REID (Research and Evaluation in Education), 10(1), 2024, 18-34
Available online at: http://journal.uny.ac.id/index.php/reid

A psychometric evaluation of an item bank for an English reading comprehension tool using Rasch analysis

Louis Wai Keung Yim[1]*; Che Yee Lye[2]; Poh Wee Koh[1]
[1]Singapore Examinations and Assessment Board, Singapore
[2]Singapore University of Social Sciences, Singapore
*Corresponding Author. E-mail: lwky100@cantab.net

ARTICLE INFO

Article History
Submitted:
22 August 2023
Revised:
21 February 2024
Accepted:
27 March 2024

Keywords
Rasch analysis; construct validity; differential item functioning; local item dependence; English reading comprehension

Scan Me:

ABSTRACT

This study reports the psychometric evaluation of an item bank for an Assessment for Learning (AfL) tool to assess primary school students' reading comprehension skills. A pool of 46 primary 1 to 6 reading passages and their accompanying 522 multiple choice and short answer items were developed based on the Progress in International Reading Literacy Study (PIRLS) assessment framework. They were field-tested at 27 schools in Singapore involving 9834 students aged between 7 and 13. Four main comprehension processes outlined in PIRLS were assessed: focusing on and retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information, and evaluating and critiquing content and textual elements. Rasch analysis was employed to examine students' item response patterns for (1) model and item fit; (2) differential item functioning (DIF) about gender and test platform used; (3) local item dependence (LID) within and amongst reading passages; and (4) distractor issues about options within the multiple-choice-type items. Results showed that the data adequately fit the unidimensional Rasch model across all test levels with good internal consistency. Psychometric issues found amongst items were primarily related to ill-functioning distractors and local dependence on items. Problematic items identified were reviewed and subsequently amended by a panel of assessment professionals for future recalibration. This psychometrically and theoretically sound item bank is envisaged to be valuable to developing comprehensive classroom AfL tools that provide information for the English reading comprehension instructional design in the Singaporean context.

This is an open access article under the **CC-BY-SA** license.

To cite this article (in APA style):
Yim, L., Lye, C., & Koh, P. (2024). A psychometric evaluation of an item bank for an English reading comprehension tool using Rasch analysis. *REID (Research and Evaluation in Education), 10*(1), 18-34 https://doi.org/10.21831/reid.v10i1.65284

INTRODUCTION

The importance of reading comprehension in primary schools has been a topic of discussion for many years in the educational fraternity. The impact of reading comprehension on early childhood development is so significant that it cannot be overstated. Despite much research on the topic of reading over the years, many children around the world are still struggling to read (OECD, 2016). The detrimental impact of reading difficulties for these children reaches far beyond the academic realm in which studies have linked poor literacy skills to social and psychological maladjustment among school-aged children (Korhonen et al., 2014; Livingstone et al., 2018; Mugnaini et al., 2009).

Increasingly, researchers and practitioners alike advocate reading comprehension as a core measure/indicator of reading literacy for two main reasons. Firstly, the fundamental purpose of reading is to decipher and make meaning of texts (RAND Reading Study Group, 2002) and the ability to comprehend translates into content mastery across different subjects (Smith et al.,

2021). Secondly, silent reading has overtaken oral reading as the predominant mode of reading activity in modern classrooms. Since the turn of the century, augmented discussions in reading comprehension (Oakley, 2011; RAND Reading Study Group, 2002) have also sparked an interest in the development of assessments that facilitate the measurement of this skill (Pearson & Hamm, 2005). In particular, Hwang and Wu (2014) suggested that formative assessments provide crucial opportunities for teachers to follow up on the data collected which could enhance students' learning.

The emphasis on reading comprehension assessment inevitably raises the question of validity and reliability associated with such assessments. This question is particularly relevant in the context of Singapore, where culturally relevant assessment instruments are scarce in reading comprehension. With a lack of readily accessible assessment instruments, teachers most likely will turn to self-prepared assessments and grade every student's comprehension ability via an overall numerical mark at the end of the assessment. This practice might be inadequate in measuring students' comprehension levels, missing the opportunity to identify the area(s) where students are genuinely weak for further remedial actions.

In this study, we report the psychometric evaluation of an item bank (or repository) for an Assessment for Learning (AfL) tool developed to assess English reading comprehension for Singapore primary students. We also examine the psychometric properties of items within these reading comprehension passages, which spanned across six test levels, and address the issues relating to validity and reliability of these assessment instruments.

The Singapore English Reading Comprehension Tool

The English reading comprehension tool developed by the Singapore Examinations and Assessment Board (SEAB) is to check on students' English reading comprehension proficiency at the end of key stages for primary school students. The tool focuses on assessing different reading comprehension skills based on the Progress in International Reading Literacy Study (PIRLS) assessment framework as well as providing qualitative feedback on students' comprehension proficiency so that teachers can identify the area(s) where they are weak at and target those areas to form remedial instructions/plans for students in an attempt to close the learning gaps. For readers' reference, PIRLS is an international assessment and research project designed to measure reading achievement at the fourth-grade level, as well as school and teacher practices related to instruction (Mullis & Martin, 2021).

Assessment Framework

Processes of Comprehension	Purposes for Reading	
	Literary Experience	Acquire and Use Information
Focus on and Retrieve Explicitly Stated Information		
Make Straightforward Inferences		
Interpret and Integrate Ideas and Information		
Evaluate and Critique Content and Textual Elements		

Figure 1. The PIRLS Reading Purposes and Comprehension Processes

Figure 1 shows the PIRLS 2021 assessment framework illustrating the two overarching purposes for reading that account for most of the reading covered by young students both in and out of schools: (1) literary experience and (2) acquiring and using information. In addition, the PIRLS assessment integrates four broad-based comprehension processes within each of the two purposes for reading: (1) focus on and retrieve explicitly stated information, (2) make straightforward inferences, (3) interpret and integrate ideas and information, and (4) evaluate and critique content and textual elements. It should be acknowledged that the purposes for reading and the processes of comprehension do not function in isolation from one another or from the context in which students live and learn (Mullis & Martin, 2021). SEAB has already acquired copyright approval from the International Association for the Evaluation of Educational Achievement (IEA) for the adoption of the PIRLS 2021 Reading Assessment Framework, i.e., *PIRLS Reading Purposes* and *Comprehension Processes* and their associated description paragraphs, as part of the AfL tool's framework of assessment.

On top of the PIRLS 2021 assessment framework, the tool also makes reference to the Singapore English Language Syllabus (Ministry of Education of Singapore, 2020) for primary schools to ascertain four learning outcomes (LOs) subsumed under the three focus areas tailor-made to the local context are fulfilled: (1) reading and viewing closely; (2) reading and viewing critically; and (3) reading and viewing widely and extensively for different purposes. Figure 2 shows the four learning outcomes (LOs) under the three focus areas for ELS 2020 for primary students. Figure 3 shows how the PIRLS reading assessment framework is dovetailed with the ELS 2020 learning outcomes. Apart from LO1, which is not assessed directly in primary schools, LOs 2, 3, and 4 map to the PIRLS framework's purposes of reading and processes of comprehension in an orderly manner.

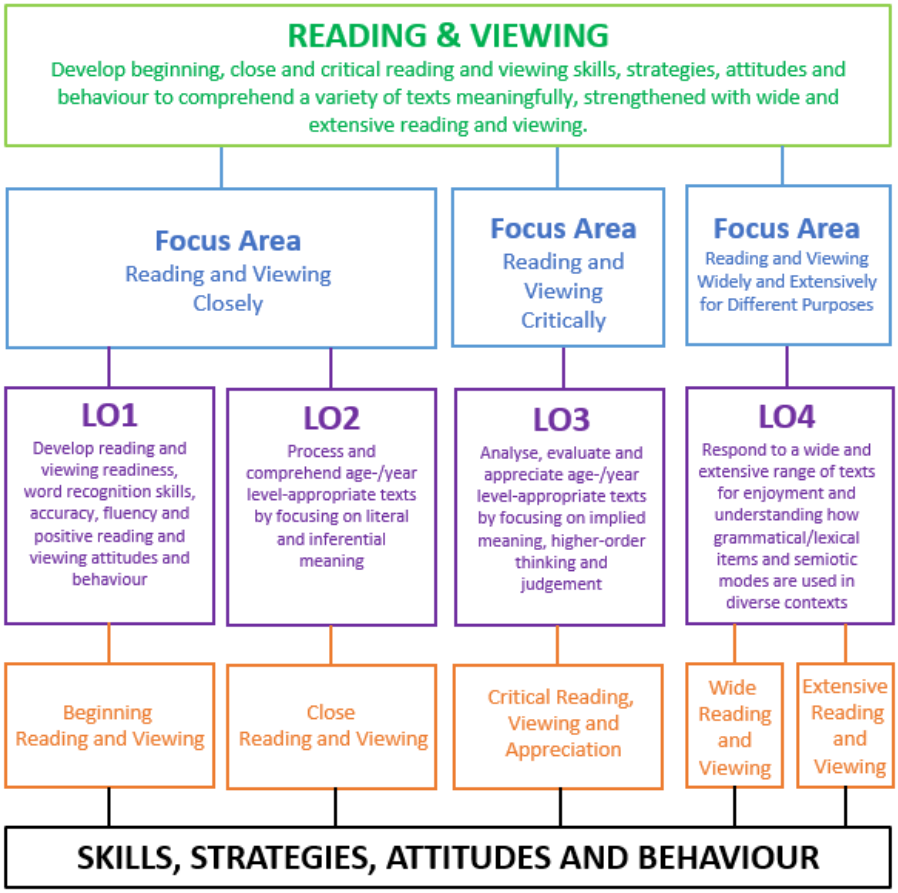

Figure 2. ELS 2020's Four Learning Outcomes for Primary Reading Comprehension (Ministry of Education of Singapore, 2020)

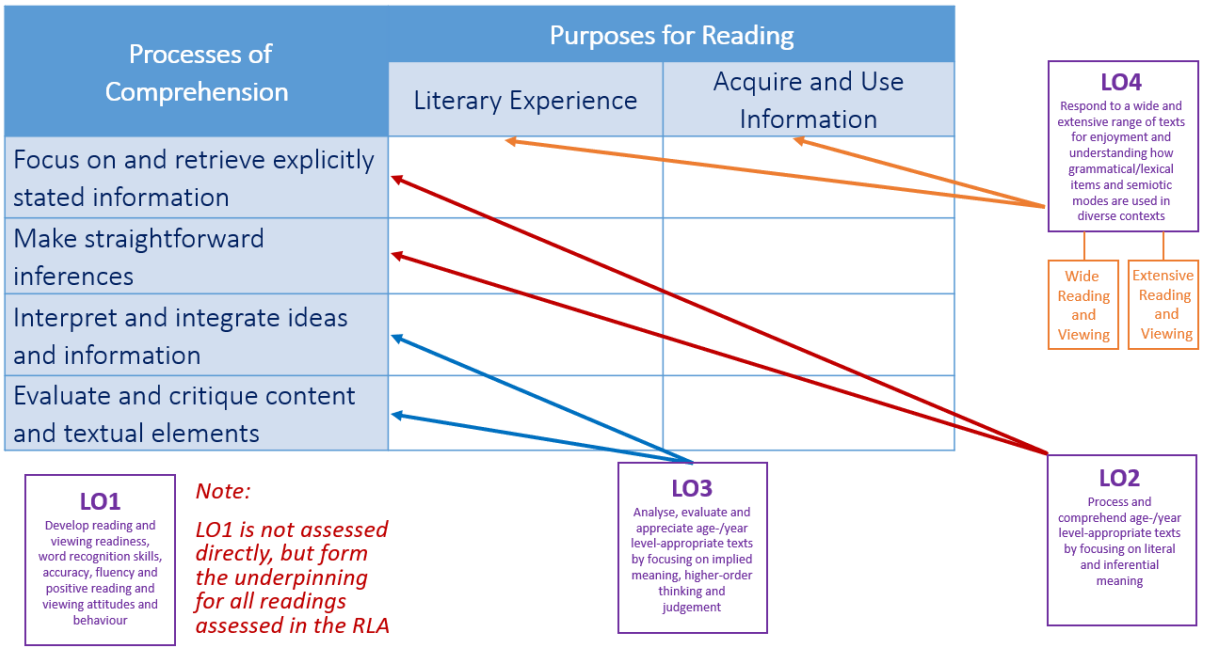

Figure 3. Mapping Between the PIRLS 2021 Assessment Framework and ELS 2020 Learning Outcomes

The assessment framework also covers the passage criteria at different test levels (from primary 1 to 6) such as word limit, item types, number of items, and test duration per text. These are put together by a group of reading professionals which comprises local experts and an international consultant after much consideration and deliberation.

Building An Item Bank

A main thrust of the AfL tool is to build a bank that comprises a repository of passages (or texts) and their associated items encompassing all primary test levels to assess students across a broad range of abilities and grade levels. The bank should, therefore, have a range of texts fulfilling the assessment framework criteria put together by the group of reading experts. The distribution of texts from primary 1 to 6 in the repository at its preliminary stage is listed in Table 2. More passages and items at different levels, especially those at key stages, i.e., primary 2, 4, and 6, will be added to the bank as the tool evolves.

A schematic diagram illustrating the process of building the item bank is shown in Figure 4. Setters are provided with passage and item exemplars at each level as references to help develop passages and items adhering to the assessment framework's criteria from scratch. These exemplars are created and endorsed by a group of reading experts to act as a set of guidelines for setters. Upon completing a first draft of a passage/item, the setter submits the materials for SEAB's moderation to which feedback will be given for his/her refinement. After a few iterations, the polished version is submitted to an Expert Panel for evaluation in which criteria of the assessment framework are scrutinised. The Expert Panel comprised English reading comprehension experts for primary schools from the Ministry of Education (MOE), SEAB, and an external consultant. Passages that require no further interventions will be added to the passage/item calibration pool pending to be field-tested at schools. The rest of the passages/items will be further edited/amended before being reunited in the passage/item calibration pool.

Before the administration of field tests at schools, passages (or tests) are assembled into different test forms for different student groups to sit the tests systematically. These are to comply with the test-linking design and Rasch sampling criteria, i.e., the recommended number of students to be field-tested per item (Hagell & Westergren, 2016) to acquire reliable calibration results. A sampling frame is established to select suitable schools to participate in the field tests.

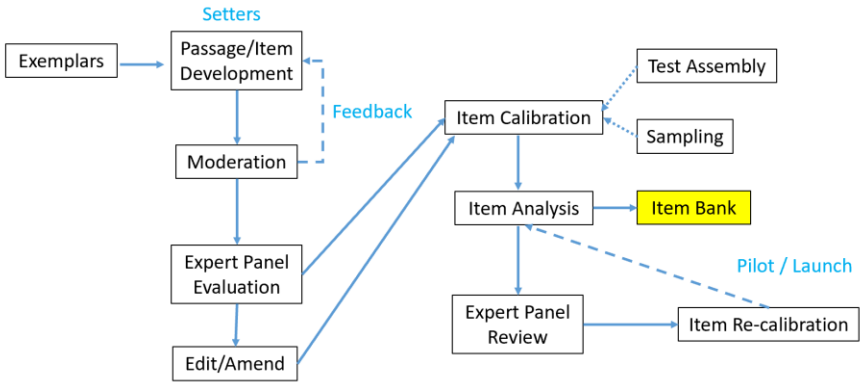

Figure 4. A Schematic Diagram Illustrating the Process of Building a Passage/Item Bank

The criteria of the sampling frame and the method of test assembly are discussed in detail in the Method section. Twenty-seven primary schools involving 9834 students aged between 7 and 13 years took part in the field tests. Upon the collection of empirical data, Rasch item analysis is carried out across the six test levels where the difficulties of items are placed on a continuum scale. Psychometric traits of individual items are scrutinised to ascertain the quality of each item. Items that do not function optimally are flagged for the Expert Panel to review. Once changes have been made, these items will be re-calibrated in the next round of field tests. Passages/items that have passed the expert panel's evaluation are deposited in the item bank for use in the assembly and administration of tests.

The Rasch Model

There has been a growing interest in and acknowledgment of the need to evaluate language assessment items using the Rasch model (Fan & Bond, 2019). The Rasch model is based on the probabilistic model that:

> … a person having a greater ability than another person should have a greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person, the probability of solving the second item is the greater one. (Rasch, 1960, p. 117)

The Rasch model encompasses a set of rigorously prescribed conditions in which the assessment items should meet to be considered productive measurement. These conditions include unidimensionality and local independence (Bond & Fox, 2015). The principle of unidimensionality in Rasch requires the assessment to measure a single underlying measurement dimension or construct at one time, and examples of dimensions in language assessments are reading ability or writing ability. As Rasch uses interval scaling which is concerned with a psychometric (empirical) construct rather than a psychological (conceptual) construct, it is possible to examine the unidimensionality of language assessment data (Fan & Bond, 2019). Local independence requires assessment items to be set in such a way that an examinee's response to an item is not affected by his/her responses to any other items in the same test. This is especially critical in a reading comprehension assessment since items are subsumed under the same reading passage, and hence, the chances of items within the same passage violating the local independence's principle become more likely. Local independence and unidimensionality are relative concepts. If the principle of local independence is violated, likely, unidimensionality is also violated (Fan & Bond, 2019).

In general, the Rasch model can facilitate the development and maintenance of an item bank. It provides interval scale measures that describe a continuum of increasing ability on the construct (i.e., item difficulty and person ability), thus allowing the linking of items or scales for item banking.

Construct Validity and the Rasch Model

The six aspects of Messick (1995) unified model of construct validity have been linked to Rasch-based analyses to provide evidence of validity for tests (e.g., Bond, 2003; Ravand & Firoozi, 2016). Messick (1995) views construct validity as a unitary concept with six distinguishable aspects, namely, content, substantive, structural, generalisability, external, and consequential. Essentially, construct validity is an overall judgment of the degree to which empirical evidence and theoretical rationales support the appropriateness, trustworthiness, and usefulness of the inferences made from test scores. The six aspects in Messick's construct validity are elaborated as follows, which also describe how each aspect links to the Rasch analyses' output.

Content

The content aspect, in general, includes content relevance, representativeness, and technical quality. It addresses the question of whether the test items appear to be measuring the designated construct. Item fit statistics are often used to provide evidence of the content aspect of construct validity. They highlight misfitting items that might represent different constructs. The person-item map can be used to examine gaps and redundancies, as well as the mismatch of the mean between items versus persons along the central tendency of the distributions.

Substantive

The substantive aspect addresses the consistency of the construct, i.e., whether the theoretical foundation underlying the construct is substantial. A commonly used example is the multiple-choice distractor analysis such as examining the proportion of examinees choosing each distractor and the mean ability of examinees for each distractor. One of the other examples includes checking the person separation reliability in establishing test consistency.

Structural

The structural aspect examines to what extent the internal structure is consistent with the domain construct. In Rasch analysis, the dimensionality test via factor analysis is usually conducted to determine the structural validity of the test items.

Generalisability

The generalisability aspect expects test score properties and interpretations to be generalised across demographics, time, and place. Examples of Rasch analyses include differential item functioning (DIF) and local item dependency (LID). The DIF test provides evidence of measurement invariance, i.e., the degree to which the items measure the same dimension across demographics, time, and place. The LID test provides evidence that examinees' responses to items are not affected by or dependent on their responses from other items in the same (or cognate) test.

External

While internal validity is addressed by the substantive and structural aspects of construct validity, external validity examines the extent to which test scores are related to other test and non-test behavior. External validity can be examined via the person-item map to check whether the test is well-targeted for the sample.

Consequential

The consequential aspect concentrates on the implications of score interpretation as a basis for action, including potential risks if scores are invalid or inappropriately interpreted. The consequential validity of a test depends on the extent of misfitting and invariance and can be established by examining the extent of item misfits and DIF. Table 1 outlines Messick's analyses'

framework conducted to establish construct validity of the reading comprehension items in the AfL tool presented here.

Table 1. Item Analyses using the Rasch Model in Association with Messick's Unified Model of Construct Validity Framework

Validity	Analysis	Examinations and Interpretations
Content	• Person-item map • Item fit	No significant gaps and redundancies along the line as acceptable. ± 2.5 indicates an adequate fit to the model; a non-significant chi-square as acceptable.
Substantive	• Distractor analysis	The proportion of respondents choosing each distractor; average ability measures of respondents choosing each distractor; high distractor measure correction as acceptable.
	• Person separation reliability	Equivalent to Cronbach's alpha, a value close to 1 indicates high reliability and model fit.
Structural	• Dimensionality	Factor analysis was conducted to evaluate Principal Component Analysis (PCA) loadings and then paired t-tests were conducted using the positively and negatively loaded items; unidimensionality is present when the percentage of significant t-test is less than 5%.
	• Local item independence (LID)	Correlation of item residuals of less than 0.30 as no evidence of LID.
Generalisability	• Differential item functioning (DIF)	Non-significant item residuals with between-groups analysis of variance (ANOVA) as no evidence of DIF.
External	• Person-item map	The distribution of personal measures and items is widely dispersed as acceptable.
Consequential	• Item fit • DIF	The consequence of the test depends on the degree of Rasch model fit and fairness for the different groups of test takers; no evidence of misfit and DIF is acceptable.

Research Questions of the Study

Messick's model of construct validity incorporating the Rasch analyses has provided a solid evaluation framework to assess the validity and reliability of an assessment instrument and forms the basis for our research questions. In particular, the following questions pertaining to the evaluation and calibration of the item bank for the tool are addressed.

(1) Do the sets of tests and their items conform to the fit item analysis requirements based on the Rasch model?

(a) item fit;
(b) unidimensionality;
(c) local item dependency (LID);
(d) differential item functioning (DIF).

(2) Do all distractors in the items show adequate evidence of consistency?

(3) To what extent are the test linking processes able to produce a calibrated item bank that meets the requirement of the item's parallel condition?

METHOD

A decile sampling approach based on each school's average score of a nationwide standardised English Language examination was employed to select participants from a multitude of Singapore primary schools. The calculated average scores of all primary schools were rank-ordered from high to low on a scale that was subsequently divided into ten equal strata based on their rank positions. Primary schools were randomly selected from each stratum to represent the wide range of schools' English Language abilities across Singapore, and individual participants were nominated by each school to take part in the study. As a result, a total of 9834 students (5524 boys and 4310 girls) aged between 7 and 13 years from schools participated in several field

tests across six test levels. To maintain confidentiality, the participants involved in this study were anonymised, and the data obtained were used only for research purposes. This also guarantees that their involvement does not influence their academic performance assessment in the future.

The reading comprehension assessment instrument was designed to measure primary school students' English proficiency. Each test comprised two reading comprehension passages with a total of 8 to 30 questions primarily of selected response type, i.e., 4 to 15 questions per passage, depending on the test level. Multiple-choice questions were the only question type for Primary 1 and Primary 2; whereas fill-in-the-blank and sequence-ordering questions were introduced from Primary Three to Primary Six on top of the multiple-choice questions.

A Step-by-step Approach to Building the Item Bank for Psychometric Evaluation

The following stages outline a step-by-step procedure for establishing an item bank for the assessment instrument.

Stage 1: Framework and Item Development

Table 2 shows the Item Specification Table for the six test levels of the English Reading Comprehension tool. The tool was designed to align with the PIRLS assessment framework concerning the two Purposes for reading and four Processes of Comprehension. Appropriate passages and items were developed by setters; and went through iterations of refinement by reading professionals. A total of 522 question items were administered at schools during the early stage of building the passage/item bank.

Table 2. Item Specification Table of the AfL from Primary 1 to Primary 6, Tabulating the Number of Passages for Two Purposes of Reading and the Number of Items for Four Processes of Comprehension at Each Test Level

Test Level	Purposes for Reading (Total no. of Passages = 46)	Comprehension Processes			
		Focusing and Retrieving Explicitly Stated Information (No. of Items)	Making Straightforward Inferences (No. of Items)	Interpreting and Integrating Ideas and Information (No. of Items)	Evaluating and Critiquing Content and Textual Elements (No. of Items)
P1	Literary (6)	18	6	-	-
P2	Literary (10)	45	25	-	-
P3	Literary (3)	16	14	7	3
	Informational (3)	19	12	4	7
P4	Literary (4)	16	19	17	5
	Informational (4)	19	19	12	6
P5	Literary (4)	14	18	16	7
	Informational (4)	18	17	12	12
P6	Literary (4)	14	20	18	7
	Informational (4)	17	18	12	10

Stage 2: Test Equating and Linking Design

As all reading comprehension items for different test levels would be put on the same continuum scale at the end of the calibration exercise, horizontal and vertical linking approaches were used to create multiple test forms with linking items prior to the test administration. Horizontal equating involves equating tests of different forms within a single test level, whilst vertical equating involves equating tests of different test levels. Another criterion factoring into the equating and linking design was to ascertain an optimal number of students who attempted each item within each passage to satisfy the analysis requirement of Rasch. As a result, 56 test forms were created involving 9834 primary school students in this study.

Table 3 shows an example of using the horizontal and vertical linking approaches to assemble different test forms. Three sets of passages and items nominated as *common passages/items*,

i.e., Sets A, B, and c, are assigned to both Primary 1 and Primary 2 levels repeatedly to form different test forms. Set B and Set c are linked horizontally within Primary 1 and Primary 2 respectively; whilst Set A is linked vertically between Primary 1 and Primary 2. It should be noted that no links are shown for Sets C, a, b, and d for specific test levels and test forms in Table 3.

Table 3. Link Design Illustration – A Visual Example of Test Forms and Common Items

Test Form	Set (Reading Passage and Items)	Set (Reading Passage and Items)	Test Level
1	A	B	P1
2	B	C	P1
3	A	a	P2
4	b	c	P2
5	c	d	P2

Stage 3: Testing of Items

Students from all sampled schools sat different test forms online for the computerised linear tests. The main purpose of conducting such tests was to collect students' raw responses so item analyses could be conducted to acquire item parameters, primarily item difficulty, for the passage/item bank.

Stage 4: Item Analysis

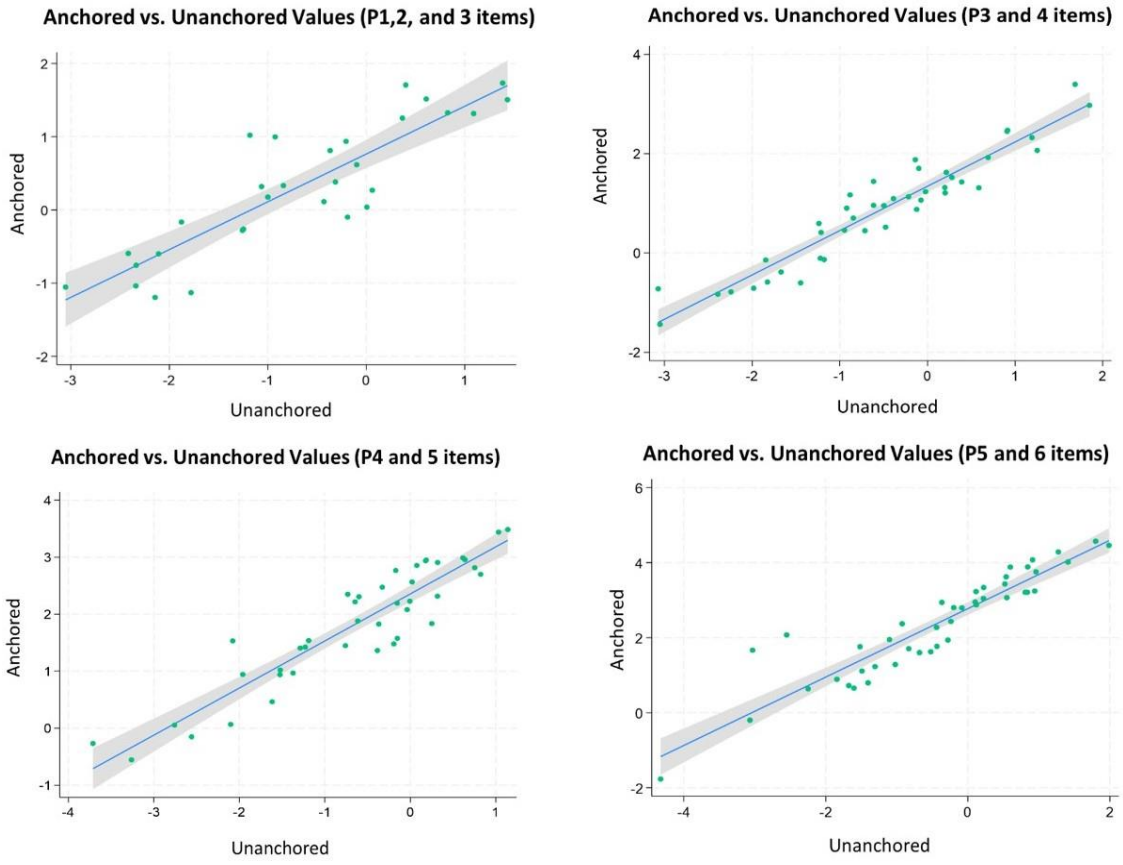

Figure 5. Correlations among Primary 1-6 Anchored and Unanchored Items: (i) P1, 2 and 3 Items, r = 0.88; (ii) P3 and 4 Items, r = 0.95; (iii) P4 and 5 Items, r = 0.92; iv) P5 and 6 Items, r = 0.92

Items were analysed based on the anchor test design, i.e., using item difficulty parameters of common items to estimate those of the unique items for different test levels. Figure 5 shows the association of item difficulty parameters between unique items and common items. The

correlation coefficients were 0.88 for P1 to P3 items, 0.95 for P3 to P4 items, 0.92 for P4 to P5 items, and 0.92 for P5 to P6 items, providing evidence of stability over the use of anchored items.

For the analyses, a score of 1 was awarded to students whose responses to an item were correct; and a score of 0 was awarded for incorrect responses. Students' raw responses were analysed using the Rasch model and the six aspects of Messick (1994) model of construct validity, namely, content, substantive, structural, generalisability, external, and consequential validity.

FINDINGS AND DISCUSSION

Findings

Item Fit

Item analyses were carried out using RUMM2030 Plus software (RUMM Laboratory Pty Ltd, 2012-2019) to examine how well the empirical data collected from field tests fit the Rasch model. The item fit was examined by means of its fit residuals (Z-scores) and item-trait interaction χ^2 statistics. Item fit residuals between $\pm$ 2.5 and non-significant χ^2 statistics (with Bonferroni adjusted level of significance) suggested that acceptable items fit the Rasch model (Pallant & Tennant, 2007). Although 113 out of 522 items across the six levels had fit residuals greater than $\pm$ 2.5, the χ^2 item-trait statistics associated with these items were not statistically significant, suggesting these items did not have significant misfit overall. Additional examinations on the standard error value of each item, individual item characteristic curves (ICC), and distractor analysis curves (also discussed later) were also carried out to identify misfit items.

Person Separation Index

The Person Separation Index (PSI) was also examined for all tests of each test level to determine the test sensitivity in differentiating between high and low performers. A PSI of 0.65 and greater implies acceptable reliability of the test in distinguishing between abilities. Table 4 shows that the PSIs for the Primary 3, 4, 5, and 6 tests are acceptable; whereas Primary 1 and Primary 2 were below 0.65, suggesting the latter two tests might not be as sensitive to distinguish between the high and low performers.

Table 4. Person Separation Indices Across All Test Levels

Test Level	Person Separation index
P1	0.45
P2	0.56
P3	0.74
P4	0.74
P5	0.72
P6	0.73

Person-item Map

The person-item map for the tests of each test level was also examined to determine the overlap between person and item threshold distributions. Figure 6 shows the person-item map for each test level from Primary 1 to Primary 6. The top half of each map shows the student ability distribution along the location axis (in logits); whereas the bottom half shows the item difficulty distribution along the same location axis. In general, there was less overlap between the two distributions at the higher ability levels across all six test levels, where there were few or no corresponding items that adequately measured these ability levels. This is because the tool primarily focused on identifying and subsequently closing the learning gaps of less able students, and hence, there is a tendency to have fewer difficult items at the tail end for higher ability students, i.e., in the region of 5.0 to 6.0 logits (unanchored).

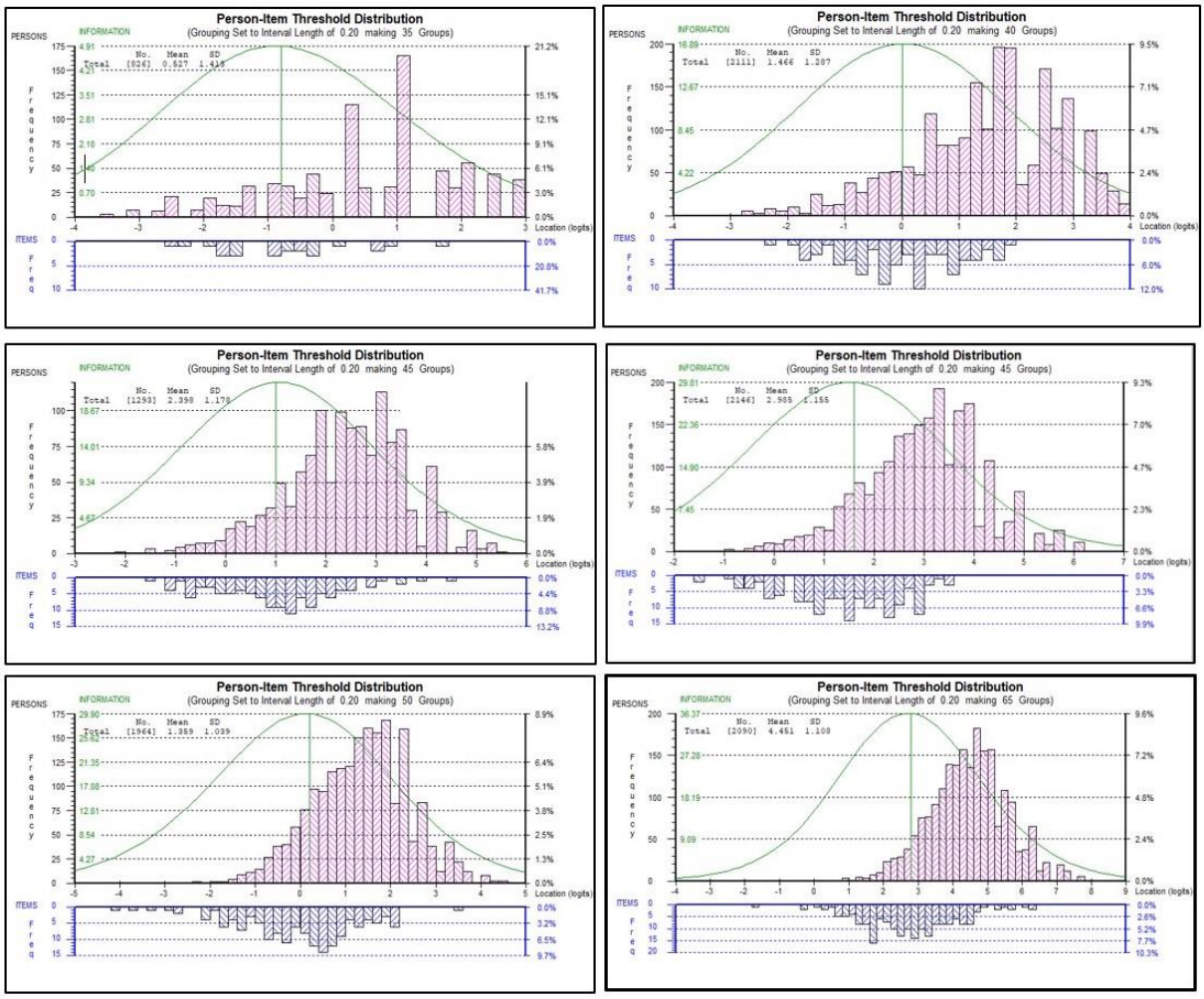

Figure 6. Person-Item Maps for Primary 1 (top-left), Primary 2 (top-right), Primary 3 (middle-left), Primary 4 (middle-right), Primary 5 (bottom-left), and Primary 6 (bottom-right) Tests

Differential Item Functioning

The analysis of DIF was performed on gender across all six test levels to examine if there were any items where male students performed better than females, and vice versa, based on the difficulty of items and the average ability of each gender group. Furthermore, DIF for the test platform was also conducted since different platforms were used to administer tests for Primary 2 and Primary 5 in an attempt to resolve some logistical issues. This was to examine if the response probabilities to items differ amongst students on different platforms despite similar abilities. Item residuals of between-group analyses of variance (ANOVAs) showed that all test items from each test level were free from DIF for gender or test platform. A further examination of the ICCs plotted for gender and test platform further reinforced the absence of DIF.

Local Item Dependence

As local item independence is an underlying assumption of the Rasch model, the response dependence amongst items for tests in each test level was also examined. Correlations of item residuals derived by means of principal component analyses (PCAs) were examined for response dependence. Correlation coefficients of less than 0.30 suggested that the assumption of local item independence is met (e.g., Christensen et al., 2017). In other words, an item response does not have a direct impact on other item responses (Cantó-Cerdán et al., 2021; Christensen et al., 2017). The analysis results showed that almost all items exhibited no signs of response dependence except for a few items in the Primary 4 test level. In the test, residual correlations of four

sequence-ordering items were found to be above 0.30, i.e., displaying evidence of item dependence, as these four items required students to order the storyline of a passage to test their overall passage comprehension, which, therefore, were closely cognate with one another in nature.

Distractor Analysis

Rasch model was applied to analyse distractors (or options) of the multiple-choice questions (MCQ) within each test. As long as the distractors are plausible options designed to understand students' misconceptions deviating from the answer key, item developers and teachers should be able to identify students' learning gaps to inform teaching and learning. The expected responses of a reasonably good MCQ item should follow a pattern similar to those shown in Figure 7, where the probability of higher-ability students choosing the distractors should be lower or much lower than that of the lower-ability students. For our distractor analyses, items with fit residuals greater than ± 2.5 would be flagged for further examination. This was followed by skimming through each MCQ item's ICC responses to surface any idiosyncratic behaviors displayed by distractors not functioning as expected. The distractor analysis results showed that three Primary 1 items, three Primary 3 items, four Primary 4 items, seven Primary 5 items, and two Primary 6 were problematic. These items shared a similar issue: the probability of one or more distractors within an item was/were higher than that of the answer key (see Figure 8a). Other issues were that the response patterns of higher and/or lower-ability students on the correct answer or distractors were not functioning as they should (see Figure 8b). These items were subsequently examined in conjunction with respective passages by the expert panel to determine if the items were to be discarded or edited for the next round of recalibration. Items not flagged for revision or had no signs of idiosyncratic ICC behavior were added to the item bank.

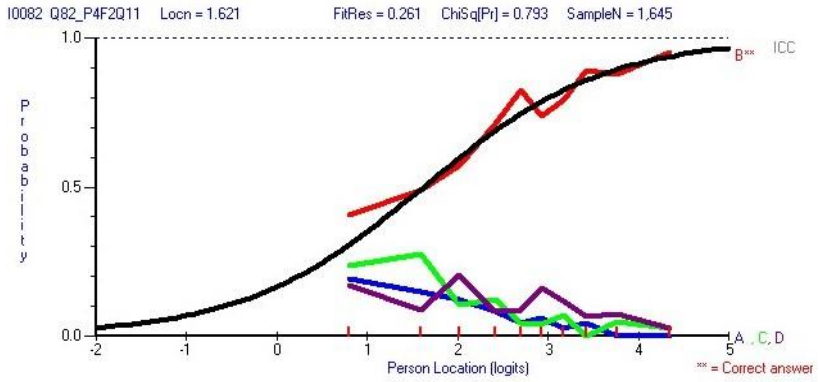

Figure 7. Item Characteristic Curve (ICC) of a Reasonably Good MCQ Item Responses

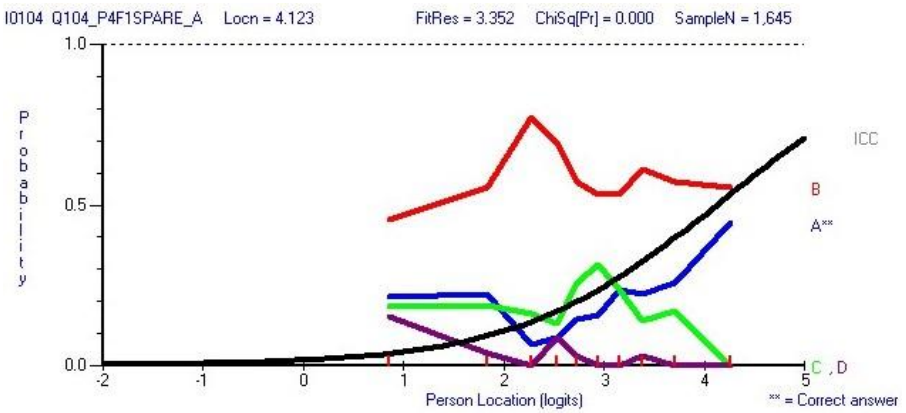

Figure 8a. Item Characteristic Curve (ICC) of a Problematic MCQ Item Response - the Probability of Distractor (B) is Higher than That of the Answer Key (A)

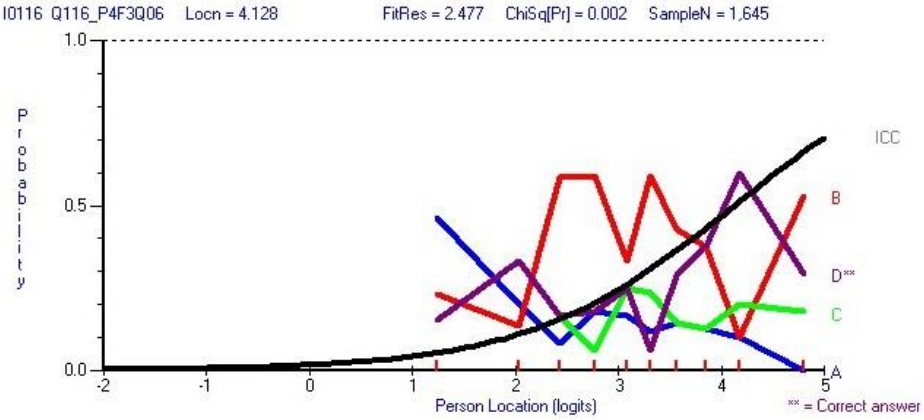

Figure 8b. Item Characteristic Curve (ICC) of a Problematic MCQ Item Response – Distractors and Answer Key are Muddled Up around the ICC

Rasch Analyses' Findings in Association with Messick's Unified Model of Construct Validity

An overview of the Rasch analyses' results concerning Messick's unified model of construct validity framework is shown in Table 5. Apart from a few items that did not function as they should and were subsequently reviewed by reading professionals for further recalibration, the rest of the items appeared to be of good quality for the item bank.

Table 5. The Item Bank Results of Rasch Analyses in Association with Messick's Model of Construct Validity

Validity	Analysis	Results of Rasch Analyses
Content	Person-item map	In general, the overlap between the person and item distributions was less at the higher ability ends across all levels. This is due to the AfL tool's design to focus primarily on identifying less able students and hence closing their learning gaps.
	Item fit	No items were identified as being significantly misfit after the examinations of fit residuals, χ^2 statistics, and standard error value.
Substantive	Distractor analysis	19 items across six test levels were identified as problematic. These items were further examined by a reading professional to determine if they should be discarded or edited for the next round of recalibration.
	Person separation reliability	The values for P3 to P5 were acceptable (>0.65); whereas those for P1 and P2 were below 0.65, suggesting the tests might not be as sensitive to distinguish between high and low performers.
Structural	Dimensionality	As no items were found to be misfit after the examinations of fit residuals, χ^2 statistics, and standard error value, the data seemed to have fit the Rasch model well, i.e., unidimensional.
	Local item independence (LID)	Four cognate sequence-ordering items were found to be violating the LID. They were subsequently merged into one item to resolve the LID issue with an improved model fit.
Generalisability	Differential item functioning (DIF)	No items were found to have an issue of DIF concerning gender and disparate test administrative platforms.
External	Person-item map	See Content (person-item map) for reference
Consequential	Item fit and DIF	See Content (item fit) and Generalisability (DIF) for reference

Discussion

The psychometric properties of 522 multiple-choice and short-answer items from 46 Primary 1 to 6 reading passages were evaluated based on an anchor-item calibration design using the Rasch model via the RUMM 2030 software. This study provided evidence of the benefits of

using Rasch in examining to what extent the data fit the model, unidimensionality, local independence, and differential item functioning. The extent to which the data fit the model can be assessed through fit statistics. The results showed that the item responses fit the Rasch model adequately as the χ^2 item-trait statistics associated with these items were not statistically significant. In other words, the responses of items operated consistently with each other in reflecting a single variable as summarised by the Rasch model. The items in the bank, therefore, aligned well with the construct of the AfL tool.

The Person Separation Index (PSI) results for Primary 1 and 2 were less than 0.65, suggesting these tests might not be as sensitive to distinguish between the high and low performers. PSI, in general, increases systematically with the number of thresholds within an item as well as with an increase in the number of items, providing the items and thresholds are functioning as required (Andrich & Marais, 2019). The small PSI values in Primary 1 and 2 items were primarily due to the small number of items within the testlets (four items in Primary 1, and six to eight items in Primary 2) and the small number of thresholds within each item, i.e., one threshold with two ordered categories (dichotomously scored). In contrast, the PSIs for Primary 3 to Primary 6 were generally much higher because there were more items within each testlet (12 to 15 items), though the number of thresholds within each item remained the same as those in Primary 1 and Primary 2.

The issues of bias in gender and test platform can be examined through item residuals of ANOVA. Items having different relative difficulties for groups and, therefore, violating invariance are referred to as bias (Andrich & Marais, 2019). The DIF results showed that there was no evidence to substantiate DIF existed amongst items in the item bank concerning the variable's *gender* or *test platform*. In other words, the items functioned in the same way for different groups of people with respect to *gender* or *test platform*, who had the same value on the trait.

Local Item Dependence (LID) was investigated through correlations of item residuals derived using PCAs. The results showed that the residual correlations of four sequence-ordering items were found to be above 0.30, i.e., displaying evidence of item dependence, as these four items required students to order the storyline of a passage to test their overall passage comprehension, which was closely cognate with one another in nature. The four items were finally grouped into one as recommended in the literature (e.g., Pallant & Tennant, 2007) to circumvent the issue of independence assumption violation as well as to improve on the overall model fit.

The distractor analysis plots showed that 19 out of 522 items across the six test levels had distractors that were not functioning as expected. In all these cases, there was at least one distractor where students of middle-to-high ability range tended to select it over the answer key (or correct answer). A subsequent review of these items by the expert panel suggested that ambiguity in the passages, item stem, and/or distractors could have contributed to the unexpected pattern of student responses. For the remaining 503 items, the probability of choosing the distractors over the answer key decreased as the student's ability increased. In other words, the distractors were functioning as expected. The probability of choosing the distractors over the answer key was also lower for easier items than for the hard ones, which was also expected. The implications of these findings are two-fold. Firstly, the value of using distractor analysis is to understand students' reading comprehension level better by means of their choice of responses because distractors were developed to provide insights into identifying gaps in students' competency when applying comprehension skills and strategies (Gierl, et al., 2017). Student responses can thus serve as useful information for the development of instructional strategies. Secondly, the distractor analysis also provides item setters with an opportunity to examine items for any ambiguity and improves item and test construction skills.

After the identification, review, and removal of unworkable items from the item bank based on the Rasch evaluation criteria, the bank was populated with English reading comprehension items whose fit statistics were adequate with appropriate distractors and free from LID and DIF issues. In other words, the item bank comprised items with strong psychometric properties.

For the item bank's quality aspect, the results also showed evidence of stability over the use of anchored items to link other items for different test levels. Based on the Rasch model, the item difficulty of the anchored items (or common items amongst the item pool) can be estimated, and these parameters can then be used to estimate the unique items in the item pool. The nature of robust item-linking could be illustrated by the high correlation coefficient between an-chored and unanchored items shown in the scatter plot in Figure 5. The practical importance of this result is that building an item bank laden with quality items whose item difficulties could be lined up on a single calibrated continuum scale requires strong linking items selected amongst different test forms within and between test levels. Moreover, as the items are linked horizontally and vertically, this has made items of similar and different test levels psychometrically comparable.

Despite the rigor imposed during the development of these test-level comprehension tests, the study is never complete without limitation. One limitation is the limited range of item types. As the test items are developed/designed to be auto-marked by the machine (or the administration platform) for the advantages of accuracy and speed, only a limited range of item types, such as MCQ, single-word answers, and short-phrase answers, were adopted in the AfL tool. The limited item formats are likely to impact the items' difficulty, which could also affect the sensitivity of the test. For example, since lower primary students are, in general, not familiar with using the keyboard, MCQ items are, therefore, the only item type included in the tests. As MCQ items are typically easier than those of open-ended, the limited item difficulty range might have contributed to the low PSI observed for Primary 1 and 2's tests, where they were not as sensitive in differentiating students with a range of abilities, on top of the reason of small number of items within a passage at lower primary level as discussed previously. The limited item difficulty range could also explain the observation in the person/item map at some test levels where the overlap between the person's ability and item difficulty distributions was not as large as expected. In other words, a proportion of students at higher ability levels would not be able to measure adequately by any item(s). Having said that, as this AfL tool is to target weaker students and identify their weaknesses to help close their learning gaps, it is acceptable that items of a certain difficulty level (typically difficult items) are limited as the item coverage of the target mastery level has been achieved.

CONCLUSION

The use of Rasch in association with Messick's unified model of construct validity in this paper has provided a solid blueprint for establishing a quality item bank for reading comprehension in the English language, which could certainly be extended to other AfL tools of a similar nature. Evidence in the paper suggested that linking items with a high correlation coefficient between anchored and unanchored items (see Figure 5) is key to building a reliable item bank with item difficulties capable of lining up on a single calibrated continuum scale. The use of Messick's unified model of construct validity across the six aspects, i.e., content, substantive, structural, generalisability, external, and consequential aspects, which manifest themselves in the quantitative output of Rasch, has provided an objective approach to screen the psychometric properties of each item and weed out any item which falls short of the Rasch criteria to win its deserving place in the item bank for live assessment. Combining Messick's unified model of construct validity, Rasch has proved itself to be a powerful agent in building a valid and reliable item bank and is gradually finding its place in the area of language assessment.

ACKNOWLEDGMENT

The authors would like to thank current and past members of the SEAB Reading Literacy Team and the expert panel, as well as Yue Lip Sin, Thong May Teng, Tay Poh Hua, and Loi Guang You for their support and feedback on this piece of research. Special thanks go to the teachers and students who participated in the study.

DISCLOSURE STATEMENT

The authors declare that there are no potential conflicts of interest concerning this article's research, authorship, and/or publication.

REFERENCES

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory. *Springer texts in education*. Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8_4

Bond, T. G. (2003). Validity and assessment: A Rasch measurement perspective. *Metodologia de las Clecias del Comportamientro, 5*(2), 179-194.

Bond, T. G., & Fox. C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Cantó-Cerdán, M., Cacho-Martínez, P., Lara-Lacárcel, F., & García-Muñoz, Á. (2021). Rasch analysis for development and reduction of Symptom Questionnaire for Visual Dysfunctions (SQVD). *Scientific Reports*, *11*(1), 14855. https://doi.org/10.1038/s41598-021-94166-9

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied psychological measurement*, *41*(3), 178-194. https://doi.org/10.1177/0146621616677520

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 83-102). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315187815-5/applying-rasch-measurement-language-assessment-jason-fan-trevor-bond

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Hagell, P., & Westergren, A. (2016). Sample size and statistical conclusions from tests of fit to the Rasch Model according to the Rasch Unidimensional Measurement Model (RUMM) Program in Health Outcome Measurement. *Journal of Applied Measurement, 17(4),* 416 – 431.

Hwang, G. J., & Wu, P. H. (2014). Applications, impacts, and trends of mobile technology-enhanced learning: A review of 2008–2012 publications in selected SSCI journals. *International Journal of Mobile Learning and Organisation*, *8*(2), 83-95. https://doi.org/10.1504/IJMLO.2014.062346

Korhonen, J., Linnanmäki, K., & Aunio, P. (2014). Learning difficulties, academic well-being, and educational dropout: A person-centred approach. *Learning and Individual Differences*, *31*, 1-10. https://doi.org/10.1016/j.lindif.2013.12.011

Livingstone, S., Mascheroni, G., & Staksrud, E. (2018). European research on children's internet use: Assessing the past and anticipating the future. *New Media & Society*, *20*(3), 1103-1122. https://doi.org/10.1177/1461444816685930

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Messick, S. (1994). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *ETS Research Report Series*, *1994*(2), i-28. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

Ministry of Education of Singapore. (2020). Pedagogy: Teaching and learning English. In *English language syllabus – Primary foundation English - Secondary normal (technical) course*. Curriculum Planning and Development Division (CPDD), pp. 30–32. https://www.moe.gov.sg/-/media/files/secondary/syllabuses-nt/eng/felnt_els-2020_syllabus.pdf

Mugnaini, D., Lassi, S., La Malfa, G., & Albertini, G. (2009). Internalizing correlates of dyslexia. *World Journal of Pediatrics*, *5*, 255-264. https://doi.org/10.1007/s12519-009-0049-7

Mullis, I. V. S., & Martin, M. O. (Eds.). (2021). *PIRLS 2021 Assessment frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://pirls2021.org/frameworks/wp-content/uploads/sites/2/2019/04/P21_FW_Ch1_Assessment.pdf

Oakley, G. (2011). The assessment of reading comprehension cognitive strategies: Practices and perceptions of Western Australian teachers. *The Australian Journal of Language and Literacy, 34*(3), 279-293. https://doi.org/10.1007/BF03651863

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. OECD Publishing. https://doi.org/10.1787/9789264266490-en

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*(1), 1-18. https://doi.org/10.1348/014466506X96931

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment*, 31-88. https://doi.org/10.4324/9781410612762

Rasch, G. (1960*). Probabilistic models for some intelligence and attainment tests.* Copenhagen Danmarks Paedagogiske Institut.

Ravand, H., & Firoozi, T. (2016). Examining the 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. https://www.ijlt.ir/article_114414.html

RAND Reading Study Group (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. RAND Corporation. https://www.rand.org/pubs/monograph_reports/MR1465.html

Smith, R., Snow, P., Serry, T., & Hammond, L. (2021). The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, *42*(3), 214-240. https://doi.org/10.1080/02702711.2021.1888348