positively.

Within the limits of the current data, the Rasch analyses suggest that the TETET reading section functions fairly across male and female test-takers, and that its items generally align with the intended design. These findings imply that the test is a reliable tool for classifying test-takers' performance. However, with limited stakeholder perspectives, evidence for consequential validity remains partial and should be strengthened in future studies.

Overall, while the Rasch analyses provided substantial evidence for content, substantive, and structural validity, the findings for external, generalizability, and consequential aspects should be regarded as preliminary, highlighting the need for additional qualitative and cross-contextual evidence that will be further discussed in the following sections.

DISCUSSION AND CONCLUSION

While the current study demonstrates how Rasch analysis can be linked to all six aspects of Messick's framework, the validity evidence presented is not equally strong across them but possibly contributes to cumulative validity arguments.

The study results demonstrate a satisfying quality of TETET on validity aspect. With respect to promoting the test quality especially on the issues of the score meaning, utility, implication and consequences, certain questionable items should be reviewed for potential improvement or removal (Shaw & Crisp, 2011). For instance, the survival reading item 4 whose outfit was not aligned with the criteria and presented a misfit result. Misfit index shows uncertain results of test-takers on that item, the takers might have relied on external knowledge, guessing or other techniques to complete the test (Bond & Fox, 2015; Linacre, 2002). Additionally, the internet reading situation 1 section which was functioning differently comparing to other sections especially the internet reading situation 2 whose test characteristics were similar. According to Bond and Fox (2015), tests sharing similarities or testing the same content should cluster together in the dimensionality diagram, with high disattenuated correlation results. The internet reading situation 1 section also needed further consideration for areas of enhancement. Apart from revisiting the test quality, the content should be reviewed to ensure that they are in the intended direction, specifically in the ESP (English for Specific Purposes) dimension.

In addition to revisiting the questionable items to enhance their quality, repeated validation should occur to promote the solidarity of test's validation process, particularly when applied to new groups of test-takers using the same test set. Although the revalidation is likely to consume significant resources, it ensures the quality in different aspects especially the bias and fairness of the test, which might affect the test-takers' utilization and their stakeholders (Kane, 2006). Furthermore, different versions of reading test should also be comparatively studied and validated in order to assure their test characteristics and functioning consistency.

Revalidation is not only available way to ensuring test quality, test designers also play an important role. They all must have thorough understanding of how each item and section functions so as to be an essential input for the future designs or improvements. When all test producers have complete understanding in the characteristics of the test, particularly items sharing similar dimensions such as internet reading situation 1 and 2, the test quality control can be held (Bond & Fox, 2015). With regard to these attempts and the study of Jaturapitakkul and Watson Todd (2018) which standardized the whole four-skill set of TETET and CEFR, TETET can be promoted as a high-stakes assessment. Additionally, the understanding and clarification on how the test in regard to the real-world contexts has been clearly established.

While the current study draws on Messick's (1989, 1995) validation framework, the validity evidence presented relies primarily on psychometric results from Rasch analysis. As highlighted by Dhakal et al. (2024), situational authenticity—particularly the authenticity of test inputs, tasks, and response processes—represents a distinct and critical form of validity evidence. Although situational authenticity overlaps with content validity, it goes further by examining

how closely test performance reflects workplace-relevant language use. This dimension of validity cannot be fully addressed by statistical evidence alone and thus remains beyond the scope of the present study.

To promote TETET as a high-stakes test, the test should also be applicable to a broader range of test-takers which can be investigated through the Wright map and DIF statistics (Noroozi & Karami, 2024; Wolfe & Smith, 2007). The Wright map indicated acceptable ranges of item difficulty and person ability. The DIF analysis in this study suggested that this test had no favor to any gender. However, to expand TETET's market, more DIF studies should be conducted with other groups of test-takers whose contexts and backgrounds are similar to Thai test-takers such as Asian or ASEAN users. Moreover, comparative studies with test-takers who share less context with Thai context should be done. The further studies can be done to ensure whether the test is effective with and fair to other groups of users in different contexts in order to endorse the test' quality and become a high-stakes one (Bachman & Palmer, 2010; Chapelle et al., 2010).

LIMITATIONS AND FUTURE DIRECTIONS

This study investigated the validity of the current version of the reading section of TETET using Rasch model in relation to Messick's validation framework. Even though the overall validation was satisfactory, the findings were subject to some limitations.

First, there were only 179 test-takers. As the intention for TETET to be a high-stakes test, more test-takers would highlight the current findings. Nonetheless, the number of test-takers was balanced and gave good insight as to the validity of the test.

Second, the attempt to combine ESP and general reading items together highlighted the difficulty of designing a test that maintained construct validity. Using Messick's framework affords us confidence in the validation, one needs to maintain skepticism going forward as there needs to be repeated validation and comparative studies across diverse test-taker groups.

Third, while the study suggests that TETET demonstrates validity and fairness, its applicability to a broader range of users, particularly those outside the Thai or ASEAN context, remains uncertain. Therefore, more studies involving participants from other countries like Cambodia, Vietnam, and Japan should be further investigated.

Fourth, the reliance on DIF analysis to assess bias is valuable, but additional studies involving test-takers from varied linguistic and cultural backgrounds are necessary to ensure the test's fairness and effectiveness in different contexts.

Fifth, the inclusion of qualitative evidence to complement the Rasch-based findings is suggested. While psychometric analysis provides strong support for content, substantive, and structural aspects of validity, other aspects—particularly consequential validity—require perspectives from test-takers, teachers, and institutional stakeholders. Qualitative methods such as

interviews, focus groups, or surveys that capture a more longitudinal perspective could shed light on how TETET results are used in practice, how they influence decision-making, and whether they produce intended or unintended consequences. Integrating such evidence with the present psychometric findings would allow for a more comprehensive validity argument and align with current recommendations in validation research (Chapelle & Voss, 2021).

Sixth, the analysis of the situational authenticity of the TETET reading tasks is absent from this study. Future research should therefore investigate how well test tasks replicate the linguistic and cognitive demands of real engineering and technological workplaces. Incorporating situational validity evidence alongside psychometric analyses would offer a more comprehensive account of TETET's validity and strengthen its claim as a domain-specific, high-stakes assessment.

Finally, the study acknowledges the resource-intensive nature of revalidation, requiring significant time, financial investment, and expertise. Institutions can adopt a phased validation approach to spread costs over time, and form collaborative partnerships to share resources. Employing targeted sampling methods and continuous data collection can also help streamline validation efforts, while seeking government or institutional funding can help secure financial support. These strategies ensure effective revalidation while minimizing costs and logistical challenges.

THE AUTHORS

Natjiree Jaturapitakkul, Ph.D. is an assistant professor at the School of Liberal Arts of King Mongkut's University of Technology Thonburi in Bangkok, Thailand where she teaches English to undergraduate and graduate students in the ELT program. She has published and presented papers on English language teaching and learning, language assessment, test development and ESP testing.
natjiree.jat@kmutt.ac.th

Paul Horness, Ph.D. is presently a professor at Soka University in Tokyo, Japan teaching in the TESOL Graduate Program. His research interest mainly includes assessment and study abroad. For assessment, he has examined listening skills and the spacing effect. As for study abroad, he mainly focuses on short-term programs (less than 4 weeks) and what effects it has on language development, cultural knowledge, and personal growth.
paul@soka-u.jp

Duangjaichanok Pansa, is a teacher at the School of Liberal Arts of King Mongkut's University of Technology Thonburi in Bangkok, Thailand where she obtained her Master's degree in Applied Linguistics (ELT). Her research interests include language testing and assessment, English language teaching and learning, integrating technology in language learning and teaching, and corpus linguistics.
duangjaichanok.pan@kmutt.ac.th

REFERENCES

Alderson, J. C., & Wall, D. (1993). Does language testing work? A study of the validity of language tests. *Language Testing, 10*(1), 1–24.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA & NCME). (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/open-access-files.html

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple-choice items. *Journal of Language Teaching and Research, 2*(5), 1052–1060.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the Human Sciences* (3rd ed.). Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education, 15*(4), 1–7. https://doi.org/10.1187/cbe.16-04-0148

Chapelle, C. A. (2012). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). *Building a validity argument for the test of English as a foreign language.* Routledge.

Chapelle, C. A., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research.* Cambridge University Press.

Chong, J., Mokshein, S. E., & Mustapha, R. (2022). Instrument validation based on the six aspects of Messick validity framework using the Rasch rating scale model. *Jurnal Penyelidikan Pendidikan, 23*(1), 117–138.

Dhakal, K. R., Watson Todd, R., & Jaturapitakkul, N. (2023). Unpacking the nature of critical thinking for educational purposes. *Educational Research and Evaluation, 28*(4-6), 130–151.

Dhakal, K. R., Watson Todd, R., & Jaturapitakkul, N. (2024). Input as a key element in test design: A narrative of designing an innovative critical thinking assessment. *rEFLections, 31*(2), 516–542.

Engelhard Jr., G., & Wind, S. A. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments.* Routledge.

ETS (2019). *Understanding TOEIC test scores*. https://www.etsglobal.org/fr/en/content/understanding-toeic-tests-scores

ETS (2023). *TOEIC 2023 report on test takers worldwide*. https://www.ets.org/pdfs/toeic/toeic-listening-reading-report-test-takers-worldwide.pdf

Ha, H. T. (2021). A Rasch-based validation of the Vietnamese version of the listening vocabulary levels test. *Language Testing in Asia, 11*, Article 16. https://doi.org/10.1186/s40468-021-00132-7

Jaturapitakkul, N. (2007), *The effects of language ability and engineering background knowledge on ESP reading ability of Thai graduate students, their test taking strategies and attitudes towards the test* [Unpublished doctoral dissertation]. Chulalongkorn University.

Jaturapitakkul, N., & Watson Todd, R. (2018). *TETET technical report*. https://sola.kmutt.ac.th/tetet/doc/TETET_Technical_report_NJ_RT_Complete.pdf

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kangwanrattanakul, K. (2025). Validation of the Thai World Health Organization quality of life-OLD (WHOQOL-OLD) among Thai older adults: Rasch analysis. *Scientific Reports, 15*, Article 12978. https://doi.org/10.1038/s41598-025-97824-4

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development* (2nd ed.). Routledge.

Linacre, J. (n.d.). *Help for facets Rasch measurement and Rasch analysis software.* Winsteps. www.winsteps.com

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

Linacre, J. (2002) What do infit, outfit, mean-square, and standardization mean? *Archives of Rasch Measurement, 16*, 871–882.

Linacre, J. M. (2005). *A user's guide to Winsteps/Ministeps Rasch model programs.* MESA Press.

Linacre, J. M. (2022). *Winsteps® Rasch measurement computer program user's guide.* Version 5.2.3. Winsteps.

Maneekhao, K., Jaturapitakkul, N., Watson Todd, R., & Tepsuriwong, S. (2006). Developing an innovative computer-based test. *Prospect-Adelaide, 21*(2), 34–46.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Noroozi, S., & Karami, H. (2024). A Rasch-based validation of the University of Tehran English Proficiency Test (UTEPT). *Language Testing in Asia, 14*(2024), 1–18.

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2009). Cross-level two-way differential item functioning analysis by multilevel Rasch modeling. *Journal of Research Methodology & Cognitive Science, 7*(1), 15–21.

Raiche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*(1), Article 1012.

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the Master's UEE using the Rasch model and the six aspects of Messick's framework. *International Journal of Language Testing, 6*(1), 1–23.

Shaw, S., & Crisp, V. (2011). Tracing the evolution of validity in educational measurement: Past issues and contemporary challenges. *Assuring in Assessment, 11*, 14–19.

Shohamy, E. G., Or, I. G., & May, S. (Eds.). (2017). *Language testing and assessment* (3rd ed.). Springer. https://doi.org/10.1007/978-3-319-02261-1

Thepsathit, P., & Tangdhanakanond, K. (2024). The development of formative assessment rubrics for enhancing students' performance on Thai percussion instruments. *International Journal of Music Education, 42*(4), 674–690. https://doi.org/10.1177/02557614231192189

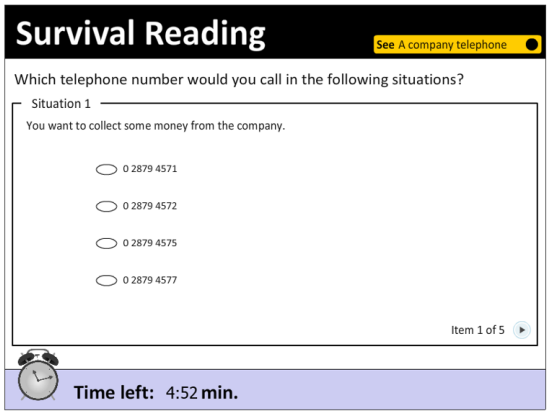
Weir, C. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models. *Journal of Applied Measurement, 8*(3), 297–318.

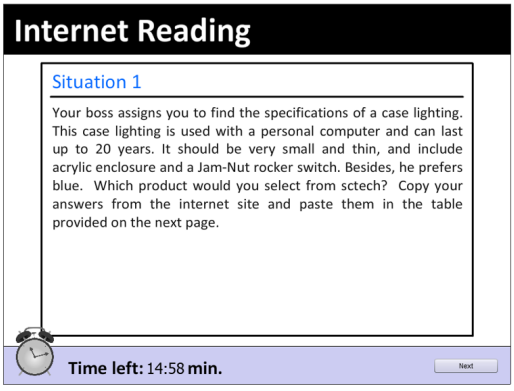
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.

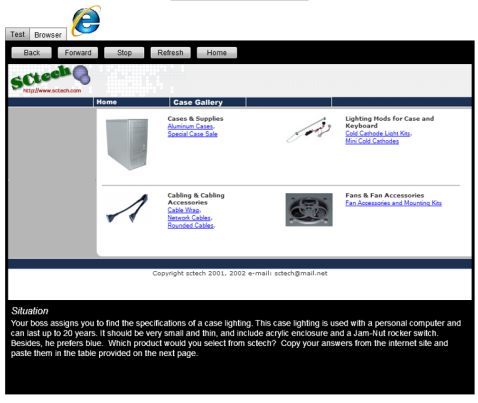
Appendix A
Sample of the reading section of Test of English for Thai Engineers and Technologists (TETET)

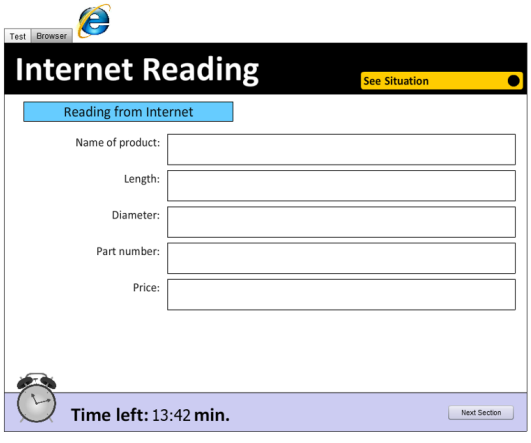
Survival reading section



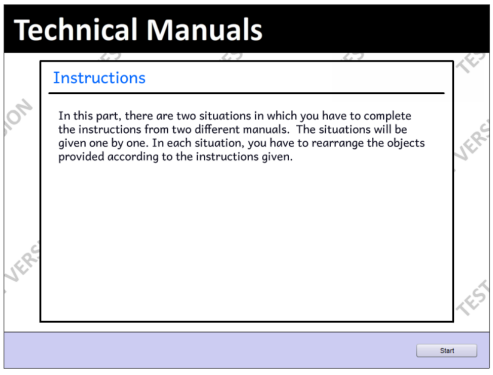
Reading from the internet section

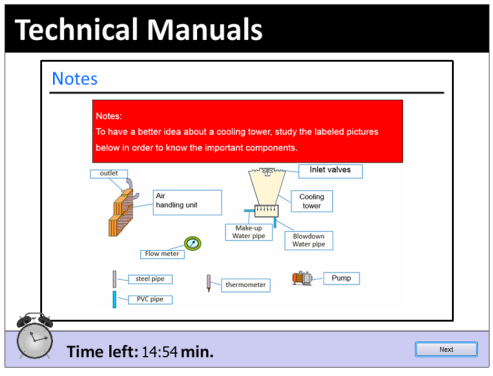


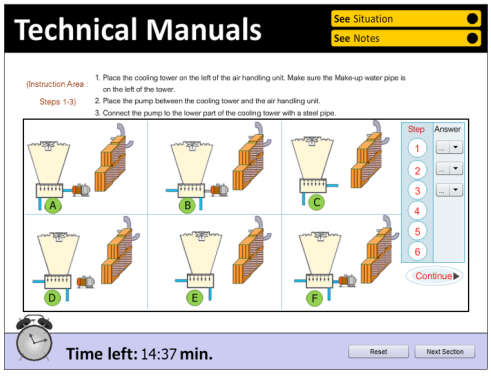


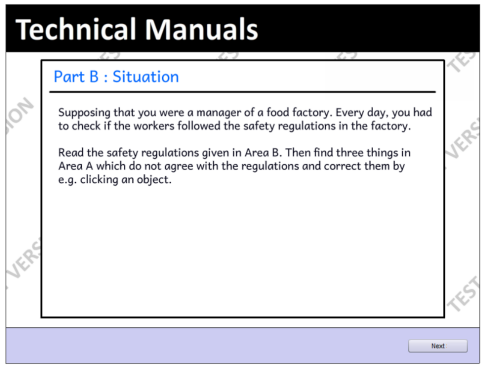


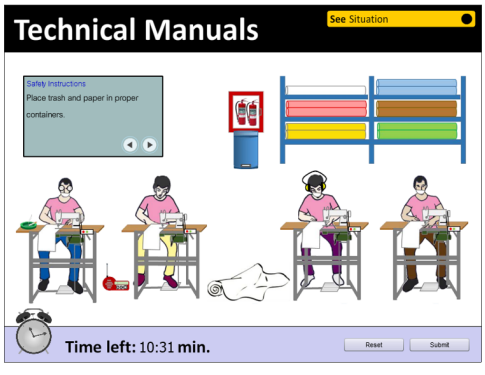
Reading technical manuals section



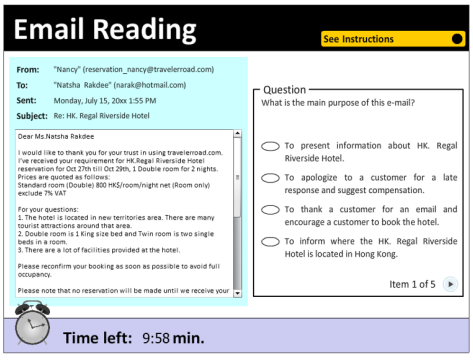








Reading email section



Appendix B
Coding scheme for each reading section

Item	Section	Item test type	Coding schema
1-5	Survival Reading	Multiple-choice	Pattern: ABCDX In each item, the participant's answer was coded as A, B, C and D as their answers.
6-15	Reading from the Internet	Copy and paste the correct information	Pattern: TFX T (True) was provided to the correct answer and F (False) was given an incorrect one.
16-21	Reading Technical Manuals	Multiple-choice (Match the instructions with the pictures).	*Part 1* Patterns: abcdefghijklmnopqrX The answer was coded as a, b, c, d, e, f, g, h, I, j, k, l, m, n, o, p, q or r due to their answer in each item.
22-30		Identify by clicking on the objects match the instructions	*Part 2* Pattern: TFX The participants' actions in the test were detected and converted as True (T) for the correct answer or False (F) answers for the incorrect one.
31-35	Reading Email	Multiple-choice	Pattern: ABCDX In each item, the participant's answer was coded as A, B, C and D as their answers.

Note: X in the coding refers to 'no answer' from the candidate in that item.