Using the Rasch Model to Calibrate Items for an English as a Second Language Reading Comprehension Computer Adaptive Placement Test

Tzemin Chung[1]✉, Mohd Zali Mohd Nor[2], Richard Yan[3], Peing Ling Loo[4]

[1] tzemin_chung@CommonTown.com
CommonTown
(Singapore)

[2] mohd.zali@my-newstar.com
Newstar Agencies
(Malaysia)

[3] richard_yan@CommonTown.com
CommonTown
(Singapore)

[4] joel@CommonTown.com,
CommonTown
(Singapore)

✉ Corresponding author

ABSTRACT: *This study aimed to calibrate items aiming at measuring English reading comprehension ability in students who learn English as a second language. A total of 571 multiple-choice items were administered to 466 participants from 14 schools via a computer-adaptive test. These participants were mainly secondary school students. Data were analyzed using the Rasch model for dichotomous items. Results indicated that the instrument was sufficiently unidimensional and was quite well targeted at the students. It was able to measure the English abilities of secondary school L2 students. Item measures were also compared to the expert's levelling of item difficulty levels. Based on the Pearson Correlation coefficient of 0.77, the items demonstrate a moderate shared variance, indicating a reasonably positive correlation with the expert's levelling.*

KEYWORDS: **Rasch model; English as a second language, reading comprehension, item calibration; computer adaptive placement test.**

➔ Received 29/04/2023 ➔ Revised manuscript received 29/07/2023 ➔ Published 30/12/2023.

1. Introduction

Accurately assessing the language proficiency of English language learners is essential for providing appropriate instruction and support (Cummins, 2000). In recent years, digital platforms have emerged as a popular means of delivering language instruction, offering a variety of resources to support learners at different levels of proficiency (Chapelle & Hegelheimer, 2004). However, accurate placement is critical to ensuring that students receive appropriate instruction that meets their needs and maximizes their potential for language acquisition (Brown, 2015).

In this paper, we describe an investigation of the psychometric properties of a placement test for English language learners based on the Rasch model. The test is designed to assess students' reading levels through a series of multiple-choice questions. Based on their performance on the test, students are then assigned appropriate books that match their reading abilities.

This study aimed to examine aspects of the reliability and validity of a placement test as a measure of language proficiency and to identify any areas where the test could be improved. By using the Rasch model, we aim to ensure that it is a reliable and valid measure of language proficiency with the potential to support the language acquisition goals of English language learners.

2. Literature review

Learning a new language can be a daunting task, but there are many different approaches one can take to acquire new language skills. Some may prefer to find a teacher and follow a structured curriculum, while others may prefer to learn on their own by reading materials that they find interesting. In fact, Stephen Krashen advocates for the latter approach with his Comprehensible Input hypothesis, which includes five different hypotheses related to language acquisition (Krashen, 2009).

Krashen's Comprehensible Input hypothesis

Krashen's Comprehensible Input hypothesis proposes several hypotheses related to language acquisition (Krashen, 1982). One of the most important hypotheses is the Input hypothesis,

which suggests that learners improve in a language when they are exposed to language that is slightly more difficult than their current level, represented by "i+1" (Krashen, 1981). Another hypothesis is the Acquisition-Learning hypothesis, which suggests that language acquisition is a subconscious process that is more effective than conscious language learning (Krashen, 1982). The Monitor hypothesis states that consciously learned language knowledge can only be used to monitor language use, but does not improve language skills (Krashen, 1982). The Natural Order hypothesis posits that language acquisition follows a predetermined order (Krashen, 1982). Finally, the Affective Filter hypothesis suggests that learners' emotions and attitudes can impact language acquisition, with negative emotions such as fear hindering language acquisition (Krashen, 1982).

Creating a reading program for English language acquisition

Based on Krashen's Input hypothesis, an effective reading program for learning English as a foreign language should include the following features:

Many books, many book series - to provide large amount of language input: Stories and narration are essential language inputs, and the more stories available, the more choices readers have to select the ones they find interesting. The themes, story plots, and illustrations should be appealing to the readers. The more they read, the more they listen, and the more they acquire the language. According to Lee (2018), optimal acquisition of the language input is achieved when reading materials are so interesting that readers are completely immersed in them and not aware that they are reading in a foreign language. When readers are fully engaged with what they are doing, they are in a state of flow (Csikszentmihalyi, 1990; Oppland, 2016).

Furthermore, Cho and Krashen (2016) suggest that readers who have a genuine interest in a specific topic or theme are more likely to engage in narrow reading, which involves reading extensively within a particular topic, author, or genre. Consequently, offering readers samples or short passages from different authors and genres may not be as effective in facilitating language acquisition as offering book series that focus on topics.

Many levels of books - for adaptive reading: According to the theory of comprehensible input, language inputs are more effective when they are comprehensible, and readers acquire language more effectively when they read materials that are slightly more challenging than their language levels (Krashen, 1985). To achieve this, books should be levelled so that readers can choose from books that are suitable for their reading levels (Mason & Krashen, 1997).

Many games - to provide compelling inputs: Comprehensible input may not get the attention of the readers unless it is compelling and interesting (Krashen, 2011). When students read stories or watch movies that are compelling, they will fully enjoy the experience and become avid readers, resulting in improved language acquisition. Krashen further notes that playing video games can be an effective way to acquire language inputs. In fact, Lewis (2020) noted significant improvement in his students' English after they played English video games.

Readin.Town - an English Reading Program

Readin.Town is an English reading program developed by CommonTown Pte Ltd in Singapore. The platform offers a collection of reading materials curated or edited by native speaker content experts from the United Kingdom. Functioning like a library, Readin.Town provides a wide range of books and book series spanning diverse topics, including adventure, classics, history, and science fiction, among others. To enrich the reading experience, the platform includes helpful aids such as native speaker narration, appealing illustrations, and an on-demand dictionary.

In addition to individual books, Readin.Town offers a selection of narrow reading support through series like Alice in Wonderland (see Figure 1), the Horace series, and the Glyn series. These stories, with a common theme or written by one author, aid readers in efficiently learning new words (Gardner, 2008). As suggested by Renanda, Krashen, and Jacobs (2018), book series are recognized as a potent tool to engage

students, featuring highly familiar language, easy-to-follow storylines, and relatable characters. By reading book series, students can enhance their reading proficiency, vocabulary, grammar, and understanding of text structure.

Figures 1. Alice in wonderland series.

The Horace series (2022-2023).
* Horace goes to Town
* Horace's Birthday Pie
* Horace and the Mean Monsters.
* Horace and his Cheese
* Horace's Treasure Hunt
* Horace goes High Up
* Horace and the Wizard

The Glyn Series (2022-2023).
* Glyn Finds Out About…. Food
* Glyn Finds Out About… Keeping Fit
* Glyn Finds Out About... Pets
* Glyn Finds Out About… Birthdays
* Glyn Finds Out About… Being Sad
* Glyn Finds Out About… Being Scared
* Glyn Finds Out About... Being Angry
* Glyn Finds Out About... Being Worried

Readin.Town offers a comprehensive range of books to suit readers of all levels, with 20 levels available, from preschool to pre-university. This is equivalent to the Common European Framework of Reference for Languages (CEFR) pre-A1 to C2 levels (see Figure 2). The platform uses a computer adaptive placement test to find out readers' English proficiency levels, ensuring that they receive book recommendations appropriate to their reading abilities.

To create a more interactive and engaging learning experience, Readin.Town incorporates a

CEFR				A1L	A1M	A1H	A2L	A2M	A2H	B1L	B1M	B1H	B2L	B2M	B2H	C1L	C1M	C1H	C2L	C2M	C2H
Readin	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figure 2. CEFR levels.

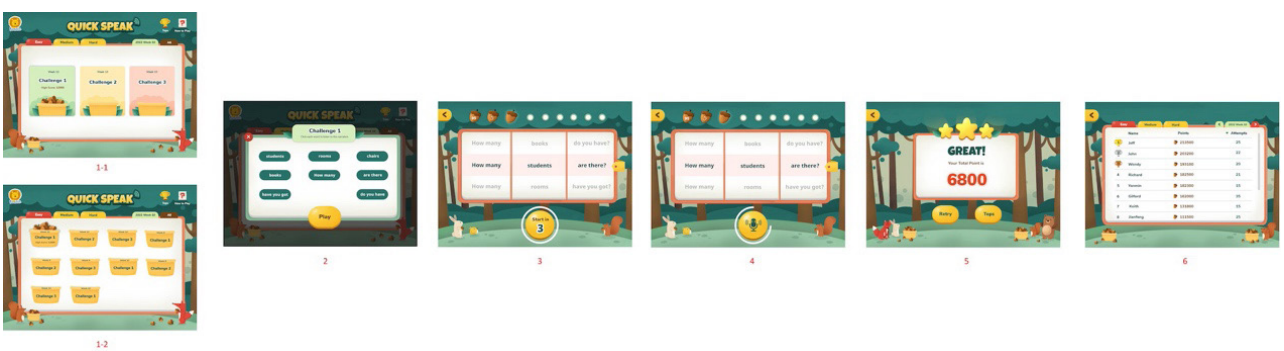

Figure 3. Sentence structure game.

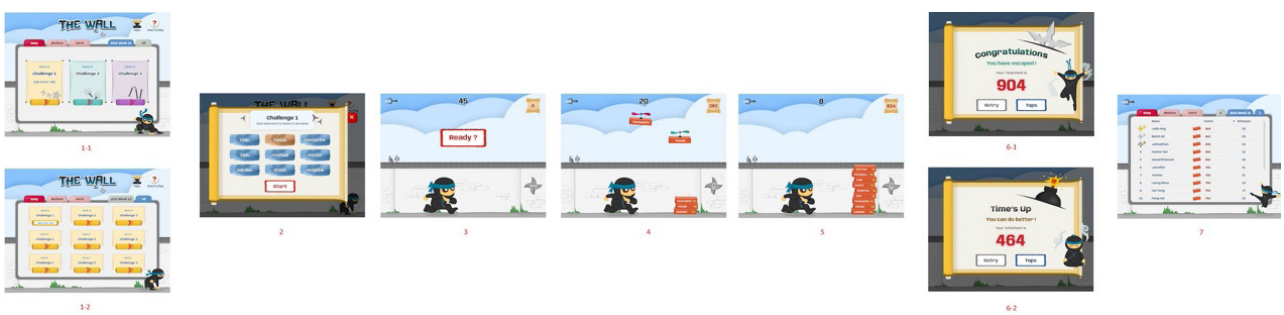

Figure 4. Word recognition game.

variety of educational games designed to enhance vocabulary acquisition and sentence structure skills (see Figures 3 and 4). These games provide a fun and immersive way for readers to enhance vocabulary acquisition and improve sentence structure skills.

As suggested by Bryan (2011), readers should choose books within their current reading level to improve fluency or slightly above their level to enhance vocabulary. To facilitate this, Readin. Town developed a placement test to assess readers' proficiency. Before using Readin.Town, readers take this computer adaptive test to find out their English proficiency level. This enables the program to recommend books suitable for their skill level. In this paper, we conducted Rasch analysis to investigate the psychometric properties of the placement test.

3. Methodology

Materials

The item pool comprised 559 items, covering difficulty levels ranging from CEFR Pre-A1, A1, A2, B1, B2, C1, C2, which correspond to levels from preschool to pre-university. When the student began taking the test, there were non-adaptive items to find out their reading level. Then, items that were appropriate to their reading levels were assigned adaptively. The items focused mainly on vocabulary and were all in a multiple-choice format, each with four distractors. For example,

"Don't _____ to send your brother a birthday card.

a. forget, b. remember,
c. remind, d. think";

"I'm sorry, I don't _____ French.

a. talk, b. say,
c. tell, d. speak."

Participants

A total of 466 participants from 14 schools took the placement test. The participants included 434 secondary school students (aged 13-19), 22 primary school students (aged 11-12), and 11 adults, including the students' teachers and school staff. Among them, 26 students learned Chinese as their native language, while the rest learned Malay. All students learned English as a second language.

Procedure

Each participant was given a 20-minute computer adaptive placement test. The participants were first administered a standardized test to determine their entry levels, after which the subsequent items were administered adaptively. This process continued until the ending criterion was met.

Data Analysis

The items were evaluated using Rasch analyses. Some items received very few responses, likely due to the small sample size and the distribution of item difficulty levels. To ensure the reliability of the analysis, items with less than 15 responses were excluded, resulting in the removal of 344 items. Winsteps software (Linacre, 2014g) version 3.81.0 was used to analyze the data.

The Rasch analysis was performed to identify items that do not contribute to useful measurement via item polarity, summary statistics, Rasch statistics for individual items, separation and reliability, item location and person measures, item discrimination, and dimensionality.

Item Polarity

Item polarity was analyzed to check if the responses to items aligned with the overall measure (Kelly et al., 2002). Positive point-measure correlations indicate that the items focused on the single construct, whereas negative or near-zero correlations indicate problematic items that needed revision or removal (Linacre, 2014a).

Summary Statistics

Summary statistics involved arranging items on an interval scale based on their difficulty level and measuring person abilities on the same continuum (Granger, 2008).

Bond and Fox (2015) indicated that fit statistics are utilized to assess how well the observed data aligns with the model being used. Among these statistics, the "infit" and "outfit" mean square values specifically examine the level of misfit present in the data. In other words, they help answer the question: To what extent does the data deviate from what is expected by the model?

The average of item measures was set at 0, by default. In addition, the acceptable ranges for infit and outfit mean squares are 0.5 to 1.5 (Bond & Fox, 2015). Mean-squares near 1.0 indicate little distortion of the measurement system (2002). However, we need to bear in mind that Rasch fit statistics (mean squares and t-statistics) are highly susceptible to sample size variation for dichotomously scored rating data (Smith et al., 2008).

Rasch Statistics for Individual Items

Rasch infit and outfit statistics are commonly used to identify problems with individual items in measurement scales. According to Granger (2008), infit is more diagnostic when the item measures are close to the person measures, while outfit is more diagnostic when the item measures are far from the person measures. Infit mean-squares greater than 1.0 indicate underfit, meaning the data is less predictable than expected, possibly due to high-ability students missing easy items. Conversely, mean-squares less than 1.0 suggest overfit, meaning the data is more predictable than expected due to redundant items. Linacre (2014d) suggests that items with values between .5 and 1.5 are the most productive for measurement.

In addition, it is necessary to examine the standardized fit statistics for the items. A standardized value greater than 0 indicates unpredictability of the items, while a value less than 0 suggests items that are overly predictable. However, if the mean-squares are deemed acceptable, the standardized fit statistics can be disregarded (Linacre, 2014d).

Items with infit and outfit mean-squares less than 0.5 and negative Z standardized values (<-2) indicate redundant responses, but they do not distort the measurement scale (Bond & Fox, 2015; Linacre, 2014d; Wright & Linacre, 1994). To find redundancy, Linacre (2000) suggests verifying whether two items with the same measures have independent responses, as this increases the local precision of person measures and is beneficial for computer-adaptive tests.

Items with infit and outfit mean-squares less than 0.5 and negative Z standardized values (>2) indicate unpredictable, erratic responses (Bond & Fox, 2015). Infit underfit issues are often related to alternative curricula or idiosyncratic groups, making them harder to diagnose and remedy than outfit issues (Linacre, 2002). Outfit underfit issues typically result from careless mistakes and lucky guesses, while overfit issues stem from imputed responses. Item writers can review these statistics to decide whether to modify or delete an item, starting with those with very high mean-squares resulting from random guessing and then checking those with low mean-squares (Wright & Linacre, 1994). According to Marais (2015), when poorly fitting items are removed from the analysis, the model is adjusted, leading to an expansion of the vertical scale, which indicates a broader range of proficiency levels. As a consequence of this adjustment, some items initially overfitting the model may now fit better, providing more meaningful information about the underlying construct being measured.

Separation and Reliability

Separation refers to *"the ability of the test to define a distinct hierarchy of items along the measured variable"* (Bond & Fox, 2015, p. 70). A higher item separation indicates that "we can place more confidence in the replicability of item placement across other samples" (Bond & Fox, 2015, p. 70). On the other hand, reliability refers to the *"reproducibility of relative measure location"* (Linacre, 2014e, para. 5).

If the item separation is less than 3 (which means items cannot be differentiated into high, medium and low levels) and item reliability is less than .9, then the data are considered to have low separation, indicating that the sample size is not sufficient to attempt a wide range of items (Linacre, 2014e).

Similarly, person separation indicates how well the items in the test are able to separate the sample into different ability levels (Linacre, 2014e). A higher reliability indicates "better separation that exists and the more precise the measurement" (Wright & Stone, 1999, p. 151). If the person separation is less than 2 (which means persons cannot be differentiated into high and low ability) and person reliability is less than .8, then the items are not sensitive enough to person abilities (Linacre, 2014e). Finally, item

location and person measures were calibrated on a shared scale of the latent construct, allowing the comparison of the average person ability and item measure. An average person measure close to the average item measure indicated a good fit of the items to the construct.

Item Location and Person Measures

The Rasch model provides a shared scale for calibrating both item measures and person abilities, allowing for the comparison of the average person ability and item measure on a common scale (Granger, 2008). The average item measure is set at 0, and an average person measure close to the average item measure indicates that the difficulty levels of the items are well targeted with respect to the sample. If mean person measure values are significantly higher or lower than the average item measure, it suggests that the items are mistargeted for that sample (Bond & Fox, 2015).

Mistargeting can also occur when items cover a broad range of the scale, but most respondents' abilities are concentrated in a different range. This concentration results in gaps in the scale, reducing the precision of person (and item) parameter estimates and leading to larger standard errors (Salzberger, 2003).

To visually inspect the distribution of person abilities and item measures, the Wright Map (Lunz, 2010) can be used. This graph helps identify if the item set is too difficult or too easy for the sample and which part of the scale lacks items.

Item Discrimination

If students who score high on an exam also correctly answer a particular item, this item is able to differentiate those who know the content from those who do not. An item with zero or negative discrimination undermines the test (Kelley et al., 2002). The useful range of item discrimination is .5 to 2 (Linacre, 2014h).

Dimensionality Check

An underlying assumption of the Rasch model is that a single latent trait accounts for test-takers performances on the set of items in the measure. Each student in the sample has an amount of the latent trait to be measured when they respond to items. Rasch analyzes whether the response patterns fit the Rasch model. It employs Principle Component Analysis (PCA) of the residuals to look for unexpected response patterns in the data that do not fit the model. If a group of items shared the same unexpected pattern, then there may be another latent trait at work (Linacre, 2014b). This additional latent trait is a "secondary dimension" that requires further investigation. The items that do not conform to the Rasch model should not be used in the measurement scale. They need to be improved or removed (Tennant & Pallant, 2006).

When using CAT to administer the test, we need to bear in mind that CAT requires a large number of items in the pool to ensure that multiple relevant items are available at various ability levels. The item pool used in a CAT should ideally reflect the unidimensionality of the construct being measured. Unidimensionality means that all items in the pool are related to a single underlying trait. Careful item selection and evaluation of the item pool's dimensionality are crucial to ensure that the CAT accurately measures the targeted construct.

Also, CAT uses adaptive item selection algorithms to choose the most informative items for each test taker based on their ability level. These algorithms need to be carefully designed to maintain measurement precision and unidimensionality throughout the test administration.

4. Results

Item Polarity

Point-measure correlations for items provided an immediate check for scoring mistakes. Items with negative point correlations have to be investigated before the Rasch fit statistics (Linacre, 2014a). Results showed that 15 items have negative point-measure correlations. Distractor analysis showed that one item has incorrect answers. Two items were too difficult, resulting in random guesses, e.g., The American runner drew _____ his deep reserves of stamina to win the race. a. back b. up c. on d. forwards. A total of 62% selected choice b, which indicated that even good students got this item wrong.

Distractors of the rest of other items were not effective. One to three distractors were not

selected, which showed that the items were easy for the students.

Model-Data Fit Analysis

In the initial analysis, the summary statistics (Table 1) indicated that infit mean-squares close to 1.0 and standardized Z scores near 0 suggest little distortion in the measurement system (Linacre, 2002). While the outfit mean square appeared to be higher than expected, the outfit Z score was still close to 0, indicating possible random guessing items. Further examination of additional statistics is required to assess the item fit and ensure the validity of the measurement model.

Next, we proceeded to examine the individual item fit statistics. Results for individual item infit (Table 2) and outfit (Table 3) statistics indicated that there were more erratic responses "far from the person measures" than responses "close to the person measures." Out of the 118 items analyzed, 26 items were underfitting (>1.5); that is, they were above the outfit mean-squares range for a productive measure, as opposed to 2 items (>1.5) outside the infit mean-square range.

The larger number of outfit items that misfit was due to higher-ability students getting the easier items wrong. For example, "_____ at that boat!" a. see, b. watch, c. listen, d. look, nobody selected choices a and c, which left only two choices, making the item easier (item measure was -3.8). However, with such an easy item, a medium-high level student got it wrong, resulting in a high outfit mean square of 6.83.

Furthermore, some items were difficult, such as "I think we got our _____ crossed. I said I couldn't give you a lift this evening." a. minds, b. string, c. wires, d. connections. Students appeared to be guessing, as evidenced by their responses (a = 62, b = 15, c = 17, d = 32), resulting in an outfit mean square of 2.07.

Two items' infit mean squares were greater than 1.5. One of them was an easy item ("Find the picture of goat", infit mean square = 1.53), where one of the distractors was not selected, making it even easier than intended. Only one medium-high ability student answered this item incorrectly, which suggests that it may not be a serious issue. To address the issue, we will begin by revising the distractor that was not selected by the respondents, and then we will retest the item with the revised distractor.

The item "bad / worst, pretty / prettiest, good / _____" a. gooder, b. worse, c. best, d. bestest with an item measure of -1.1 and an infit mean square of 1.56 was leveled by an expert to be appropriate for primary 3 level and was not considered a difficult item. However, a notable number of high-ability and medium-ability secondary school students (12 and 15, respectively) answered it incorrectly (A=31, B=17, C=95, D=4). The high number of incorrect responses to the item testing superlatives makes it unclear whether the issue is related to how the item was presented, students forgetting the taught material, or the possibility that some students were never taught about superlatives in school. If the latter is the case, it could indicate a more serious concern regarding missed learning opportunities or belonging to an idiosyncratic group.

Based on the results of this study, we will make necessary modifications to the items and conduct a re-trial to improve the Rasch measurement properties of the test.

Separation and Reliability

The high item reliability of .94 suggests that the order of item estimates can be confidently replicated when administering these items to other appropriate samples. However, it should be noted that high item reliability can be influenced by large sample sizes (Bond & Fox, 2015, p.70).

Table 1. Summary statistics from Rasch Analysis.

	Measure (logit)	**SE**	**Infit MNSQ**	**Infit ZSTD**	**Outfit MNSQ**	**Outfit ZSTD**	**Separation**	**Reliability**
Person	.17	.61	.97	-.2	1.08	.1	2.04	.81
Item	.00	.47	.98	-.1	1.22	.2	4.12	.94

Table 2. Infit Mean-squares item distribution.

Infit Mean-squares	Item
< 0.50	0
0.50 - 0.70	3
0.71 - 0.90	55
0.91 - 1.10	99
1.11 - 1.30	22
1.31 - 1.50	7
1.51 - 1.70	2
1.71 - 1.90	0
1.91 - 2.10	0
2.11 - 2.30	0
> 2.30	0
Total	188

Table 3. Outfit mean-squares item distribution.

Outfit Mean-squares	Item
0.00 - 0.49	6
0.50 - 1.00	104
1.01 - 1.50	52
1.51 - 2.00	9
2.01 - 2.50	6
2.51 - 3.00	5
> 3.00	6
Total	188

To further validate the item hierarchy, we also examined the item separation index (4.16), which confirmed a reliable hierarchy of item difficulty across four levels. However, since the sample consisted mainly of secondary school students, a wider ability range of individuals, including primary school students, would be needed to separate more difficulty levels (Linacre, 2014e).

The reliability of person ability estimates was .81, with a person separation index of 2.04. While these values indicate that the items were sensitive enough to distinguish between high and low performers, the separation between high and low students is not adequate. To achieve a more precise measurement, we need more well targeted items to separate more levels of students.

Item Location and Person Measures

The Wright map (Figure 5) visually shows how well test questions match candidates' abilities on the same measurement scale. This helps to assess the test's effectiveness in measuring candidates' abilities and the appropriateness of the questions.

Most of the items showed a good range of difficulty, widely spread along the logit axis. However, due to inadequate responses, we lacked sufficient students at certain difficulty levels. Only 8 students attempted items at the lower end of the axis (-6 to -3 logits), and merely 26 students attempted items from 2.2 logits onwards. There were no students beyond 3.8 logits. The key

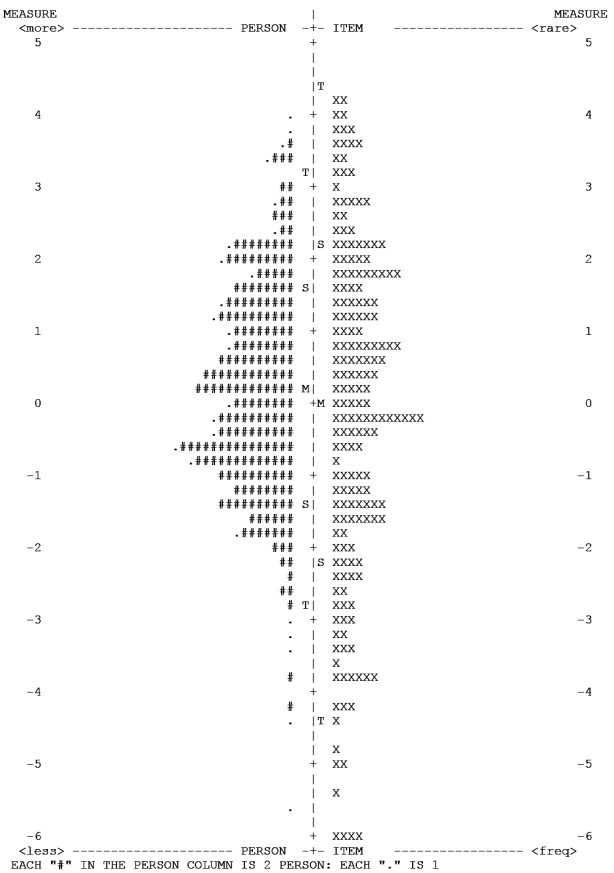

Figure 5. The Wright Map of student abilities and item measures.

issue here is the insufficient number of students to accurately estimate item difficulty in these ranges, highlighting the need for more students at those ability levels to properly test the items. In this context, a rectangular distribution of students is preferred over a normal distribution. A rectangular distribution ensures an even representation of students across the ability range, including both very high and very low

ability students. This will ensure better coverage of the entire range of item difficulties.

We can enhance the test by creating additional items, particularly at the top end of the logit axis (4-5 logits), to extend the range of difficulty levels and effectively measure high ability students

Item Discrimination

Item discrimination analysis identified five items (out of xxx) with zero or negative discrimination (Table 4). These five items were answered incorrectly by higher-ability students. They require immediate revision to improve the overall test quality.

Dimensionality

The dimensionality analysis results (Table 5) revealed that 40.7% of the variance was explained by the latent trait, which is higher than the recommended guideline of 29.5%

Table 4. Items with Negative and Zero Discrimination Indices.

Item Measure	Expert Level	Discr Index	Item	Distractor	No of Responses	Answer	Zero Response	Remark
-1.4	6.8	-0.7	In summer the weather is _____. a. much hotter, b. more hotter, c. many hotter, d. hoter	D B A	2 4 9	A	C	wrong spelling "hoter". 2 medium ability students answered it wrongly. Weak students answered it correctly
-1.1	5.5	-0.5	bad / worst, pretty / prettiest, good / _____ a. gooder, b. worse, c. best, d. bestest	B D A C	17 4 31 95	C		Low level item. Infit = 1.56: 12 high ability and 15 medium ability students answered it incorrectly.
-0.3	9.2	-0.4	If you're allergic to something, try not to _____, as it can become infected. a. Itch, b. rub, c. scratch, d. mark	B A C	2 6 13	C		While 2 high ability and 6 medium ability students answered it incorrectly, 2 low ability students answered it correctly
-5.0	2.5	0.0	Dan hopes to fly an airplane when he grows up. (Text and Audio) a. an airplane (picture), b. a rocket (picture), c. a ship (picture), d. a train (picture)	D B A	1 2 14	A	C	Easy question - one medium ability student answered it incorrectly
-1.6	4.5	0.0	Trees are not _____ high _____ mountains. A. like, b. tall, c. too, d. as	C B D	4 3 10	D	A	one medium ability student answered it incorrectly

for computer adaptive tests (Linacre, 2014f). However, there was still a significant amount of randomness in the data (59.3%) (Table 6), which led us to investigate the decomposed unexplained variance to find out if a second dimension had a substantial effect on the scale. The strength of the second dimension is indicated by the Eigenvalue (Linacre, 2014c). In our analysis, the Eigenvalue for the unexplained variance in the first contrast was 8.0, which was much higher than the recommended value of 2, and the variance explained by the first contrast was 2.5%. We then examined the contrast between the items at the top and bottom of the contrast plot (Figure 6) to see if they were different enough to warrant a second construct (Linacre, 2014c). There were eleven items at the top of the plot and seven at the bottom (Table 6). Our content expert investigated these items and found no particular structure or item explaining a second construct. These results indicated that the data could be accounted for by only one dimension, which is the latent trait of reading ability.

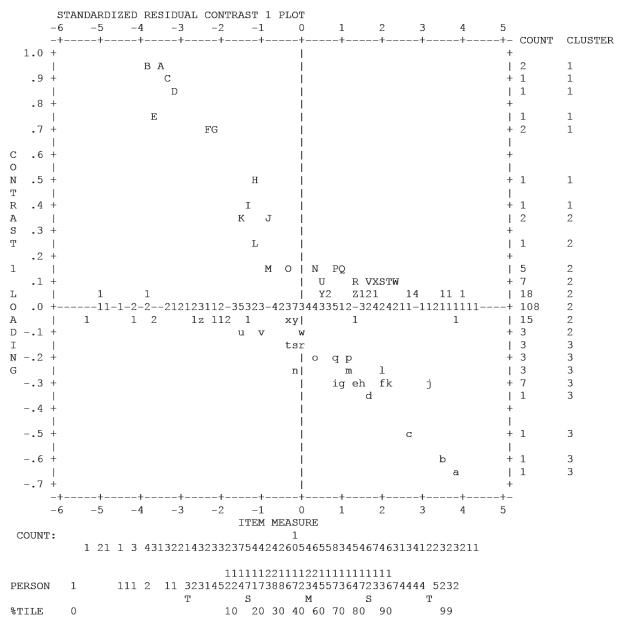

Figure 6. The standardized residual contrast 1 plot.

Concurrent Validity and Invariance Analysis

In their work, Wright, Huber, O'Neill, and Linacre (2000) argued that "If the difficulty of an item were not invariant over some useful domain, then the term difficulty would have no useful meaning." Consequently, to assess the utility of item difficulty, we investigated whether item measures (Mean = -0.12, SD = 2.36) aligned with estimates of item difficulty provided by a proficient English as a foreign language teacher who is a native speaker (Mean = 9.06, SD = 3.81). A moderately strong positive correlation of .77 was found between the two measures, indicating that as item measures increased, expert estimates also tended to increase. However, two items

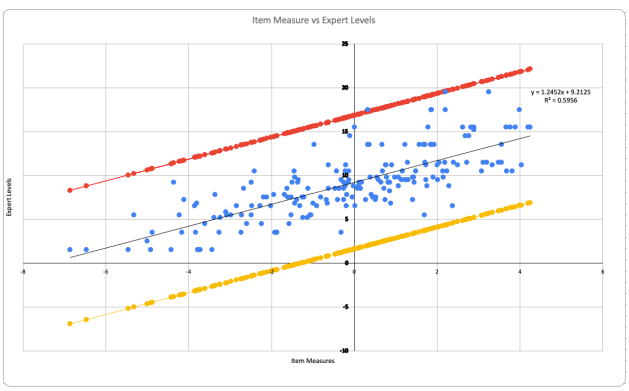

Figure 12. Pearson correlation between item measures and expert's estimates.

Table 5. Standardized residual variance (in Eigenvalue units).

	Eigenvalue	Observed	Expected
Total raw variance in observations	316.9	100.0%	100.0%
Raw variance explained by measures	128.9	40.7%	39.1%
Raw variance explained by persons	47.5	15.0%	14.4%
Raw Variance explained by items	81.4	25.7%	24.7%
Raw unexplained variance (total)	188.0	59.3%	60.9%
Unexplained variance in 1st contrast	8.0	2.5%	4.3%
Unexplained variance in 2nd contrast	4.3	1.4%	2.3%

Table 6. Items loadings and at the top and bottom of the contrast plot.

A	.96	guitar, look for the correct picture a. cello b. guitar c. piano d. violin	a	-.65	I'm afraid your suggestions don't __ the fundamental problem. a. deliver b. address c. argue with d. support
B	.93	Your uncle's son or daughter is your _____. a. nephew b. cousin c. step brother / sister d. grandmother / grandfather	b	-.58	I've _____ to invite you out for a long time. a. been desiring b. been meaning c. decided d. been thinking
C	.89	How _____ are you? a. age b. old c. much d. will	c	-.51	I'm always mixing adjectives _____ with adverbs. a. through b. out c. up d. down
D	.87	You should apologize for _____ impolite. a. bee b. being c. to be d. been	d	-.37	The police decided the suspect _____ committed the murder as he was out of the country. a. must have b. has c. would have d. couldn't have
E	.73	Tablet, look for the correct picture. a. desktop b. notebook c. tablet d. monitor	e	-.32	This door _____ . It won't close properly. a. needs fixing, b. wants fixed, c. needs to be fixing d. has to fix
F	.70	Can you pass me that book? I can't _____ it. a. find b. reach c. have d. hold	f	-.31	It was obvious he _____ smoking, he smelled strongly of tobacco. a. was b. had been c. will be d. might be
G	.68	The daughter of a king is a _____. a. prince b. queen c. princess d. knight	g	-.30	My sister _____ herself on her punctuality. a. loves b. boasts c. tries d. prides
H	.51	You need to book the train ticket in _____. a. advance b. before c. forward d. ahead	h	-.29	If I _____ my mum's size, I wouldn't have bought her the wrong shoes. a. had known, b. knew, c. have known, d. was known
I	.40	I need to _____ a hole in the wall to hang this picture. a. slice b. hammer c. push d. drill	i	-.29	_____ recognition of your hard work, we would like to give you this present. a. out, b. in, c. by, d. at
J	.34	My friend put his hand under the hot tap and got a bad _____. a. cut b. wound c. sore d. burn	j	-.29	I think we got our _____ crossed. I said I couldn't give you a lift this evening. a. minds b. string c. wires d. connections
K	.34	This vocabulary test is extremely _____. a. hard b. hardly c. impossible d. perfect	k	-.28	The billionaire had already _____ a large fortune by the time he was twenty. a. compiled b. amassed c. completed d. aggregated
			l	-.26	They need to _____ that old office building, it's dangerous. a. knock down, b. fall down, c. push down, d. destroy down
			m	-.24	Go and do the shopping, and _____ I'll wait here and have a cup of coffee. a. at the meantime, b. in the moment, c. at the moment, d. in the meantime
			o	-.21	When the wolf came to the well and started to drink the water, the heavy stones in his stomach made him ____ in. a. jump, b. fall, c. go, d. run

			p	-.21	She _____ the necessary experience for the job. a. loses, b. lacks, c. misses, d. does
			q	-.21	This business will never be _____ without proper investment. a. profit, b. profitted, c. profitable, d. profit

Table 7. Item for which Rasch measures and expert levels that did not correlate well.

Item	Item Measure	Expert Level	Action
The thing _____ I love most about her is her sense of humour. a. that, b. what, c. which, d. who	.33	17.5 (~CEFR C1 High)	Removed (due to no replacement for option d)
Although everybody had been drilled in what to do in the event of a fire, when it actually happened there was _____. a. pandemonium b. commotion c. histrionics d. turbulence *Revised item* Despite being told what to do, when we had a fire in the basement, the ensuing _____ took us all by surprise. a. pandemonium, b. hubbub, c. histrionics, d. turbulence	3.25	19.5 (~CEFR C2 Medium)	Revised

showed discrepancies between their measures and expert estimates, suggesting a need for revision and further trialling.

5. Conclusions

This study aimed to calibrate the items for an English as a second language placement test using Rasch analysis. The results showed that the items were productive and relatively well-targeted at the students' ability levels, albeit slightly easy. Sufficient unidimensionality of the data was confirmed by the fit statistics and dimensionality analysis, and the items separated students into high- and low-ability groups. The items were also found to have a hierarchy of four levels of difficulty, corresponding with secondary 1 to 4 (Grades 7 to 10) levels of English skills.

Moving forward, we must prioritize continued testing to refine items in our item pool, including the 34 (18%) items requiring revision and any new items. These iterative testing endeavors are vital for effectively separating students into more ability groups as well as enhancing the test's quality.

References

Bond, T. & Fox, C. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). New York: Routledge.

Brown, H. D. (2015). *Principles of language learning and teaching* (6th ed.). Pearson Education.

Chapelle, C. A., & Hegelheimer, V. (2004). Theorizing and testing technology-mediated second language learning: Issues and hypotheses. *Annual Review of Applied Linguistics, 24,* 223- 247.

Cho, K. S. & Krashen, S. (2016). What does it take to develop a long-term pleasure reading habit? *Turkish Online Journal of English Language Teaching (TOJELT), 1*(1), 1-9. http://www.sdkrashen.com/content/articles/2016_cho_and_krashen_long-term_reading.pdf

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience.* Harper Perennial.

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire.* Multilingual Matters.

Gardner, D. (2013). Vocabulary recycling in children's authentic reading materials: A corpus-based investigation of narrow reading. *Reading in a Foreign Language, 25*(2), 234-253.

Granger, C. (2008). Rasch Analysis is important to understand and use for measurement. *Rasch Measurement Transactions, 21*(3) p. 1122-3. Retrieved from https://www.rasch.org/rmt/rmt213d.htm

Kelley T., Ebel R., & Linacre, J. M. (2002). Item discrimination indices. *Rasch Measurement Transactions, (16*)3, p.883-4.

Krashen, S. D. (1981). *Second language acquisition and second language learning.* Oxford: Pergamon.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications.* New York: Longman.

Krashen, S. D. (2009). *Principles and Practice in Second Language Acquisition.* http://www.sdkrashen.com/content/books/principles_and_practice.pdf

Krashen, S. D. (2011). *The Compelling (not just interesting) Input Hypothesis.* The English Connection (KOTESOL). A Publication of KOTESOL, 15(3). http://www.sdkrashen.com/content/articles/the_compelling_input_hypothesis.pdf

Lee, S. Y. (2018). The Power of Story in SLA: Insights from Research [monograph]. *Reconceptualizing English Language Teaching and Learning in the 21st Century A Special Monograph in Memory of Professor Kai-chong Cheung.* http://www.sdkrashen.com/content/articles/35-sy-ying_lee_(final).pdf

Lewis, R. (2020, November 4). *What Is Comprehensible Input and Why Does It Matter for Language Learning?* https://www.leonardoenglish.com/blog/comprehensible-input

Linacre, J.M. (2000). Redundant items, overfit and measure bias. *Rasch Measurement Transactions, 14*(3), 755. Retrieved from https://www.rasch.org/rmt/rmt143a.htm

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), p.878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2014a). Correlations: Point-biserial, point-measure, residual. *Help for Winsteps Rasch Measurement Software: www.winsteps.com.* Retrieved from http://www.winsteps.com/winman/correlations.htm

Linacre, J. M. (2014b). Dimensionality: Contrasts & variances. *Help for Winsteps Rasch Measurement Software: www.winsteps.com.* Retrieved from http://www.winsteps.com/winman/principalcomponents.htm

Linacre, J. M. (2014c). Dimensionality: When is a test multidimensional? *Help for Winsteps Rasch Measurement Software: www.winsteps.com.* Retrieved from http://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (2014d). Fit diagnosis: Infit outfit mean-square standardized. *Help for Winsteps Rasch Measurement Software: www.winsteps.com.* Retrieved from http://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2014e). Reliability and separation of measures. *Help for Winsteps Rasch Measurement Software: www.winsteps.com.* Retrieved from http://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (2014f). Table 23.0 variance components for items. *Help for Winsteps Rasch Measurement Software: www.winsteps.com.* Retrieved from http://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (2014g). *Winsteps Rasch measurement software [Computer software].* Retrieved from www.winsteps.com

Linacre, J. M. (2014h). Item discrimination or slope estimation. *Help for Winsteps Rasch Measurement Software: www.winsteps.com.* Retrieved from https://www.winsteps.com/winman/discriminationestimation.htm

Lunz, M. (2010). Using The very useful Wright Map. *Measurement Research Associates Test Insights.* https://www.rasch.org/mra/mra-01-10.htm#:~:text=The%20Wright%20Map%20provides%20a,how%20appropriately%20the%20test%20measured.

Marais, I. (2015). Implications of Removing Random Guessing from Rasch Item Estimates in Vertical Scaling. *Journal of Applied Measurement, 16*(2), 113-128.

Mason, B., & Krashen, S. (1997). Extensive reading in English as a foreign language. *System, 25*(1), 91-102.

Oppland, R. (2016). Flow in the foreign language classroom. *The Language Learning Journal, 44*(4), 418-428.

Renandya, W., Krashen, S., & Jacobs, G. (2018). The Potential of Series Books: How Narrow Reading Leads to Advanced L2 Proficiency. *RELC Journal, 49*(2), 199-211. doi: 10.1177/0033688218778549

Salzberger, T. (2003). Item Information: When Gaps Can Be Bridged. *Rasch Measurement Transactions, 17*(1), 910-911.

Smith, A.B., Rush, R., Fallowfield, L.J. *et al.* (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodol 8*(33). https://doi.org/10.1186/1471-2288-8-33

Tennant A., & Pallant J.F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions, 20*(1), p. 1048-51. Retrieved from https://www.rasch.org/rmt/rmt201c.htm

Wright, B.D., Huber, M., O'Neill, T., & Linacre, J.M. (2000). The problem of measure invariance. *Rasch Measurement Transactions, 14*(2), 745. Retrieved from https://www.rasch.org/rmt/rmt142d.htm

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. Retrieved from https://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Stone, M. (1999). *Measurement Essentials* (2nd ed.). Wide Range, Inc.