How to Create Valid Reading Comprehension Tests to Measure Improvement

Sachi Oshima
Chuo Gakuin University

Reference Data:

Oshima, S. (2025). How to create valid reading comprehension tests to measure improvement. In B. Lacy, M. Swanson, & P. Lege (Eds.), *Moving JALT Into the Future: Opportunity, Diversity, and Excellence*. JALT. https://doi.org/10.37546/JALTPCP2024-08

To measure improvement in students' reading performance, teachers and researchers often administer multiple reading tests at different points during a course—typically a pretest, midterm, and posttest. Although using an identical test enables researchers to compare test scores obtained at multiple points directly, there is a "testing threat" that negatively affects validity (Trochim et al., 2016). If the same test is used repeatedly, students might improve their scores; however, this does not necessarily mean their reading performance has improved because they might remember the content of the pretest reading texts and items, lowering the difficulty of the posttest. Different tests consisting of different texts are expected to address this validity threat; still, it introduces another problem: if the reading passages vary considerably in difficulty, then the tests cannot validly measure whether students' reading performance has actually improved. The purpose of this paper is to demonstrate a five-step solution to address this issue: (a) selecting the reading passages used for the tests, (b) analyzing and adjusting the lexical and readability level of the passages, (c) creating question items based on the difficulty level of questions (Burrows, 2012; Lumley, 1993), (d) conducting alpha and beta testing (Fulcher & Davidson, 2007), and (e) employing Rasch analysis to ensure comparable difficulty estimates among multiple reading tests.

生徒の英語リーディング（文章読解）力向上を測定するため、教師や研究者は複数回のテスト（コース開始時のpretest、中間のmidterm、終了時のposttestなど）を実施することが多い。同一のテストを複数回実施することで、スコアの直接比較は可能となるが、testing threat (Trochim et al., 2016)　が妥当性にもたらす影響を考慮する必要がある。同一のテストを繰り返し実施する場合、生徒のスコアは向上するかもしれないが、それが必ずしも生徒の英語リーディング力向上を意味するとは限らない。生徒がpretestの内容を記憶していることでposttestの難易度が下がることもあり得るからである。そこで妥当性を担保するため、異なる文章を用いたテストを準備することが望ましいが、また別の問題が生じる。文章の難易度がそもそも異なる場合、難易度の異なるテストを実施したところで、生徒のリーディング力向上を検証するのは妥当ではない。そこで、本稿では具体的な解決策として、5つのステップ―(a) テストに使用する文章の選定、(b) 語彙・可読性のレベル分析・調整、(c) 設問の難易度 (Burrows, 2012; Lumley, 1993) を考慮した設問作成、(d) アルファテスト・ベータテスト (Fulcher & Davidson, 2007) の実施、(e) 複数のテストが同等の難易度であることを担保するためのラッシュ分析実施―を紹介する。

Many English teachers in Japan are facing a growing need to help low-proficiency university students improve their English skills, including reading. Today, Japanese university students, the majority of whom are studying English as a foreign language (EFL), have generally received ten years of English education by the time they graduate from senior high school: four years in elementary school, three years in junior high school, and three years in senior high school (Mochizuki et al., 2018). However, despite the substantial time spent on English education, the English proficiency of many senior high school graduates is remarkably low. A recent survey conducted by the Ministry of Education, Culture, Sports, Science, and Technology of Japan (MEXT, 2023) has provided further evidence of these low proficiency levels. The survey identified the Eiken Grade Pre-2 English Proficiency Test (Eiken Foundation of Japan, n.d.) at the A2 level of the Common European Framework of Reference for Languages (CEFR) as the expected English proficiency level for senior high school graduates. However, less than half (48.70%) of senior high school graduates are estimated to reach this benchmark. This figure comprises 30.20% who had obtained the CEFR A2-level test score, and the rest were students whose teachers estimated they would reach the CEFR A2 level if tested (MEXT, 2023). In other words, 51.30% were estimated to be at the A1 level, indicating that a significant number of A1-level students enter university every year.

Japanese EFL university students with low English proficiency have not received much attention from previous researchers. For instance, in Sun et al.'s (2021) meta-analysis examining the relationship between reading strategies and reading comprehension, 48 empirical studies (N = 21,548) published from 1998 to 2019 were analyzed; however, only one of the studies concerned Japanese EFL university students with proficiency at the CEFR A2-B1 level (Hayashi, 1999).

Previous researchers have called for more studies on reading strategy instruction to help learners, including those with low proficiency, improve their English reading skills (e.g., Chamot, 2008; Grabe & Stoller, 2011; Yapp et al., 2023). Findings support the positive effects of such instruction on the reading comprehension of L2 learners, which is their ability to process, understand, and interpret written text (e.g., Aghaie & Zhang, 2012; Li et al., 2022; Macaro & Erler, 2008; Yapp et al., 2023). In these studies, researchers implemented reading comprehension tests two or three times during their experiments, tracked changes in participants' test scores, and utilized the results to justify improvements in reading comprehension.

As shown in these previous studies, teachers or researchers prepare multiple reading tests, often administered several times during the intervention, to measure improvement in learners' reading performances. The problem here is that if the difficulty of the tests varies considerably, they cannot be considered valid for examining whether students' reading performance has improved. The purpose of this paper is to address this issue by presenting a five-step solution; more specifically, I describe how I created multiple reading tests for low-proficiency university students while ensuring that the results produced by different texts could be validly compared.

Literature Review

Avoiding Testing Threat and Ensuring Equal Difficulty

In pre-post and other repeated-measures research designs, researchers aim to evaluate the effectiveness of an intervention or treatment by comparing test scores collected at multiple time points (Mackey & Gass, 2015). In the case of reading comprehension tests, if participants' posttest scores are higher than the pretest scores, this result can be interpreted to mean that the intervention improved the participants' reading comprehension. Because the comparison of test scores measured before and after an intervention is essential, Yapp et al. (2023) suggested that "the instruments of measurement must be of equal difficulty, due to the fact that we [researchers] wish to rule out differences in difficulty as a possible explanation for observed differences" (p. 10).

Ensuring that the instruments are of equal difficulty is easy to say but difficult to achieve. The easiest approach might be to use identical test items at each time point, which enables researchers to conduct direct comparisons. However, this approach can cause what is called a "testing threat" that negatively affects validity (Trochim et al., 2016). For example, by utilizing identical reading test items, participants might remember the content of the pretest reading texts and items, which lowers the difficulty of the posttest. The participants might also remember their answers to the pretest questions and repeat them, which hinders measurement accuracy.

Considering the testing threat described above, researchers should use tests consisting of different texts and questions. However, this approach also has an issue that needs to be addressed. If the difficulty level varies considerably depending on the text and questions used, the test results might not be comparable for examining whether students' reading performance has improved. More specifically, if the posttest is easier for participants than the pretest, they will obtain higher posttest scores, but this result does not necessarily indicate an improvement in their reading comprehension.

Reading Comprehension Tests Used in Previous Studies

Researchers of prior studies have utilized various types of reading comprehension tests in pre-post or repeated-measures research designs. According to Grabe and Yamashita (2022), both standardized assessment instruments and researcher-developed measures can be used for research purposes. Still, as far as previous reading strategy studies targeting junior college or university students are concerned, researchers have frequently used standardized English tests. For instance, Yapp et al. (2023) used the Cambridge Advanced English (CAE) reading comprehension tests (Cambridge University Press & Assessments, n.d.) for first-year university students (aged 17–22) in the Netherlands whose English proficiency was at the CEFR B2 level. Shih and Reynolds (2018) used the intermediate level of the General English Proficiency Test (GEPT), estimated as the CEFR B1 level by the Language Training and Testing Center (LTTC, n.d.), for Taiwanese first-year junior college students (aged 16–17). Li et al. (2022) used the reading comprehension section of the College English Test Band 4, which is a national standardized test delivered by the National College English Testing Committee (Cheng & Curtis, 2010), for first-year Chinese university EFL students (aged 17–21). In the Japanese context, Hayashi (1999) used the scores of the TOEFL Institutional Testing Program (TOEFL ITP®; ETS, n.d.) for Japanese second-year university students (their estimated ages were 18–20). Their pretest TOEFL ITP® scores ranged from 451 to 497, which is regarded as intermediate English proficiency (i.e., CEFR A2 to B1 level).

The use of standardized tests is considered reasonable because they are "far more constrained by concerns of validity, reliability, time, cost, usability, and consequence" (Grabe & Yamashita, 2022, p. 465). However, Grabe and Yamashita (2022) also emphasized that standardized tests and their tasks are not necessarily valid for populations at much higher and lower proficiency levels. Indeed, considering that 51.30% of Japanese senior high school graduates are at the CEFR A1 level (MEXT, 2023) and that

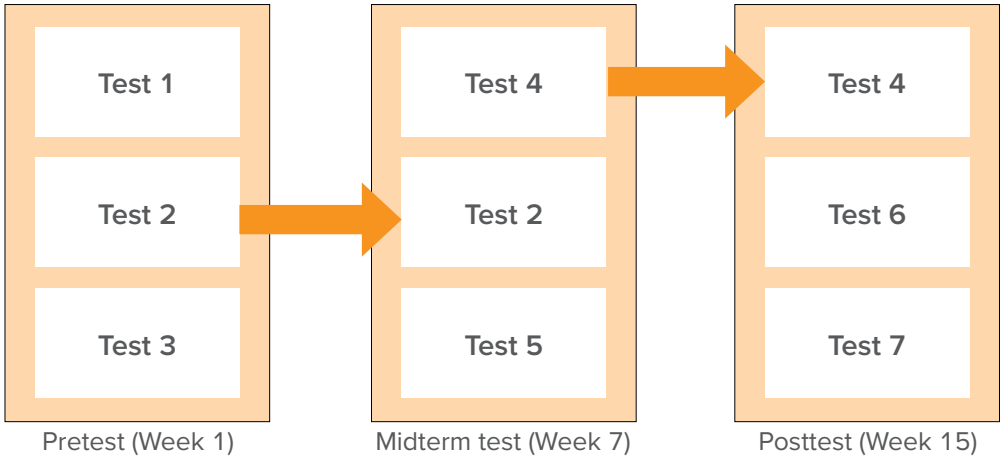
many Japanese university students find it difficult to keep up in their English classes due to their lack of basic English skills, standardized tests might not always function well, especially for low-proficiency students. In other words, standardized tests, such as those employed in prior studies, are too difficult for low-proficiency learners, preventing them from demonstrating improvements in reading comprehension over the course. This problem highlights the need for tailored assessments. The purpose of this paper is to demonstrate how to create valid reading comprehension tests while avoiding a testing threat and ensuring equal difficulty.

A Proposed Five-Step Method for Creating Reading Comprehension Tests

To illustrate this method, I present a detailed example using the following context: assessing reading comprehension changes among low-proficiency Japanese EFL university students (CEFR A1 level, typically aged 18–22) enrolled in a compulsory 15-week English reading course (90 minutes per week). The assessment design involves administering reading comprehension tests at three time points—Week 1 (pretest), Week 7 (midterm test), and Week 15 (posttest)—to measure improvement over the semester (see Figure 1).

Figure 1

Reading Comprehension Tests



Note. Tests 2 and 4 are used for anchoring.

As Figure 1 shows, the testing process involves three phases: the pretest, the midterm test, and the posttest, each with a set of three reading passages and comprehension questions. The pretest (Tests 1, 2, and 3) is administered in Week 1, the midterm test (Tests 2, 4, and 5) is administered in Week 7, and the posttest (Tests 4, 6, and 7) is administered in Week 15. Tests 2 and 4 are used twice as part of a Rasch item anchoring technique called common-item linking (Bond et al., 2021) using the Rasch model (Rasch, 1960) to enable the difficulty estimates of two different sets of items (e.g., the pretest and midterm test) to be plotted on a single measurement scale. The anchoring technique is described in detail in Step 5 below.

Step 1: Selecting the Reading Passages Used for the Tests

The reading passages on each test were taken from *Climate Change* (Newbolt, 2009), a Stage 2 book in the *Oxford Bookworms Library*, a series of graded readers published by Oxford University Press. These books are categorized into seven levels, from Starter to Stage 6, corresponding to CEFR levels, IELTS band scores, and the grades of the Eiken English Proficiency Test in Japan (Oxford University Press, 2024). The Stage 2 books are at the CEFR A2 level and the Eiken Pre-2 Grade English Proficiency Test—the benchmark of MEXT's (2023) survey on the English proficiency of senior high school graduates. Burrows (2012) suggested that expository passages are "deemed appropriate for the types of questions introduced on the reading comprehension test" (p. 103). Following this suggestion, among the Stage 2 books in this graded readers series, seven passages (Table 1) were chosen from *Climate Change* (Newbolt, 2009) because they are expository passages. In addition, the topic of climate change is considered familiar to Japanese students regardless of gender or major.

Table 1

Seven Passages Used for Tests 1 to 7

Test	Title	Pages
1	Getting warmer	8–11
2	Wetter—and drier	12–16
3	Extreme weather	17–19
4	How bad will it get?	24–27
5	Is it all bad?	28–31

Test	Title	Pages
6	Carbon	32–34
7	What are our governments doing?	35–38

Note. The seven passages come from Newbolt (2009).

Step 2: Analyzing and Adjusting the Lexical and Readability Level of the Passages

After the seven passages shown in Table 1 were chosen, the difficulty level of each passage was controlled in terms of the number of running words (tokens), text readability, text complexity, vocabulary frequency, and lexical diversity (Table 2). Readability estimates based on the Flesch-Kincaid scale were produced using a function in Microsoft Word. The Rasch-based Lexile text measures, which show the difficulty of a reading text, were produced by employing the Lexile Text Analyzer (MetaMetrics, 2024).

Vocabulary frequency was checked using the New Word Level Checker (NWLC; Mizumoto, 2021) with the Scale of English Word Knowledge—Japanese (SEWK-J) for the following three reasons. First, the NWLC, together with SEWK-J, is optimized for Japanese EFL learning contexts (Mizumoto et al., 2021), which fits the educational context of this study. Second, the SEWK-J is designed to represent learners' actual vocabulary knowledge. Mizumoto et al. (2021) suggested that "when matching learners with texts, we should consider basing test and lexical profilers on what learners *do* know, rather than…what learners *should* know" (p. 10). The purpose of analyzing the texts is to gauge what vocabulary the students comprehend; therefore, the SEWK-J fits the aim of this study. Third, by using the SEWK-J, the vocabulary can be analyzed using finer 500-word bands compared to the 1,000-word bands used by other profilers, such as VocabProfilers (Cobb, n.d.) and the Lexical Frequency Profile (LFP; Laufer & Nation, 1995). These finer word frequency bands allow researchers to equalize vocabulary difficulty across multiple reading passages more precisely.

Moreover, the Moving Average Type-Token Ratio (MATTR; Covington & McFall, 2010) and the Measure of Textual Lexical Diversity (MTLD; McCarthy & Jarvis, 2010) were also measured by employing the Tool for the Automatic Analysis of Lexical Diversity (TAALED, version 1.4.1) (Kyle et al., 2021). The MATTR is the average of all type-token ratio (TTR) values across the text by using a 50-word window, whereas MTLD represents the average number of words required for the text to reach a point of TTR stabilization (Covington & McFall, 2010; Kyle et al., 2021; McCarthy & Jarvis, 2010). Higher scores of MATTR and MTLD indicate greater lexical diversity.

As a result of these analyses, it was found that Tests 2 and 5 initially appeared easier than the other five tests. More specifically, Tests 2 and 5 had Flesch-Kincaid readability scores of 5.20 and 5.50, respectively, whereas those of the other five tests ranged from 6.30 to 7.50. Test 2 also had lower Lexile ranges (410-600) compared to the other tests. Moreover, Tests 2 and 5 include higher percentages of high-frequency vocabulary (81.97% and 82.40% in the L1 band, respectively). Based on these multiple indicators falling outside the range of the other tests, I modified Tests 2 and 5. These modifications included using lower-frequency vocabulary, employing synonyms, and making sentences longer with conjunctions. After the modifications, the difficulty of the seven test forms was similar (Table 2).

Step 3: Creating Question Items Based on the Difficulty Level of Questions

After controlling the difficulty level of each passage, I created eight comprehension questions for each passage; therefore, a total of 24 questions were included in the pretest, midterm test, and posttest. Lumley (1993) suggested that the difficulty of comprehension questions varies depending on the type of question asked. More specifically, the questions asking learners to identify explicitly stated information in a text are easier than those asking the learners to identify and synthesize ideas or to draw an inference. Based on the difficulty of particular item types suggested by Lumley (1993) and Burrows (2012), eight questions were written that covered the following four types of questions:

- Two questions asking for specific information clearly stated in the reading passage (e.g., *wh*-questions starting with *When* and *Where*).
- Two questions asking for less specific information stated in the reading passage (e.g., *wh*-questions starting with *Why* and *What causes*).
- Two questions asking for information not directly stated in the reading passage so that students are required to draw an inference.
- Two questions asking for the main idea of individual paragraphs or the whole reading passage.

Table 2

Results of Analyzing the Passages Used for the Seven Tests

Original		Test 1	Test 2		Test 3	Test 4	Test 5		Test 6	Test 7
			Original	Modified			Original	Modified		
Flesch-Kincaid readability[a]		7.50	5.20	6.30	7.30	7.50	5.50	6.30	6.80	7.50
Lexile Text Analyzer[b]		610-800	410-600	610-800	810-1,000	810-1,000	610-800	610-800	610-800	810-1,000
New Word Level Checker[c]										
Token (Frequency)		577	588	603	570	601	551	598	563	572
SEWK-J (Freq. %)	L1 (1-500)	73.31	81.97	81.09	72.63	74.54	82.40	80.94	77.80	73.25
	L2 (501-1,000)	6.07	7.31	7.96	9.47	10.82	7.44	8.19	5.33	6.29
	L3 (1,001-1,500)	3.12	1.02	1.33	4.04	2.00	3.63	3.18	4.09	4.20
	L4 (1,501-2,000)	0.52	0.85	0.83	0.88	0.50	0.54	1.00	0.71	1.05
	L5 (2,001-2,500)	1.73	1.53	1.66	0.18	0.67	0.91	0.84	2.49	0.52
	L6 (2,501-3,000)	0.00	0.51	0.50	0.18	0.00	0.18	0.17	0.00	0.00
	L7 (3,001-3,500)	0.35	0.51	0.50	0.18	0.17	0.00	0.00	0.00	0.35
	L8 (3,501-4,000)	0.00	0.34	0.50	0.00	0.00	0.00	0.17	0.00	0.00
	L9 (4,001-4,500)	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.17
	L10 (4,500-5,000)	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.18	0.00
	Over 10	0.35	0.85	0.83	0.70	0.50	0.00	0.33	0.18	0.52
	Proper noun & No.	14.56	5.10	4.81	11.58	10.48	4.90	5.18	9.24	13.64
TAALED[d]										
MATTR		.67		.67	.71	.70		.73	.68	.73
MTLD		34.66		31.44	44.44	44.59		48.32	37.63	49.03

Note.

[a] Flesch-Kincaid Readability Statistics were obtained by Microsoft Word function.

[b] Due to the word limit of Lexile Text Analyzer (MetaMetrics, 2024), the first 500 words of each text were analyzed.

[c] The number of tokens (running words) and vocabulary frequency are analyzed by using the New Word Level Checker (NWLC; Mizumoto, 2021) with the Scale of English Word Knowledge—Japanese (SEWK-J).

[d] TAALED = the Tool for the Automatic Analysis of Lexical Diversity (Kyle et al., 2021); MATTR = Moving Average Type-Token Ratio (50-word window); MTLD = Measure of Textual Lexical Diversity

Step 4: Conducting Alpha and Beta Testing

After creating the test items, prototyping was pursued. Fulcher and Davidson (2007) explained that prototyping consists of two parts: *alpha testing* conducted by experts and *beta testing* by a group of test-takers regarded as representatives of the future test-takers. First, alpha testing was conducted to judge whether the test items were adequate and that there were no problems in terms of context, correct responses, and linguistic sophistication. In alpha testing, a small number of test items are usually required (Fulcher & Davidson, 2007); however, to enhance the validity of all the questions, I asked two Japanese native speakers who teach English to Japanese university students to check whether any questions were unclear and whether the difficulty level increased from Question 1 to Question 8 in each test. Both teachers confirmed that Japanese students would clearly understand all test items and that test items become more difficult from Question 1 to Question 8.

Second, beta testing was conducted to determine whether test-takers understood each item appropriately, whether any unexpected or illogical responses occurred, and whether any items required revision or elimination. Thirty Japanese EFL university students at the CEFR A1 level completed Tests 1 to 7. The data obtained from these 30 students were entered into a Microsoft Excel spreadsheet, exported to Winsteps 4.4.7 (Linacre, 2022), and analyzed using the Rasch rating scale model (Andrich, 1978).

The Rasch rating scale model (Andrich, 1978) can be used to examine whether test items function properly and whether any items are candidates for modification or for being discarded (Bond et al., 2021). Therefore, the Rasch item reliability estimate and the fit of each item to the Rasch model were inspected for the pretest, midterm test, and posttest. Fisher (2007) established the following criteria: item reliability estimates are classified as excellent (> .94), very good (.91–.94), good (.81–.90), or fair (.67–.80); whereas item infit mean square (MNSQ) values are considered excellent (0.77–1.30), very good (0.71–1.40), or good (0.50–2.00).

According to Fisher's (2007) criteria, the item reliability estimate and the fit of each item were both evaluated as good in all the pre-, midterm, and posttests (Table 3). The item reliability estimate of the pretest was .84 and the infit MNSQ statistics ranged from 0.65 to 1.37, both of which were considered good. The item reliability estimate of the midterm test was .80 and the infit MNSQ statistics ranged from 0.65 to 1.43, which was also evaluated as good fit to the Rasch model. The posttest had an item reliability estimate of .80 and the infit MNSQ statistics ranged from 0.68 to 1.35, also suggesting good fit. Overall, the results indicated that the three tests functioned appropriately for measuring students' reading comprehension.

Table 3

Item Reliability and Infit MNSQ Values of the Pretest, Midterm, and Posttests

Test	Tests included	Item reliability	Infit MNSQ
Pretest	Tests 1, 2, and 3	.84	0.65 to 1.37
Midterm test	Tests 2, 4, and 5	.80	0.65 to 1.43
Posttest	Tests 4, 6, and 7	.80	0.68 to 1.35

Note. Tests 2 and 4 serve as anchoring items. MNSQ = mean square

Step 5: Employing Rasch Analysis to Ensure Comparable Difficulty Estimates

As shown in Figure 1, a pretest, midterm test, and posttest will be conducted in the context described above. Each test consists of three reading passages with eight questions each for a total of 24 questions. The pretest is made up of Tests 1, 2, and 3, the midterm test includes Tests 2, 4, and 5, and the posttest is made up of Tests 4, 6, and 7.

The questions accompanying Tests 2 and 4 serve as anchor items (Bond et al., 2021) that ensure the comparability of the person ability estimates produced by each test form. When students complete the three reading tests, the difficulty of each test inevitably differs; thus, more than a simple comparison of the raw scores is needed to adequately estimate changes in reading comprehension over time. The Rasch rating scale model (Andrich, 1978) will be employed to address this issue because it provides person estimates, difficulty threshold estimates for each item, and a single "rating scale threshold structure that is common for all of the items" (Bond et al., 2021, p. 97). In other words, by utilizing difficulty estimates obtained from the Rasch analysis instead of the raw test scores, the students' pretest, midterm test, and posttest results can be placed on the same measurement scale for comparison.

The Rasch model was used to determine whether the questions on Tests 2 and 4 can serve as anchor items. For example, to examine whether Test 2 items function appropriately as anchors, I obtained the difficulty estimates for the Test 2 questions from the initial pretest analysis (see Step 4 above) and utilized those difficulty estimates for the midterm test analysis using the Winsteps IAFILE (Item Anchor File) command (Linacre, 2025, p. 143). Researchers have pointed out that displacement values should be less than 0.50 logits to ensure that the test items are functioning adequately well as anchors (O'Neill et al., 2013). As Tables 4 and 5 show, the displacement values of all eight items of

Table 4

Rasch Descriptive Statistics for Test 2 Anchor Items

Item	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Displacement
T2Q1	-1.28A	0.56	1.02	0.18	0.93	0.13	-0.06
T2Q2	-2.11A	0.74	1.11	0.38	1.19	0.53	-0.07
T2Q3	-0.99A	0.52	1.02	0.16	1.18	0.49	-0.06
T2Q4	-0.12A	0.44	1.32	1.54	1.78	1.91	-0.03
T2Q5	1.57A	0.42	1.07	0.46	1.01	0.13	0.02
T2Q6	0.24A	0.42	1.06	0.40	0.92	-0.19	-0.02
T2Q7	0.74A	0.41	0.80	-1.27	0.71	-1.29	0.00
T2Q8	0.57A	0.41	1.06	0.40	1.03	0.19	0.00

Note. T2Q1 stands for Test 2 Question 1. MNSQ = mean square; ZSTD = z-standardized. The "A" in the Rasch measure column indicates anchor items.

Table 5

Rasch Descriptive Statistics for Test 4 Anchor Items

Item	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Displacement
T4Q1	-2.96A	1.01	1.03	0.33	1.10	0.54	-0.05
T4Q2	-0.57A	0.47	0.98	-0.02	0.79	-0.37	-0.05
T4Q3	0.39A	0.42	0.81	-1.07	0.76	-0.94	-0.04
T4Q4	0.39A	0.42	0.71	-1.72	0.63	-1.61	-0.04
T4Q5	-0.35A	0.45	0.85	-0.69	0.72	-0.69	-0.05
T4Q6	0.89A	0.41	1.14	0.81	1.16	0.75	-0.02
T4Q7	-0.16A	0.44	0.74	-1.37	0.60	-1.26	-0.04
T4Q8	0.21A	0.42	1.21	1.12	1.24	0.90	-0.04

Note. T4Q1 stands for Test 4 Question 1. MNSQ = mean square; ZSTD = z-standardized. The "A" in the Rasch measure column indicates anchor items.

Test 2 ranged from -0.07 to 0.00, and those of Test 4 ranged from -0.05 to -0.02. These results confirmed that the items on Tests 2 and 4 work well as anchors.

Using the five-step sequence as described above, teachers and researchers can equalize the difficulty of reading passages and test items as well as check the validity of test items. With a comprehensive inspection and assurance of validity, the set of reading comprehension tests is now ready for utilization in a repeated-measures design.

Conclusion

The purpose of this paper was to demonstrate how to create multiple reading comprehension tests with the same difficulty to measure learners' improvement in pretest-posttest or repeated-measures research designs. The use of standardized English tests is common (e.g., Hayashi, 1999; Li et al., 2022; Shih & Reynolds, 2018; Yapp et al., 2023). However, when assessing learners with low English proficiency, standardized tests are often too difficult, which makes teacher- and researcher-developed tests necessary.

This paper addresses this issue by describing a method that enhances the validity of self-made reading comprehension tests through a five-step sequence:

1. Select the reading passages used for the tests.
2. Analyze and adjust the lexical and readability level of the passages.
3. Create question items based on the difficulty level of questions.
4. Conduct alpha testing with experts and beta testing with future test-taking populations.
5. Employ Rasch analysis to ensure comparable difficulty estimates.

When assessing learners with low English proficiency, teacher- or researcher-developed measures can be more appropriate than standardized tests (Grabe & Yamashita, 2022); however, teachers and researchers must be careful to account for differences in the difficulty of the reading passages and test items caused by varying lexical and readability levels, because results based on the tests at different difficulty levels are not valid estimates of improvement in students' reading performances. This five-step sequence can provide teachers and researchers with a practical method for assessing learners' reading comprehension more accurately.

Acknowledgements

I would like to thank David Beglar for his suggestions, guidance, and encouragement in compiling this study. My thanks also go to the anonymous reviewers and all who provided me with their insightful comments and feedback on my data and an earlier version of this study.

Bio Data

Sachi Oshima is an associate professor at the Faculty of Law, Chuo Gakuin University, Chiba. She is also a research fellow at the Research Institute for Policy Studies, Tsuda University, Tokyo. After working for the Japan Foundation as a chief officer, she has been an English teacher at several universities in the Kanto area, Japan. Her research interests are related to TESOL and intercultural exchange projects. <oshima.s@mc.cgu.ac.jp>

References

Aghaie, R., & Zhang, L. J. (2012). Effects of explicit instruction in cognitive and metacognitive reading strategies on Iranian EFL students' reading performance and strategy transfer. *Instructional Science*, *40*, 1063–1081. https://doi.org/10.1007/s11251-011-9202-5

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Burrows, L. (2012). *The effects of extensive reading and reading strategies on reading self-efficacy*. [Doctoral dissertation, Temple University]. https://doi.org/10.34944/dspace/868

Cambridge University Press & Assessment. (n.d.). *Cambridge English Qualifications: C1 Advanced*. https://www.cambridgeenglish.org/exams-and-tests/advanced/

Chamot, A. U. (2008). Strategy instruction and good language learners. In C. Griffiths (Ed.), *Lessons from good language learners* (pp. 266–281). Cambridge University Press.

Cheng, L., & Curtis, A. (Eds.). (2010). *English language assessment and the Chinese learner* (1st ed.). Routledge. https://doi.org/10.4324/9780203873045

Cobb, T. (n.d.). *VocabProfilers* [Computer software]. https://www.lextutor.ca/vp/comp

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Eiken Foundation of Japan. (n.d.). *Eiken grades*. https://www.eiken.or.jp/eiken/en/grades/

ETS. (n.d.). *About the TOEFL ITP® assessment series*. https://www.ets.org/toefl/itp/about.html

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095. https://www.rasch.org/rmt/rmt211m.htm

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. Routledge.

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). Routledge. https://doi.org/10.4324/9781315833743

Grabe, W., & Yamashita, J. (2022). *Reading in a second language: Moving from theory to practice* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108878944

Hayashi, K. (1999). Reading strategies and extensive reading in EFL classes. *RELC Journal*, *30*(2), 114–132. https://doi.org/10.1177/003368829903000207

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Li, H., Gan, Z., Leung, S. O., & An, Z. (2022). The impact of reading strategy instruction on reading comprehension, strategy use, motivation, and self-efficacy in Chinese university EFL students. *Sage Open*, *12*(1), 1–14. https://doi.org/10.1177/21582440221086659

Linacre, J. M. (2022). *Winsteps®* (Version 4.4.7) [Computer software]. Winsteps.com. https://www.winsteps.com/

Linacre, J. M. (2025). *A user's guide to Winsteps® Ministep Rasch-model computer programs.* Winsteps.com. https://www.winsteps.com/manuals.htm

Language Training and Testing Center (LTTC). (n.d.). *The general English proficiency test: Intermediate*. https://www.gept.org.tw/Eng/intermediate.html

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, *10*(3), 211–234. https://doi.org/10.1177/026553229301000302

Macaro, E., & Erler, L. (2008). Raising the achievement of young-beginner readers of French through strategy instruction. *Applied Linguistics*, *29*(1), 90–119. https://doi.org/10.1093/applin/amm023

Mackey, A., & Gass, S. (2015). *Second language research: Methodology and design* (2nd ed.). Routledge. https://doi.org/10.4324/9781315750606

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. https://doi.org/10.3758/BRM.42.2.381

MetaMetrics. (2024). *Lexile text analyzer* [Web application]. https://hub.lexile.com/analyzer

Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). (2023). *Reiwa 4-nendo "Eigo kyōiku jisshi jōkyō chōsa" gaiyō* [令和４年度「英語教育実施状況調査」概要] [Results of "Survey on English education" in the fiscal 2022]. https://www.mext.go.jp/content/20230516-mxt_kyoiku01-00029835_1.pdf

Mizumoto, A. (2021). *New word level checker* [Web application]. https://nwlc.pythonanywhere.com

Mizumoto, A., Pinchbeck, G. G., & McLean, S. (2021). Comparisons of word lists on new word level checker. *Vocabulary Learning and Instruction*, *10*(1), 1–12. https://doi.org/10.7820/vli.v10.2.mizumoto

Mochizuki, A., Kubota, A., Iwasaki, H., & Ushiro, Y. (2018). *Shin gakushū shidō yō*ryō ni motozuku *eigo-ka kyōikuhō* [新学習指導要領に基づく英語科教育法] [Teaching English based on the latest course of study] (3rd ed.). Taishukan.

Newbolt, B. (2009). *Climate change*. Oxford University Press.

O'Neill, T., Peabody, M., Tan, R. J. B., & Du, Y. (2013). How much item drift is too much? *Rasch Measurement Transactions*, *27*(3), 1423–1424. https://www.rasch.org/rmt/rmt273.pdf

Oxford University Press. (2024). *Oxford bookworms library*. https://www.oupjapan.co.jp/en/gradedreaders/bookworms.shtml

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

Shih, Y., & Reynolds, B. L. (2018). The effects of integrating goal setting and reading strategy instruction on English reading proficiency and learning motivation: A quasi-experimental study. *Applied Linguistics Review*, *9*(1), 35–62. https://doi.org/10.1515/applirev-2016-1022

Sun, Y., Wang, J., Dong, Y., Zheng, H., Yang, J., Zhao, Y., & Dong, W. (2021). The relationship between reading strategy and reading comprehension: A meta-analysis. *Frontiers in Psychology*, *12*, 1–11. https://doi.org/10.3389/fpsyg.2021.635289

Trochim, W., Donnelly, J.P., & Arora, K. (2016). *Research methods: The essential knowledge base* (2nd ed.). Cengage Learning.

Yapp, D., de Graaff, R., & van den Bergh, H. (2023). Effects of reading strategy instruction in English as a second language on students' academic reading comprehension. *Language Teaching Research*, *27*(6), 1456–1479. https://doi.org/10.1177/1362168820985236