Cloze-elide as a classroom reading test

Trevor A. Holster
holster@fukuoka-u.ac.jp
Fukuoka University

Abstract

The cloze-elide procedure, in which test-takers must identify redundant words that have been added to a reading text, provides a process-oriented test of reading ability. This research investigated the suitability of cloze-elide as a summative classroom test focusing on speeded processing of written text. The cloze-elide format was found to be simple and practical to construct, and Rasch analysis found reliability and data-model fit acceptable for classroom use. However, confirming reports from earlier researchers, test-takers often mistakenly identified correct sentences as containing redundant words, leading to Rasch data-model misfit and the need for a scoring formula to account for test-takers who are excessively cautious or careless in their response patterns.

Keywords: Cloze-elide, cloze test, reading assessment, Rasch analysis, process oriented assessment

A crucial distinction in reading assessment is between product orientation and process orientation (Koda, 2004). Traditional reading comprehension tests treat comprehension as a product of reading, manifested by the ability to recall details of the text, resulting in conflation of comprehension and long-term memory. In contrast, process oriented assessments attempt to assess language processing as it occurs and before information is coded into long-term memory, emphasizing working memory. Of course, test designers can design product oriented tasks with the intention of encouraging specific types of processing, but the resulting inferences are necessarily indirect, so there is a need for assessments that are specifically designed to focus on processing.

Rauding and Process Flexibility

The importance of processing considerations in reading assessment design is implicit in Carver's description of "reading gears" (1992, p. 85), each representing a different reading process: memorizing, learning, rauding, skimming, and scanning. *Rauding*, a portmanteau of *reading* and *auding* (i.e. listening), was seen as the core reading process, with two major components: visual word decoding and comprehension, based on the *simple view* of reading (Carver, 1993; Hoover & Gough, 1990). Although Carver's work focused on first language (L1) reading, evidence of the applicability of the simple view of reading and the rauding model to second language (L2) readers comes from Yamashita and Shiotsu's (2017) finding that L2 listening comprehension was the strongest predictor of L2 reading comprehension.

Carver (1992) reported that rauding, or normal reading, in native speaker university students operated at about 300 words per minute (wpm), and required fluent syntactic parsing in order to decode meaning at the level of sentences and then integrate sentence meanings into a text model. *Scanning*, the identification of target words in a text, operated at 600 wpm or more, while *skimming* operated at 450 wpm in order to generate an overview of the text without syntactically parsing entire sentences. *Learning* typically operated at 200 wpm, with slower processing required to comprehend unfamiliar ideas or language, while *memorizing* operated at less than 150 wpm in order to commit details of a text to long-term memory for later recall. Although Carver (1992) saw rauding as being the key process in reading, proficient readers switch between different processes in response to task demands, hence displaying *process flexibility*. The design of reading assessments should therefore include considerations of the relevant processes for the assessment purpose and the degree of process flexibility required.

Intensive versus Extensive Reading

Related to process flexibility is the distinction between *intensive reading* and *extensive reading*:

> In the former case each sentence is subjected to a careful scrutiny, and the more interesting may be paraphrased, translated, or learnt by heart. In the latter case book after book will be read through without giving more than a superficial and passing attention to the lexicological units of which it is composed (Palmer, 1917, p. 205).

Rauding is the core process practiced in extensive reading (Yamashita, 2008), while learning and memorizing represent intensive reading processes. However, speaking anecdotally, second language reading textbooks often put little emphasis on rauding, instead focusing on introducing new language embedded in short reading passages and skimming and scanning tasks to practice reading strategies. Reading tasks often take the format of tests such as the superseded pencil-and-paper *TOEFL* (ETS, 2008), the reading section of which comprised short passages followed by traditional product oriented comprehension questions. Reading instruction in Japanese classrooms is still dominated by traditional methods such as grammar translation, which focus on slow and deliberate analysis of assessment products instead of fluent language processing, resulting in many students assuming that reading English means translation into Japanese (Maruo, 2012; Sakurai, 2015). The core process of rauding may therefore be overlooked by teachers and students due to focus on intensive linguistic analysis, contributing to *readicide*, the "systematic killing of the love of reading, often exacerbated by the inane, mind-numbing practices found in schools" (Gallagher, 2009, p. 2).

Reading Speed Assessment

The use of process-oriented classroom assessment is one potential way to encourage students to practice fluent rauding and develop implicit procedural knowledge rather than focusing on explicit linguistic analysis. Ideally, classroom assessment tasks should be closely linked to the content of the teaching curriculum, be criterion referenced to curriculum goals, and have formative benefits (H. D. Brown, 2004; J. D. Brown, 2005; Hughes, 2003). Quinn, Nation, and Millett (2007) provided an example of such a task in the form of reading speed tests. Rather than focusing exclusively on the product of reading comprehension, Quinn et al. (2007) provided 20 reading texts of approximately 500 words each, limited to very high-frequency vocabulary and simple grammatical structures. Students were required to read as quickly as possible while maintaining adequate comprehension, and then to record their reading time. After having finished reading, students answered 10 comprehension check questions. However, the comprehension check questions were not the primary task, but rather a secondary task used to encourage self-assessment by students as to whether their comprehension was adequate. Answering all the comprehension check questions correctly would suggest reading too slowly in order to memorize details, so students were advised that adequate comprehension only required answering a majority of questions correctly and that it was not desirable to try to memorize all the details of the texts.

Although self-reported reading speed is useful as both a formative classroom assessment and a research tool (e.g., Beglar, Hunt, & Kite, 2012), classroom teachers are typically required to provide summative end-of-course assessments that determine whether students are awarded passing or failing grades. This makes it desirable to use assessment tasks that are suitable for both formative and summative purposes. The self-reported reading speeds provided by Quinn et al. (2007) are unsuitable for summative purposes, while using the comprehension check questions for this purpose would lead to the same problems of product-orientation that raise concern with other traditional tests. This makes it desirable to use a process oriented task that can provide both summative scores for use as final grades and also provide formative benefits through encouraging rauding practice. One format that has potential for this is the cloze-elide (CE) procedure, also called an intrusive words test (Clapham, 1996) and an editing test (Bowen, 1978).

Cloze-elide Tests

Two reports by Davies (1967, 1975) on the English Proficiency Test Battery (EPTB), introduced in 1964 (Clapham, 1996), provided the first documented analysis of CE as a test of speeded reading. In a standard cloze format test, words are deleted from a text and test-takers must provide an acceptable word to complete each gap. In CE tests, words are added to the text and test-takers must delete, or elide, the redundant words, or planned items (PI) while not eliding the other words, or unplanned items (UI), a format which results in every word in the text functioning as a test item, but also raises the question of how best to combine the PI score and UI score into a single total. Davies' (1975) analysis of different EPTB administrations found CE section reliability consistently above .90, with correlations between CE and a speeded cloze test ranging from .5 to .7 for different populations. Each PI located was scored as 1 point, with 1 point deducted for each UI marked incorrectly, up to a maximum of 4 deductions per line of text. Davies (1975) reported that PIs were located at random locations in the text, but provided no details of how this was done.

Bowen (1978) argued for CE as an integrative test of reading, using manually scored pencil-and-paper tests, with three dice used to determine the interval between PIs. Scoring was done by subtracting points for mistakes on either PIs or UIs. Administered to 145 applicants to an Egyptian university, Bowen’s test of 40 PIs and 450 UIs produced KR-21 reliability of .95 and a useful level of discrimination between high-ability and low-ability test takers. It also showed correlations of approximately .70 with the overall admissions test battery, presented as evidence of construct validity as a measure of general language proficiency. However, Bowen’s results left concerns about the functioning of the UI items, with 55 selected by more than one test-taker, 109 selected by only one, and 286 not selected at all. This suggests that some UI items may function systematically because of language features that confuse test-takers, others may just reflect random careless responses, and the majority play no role in the results. Trace, Brown, Janssen, and Kozhevnikova (2017) investigated item difficulty in standard cloze tests, finding that a complex range of variables contributed to difficulty. It is reasonable to hypothesize that a similarly complex range of features will affect the functioning of CE items. It is possible that PI and UI items address different constructs related to comprehension and confusion, raising concerns about whether Bowen’s (1978) scoring system could adequately summarize such a complex construct.

Manning (1987) developed machine readable CE tests, with a table of random numbers ranging from seven to 14 used to determine the location of PIs and optical scanning used to detect the pencil marks indicating elided words. Manning noted that some redundant words were more obviously incorrect than others, meaning that item difficulty would be expected to vary independently of the overall difficulty of the text, consistent with Trace, et al. (2017). Manning also introduced a scoring formula to address the issue that CE produces two partially dependent scores: the number of correct elisions and the number of incorrect elisions. The resulting scoring formula produced scores of 0 for the limiting cases of test-takers who elided every word in the text or who elided no words, and an expected average score of 0 for purely random responses. Manning also found high intercorrelations between CE scores from different reading passages, with scores from a multiple-choice cloze, and with TOEFL scores. Factor analysis and correlational results showed a "systematic pattern that is strongly supportive of the validity of cloze-elide tests" (Manning, 1987, p. 60), with a higher correlation with a reading/structure factor than with listening.

Klein-Braley (1997) used Manning’s (1987) CE tests in a similar investigation of a battery of tests, finding a moderate level of reliability. However, Klein-Braley (1997) used manual scoring rather than the machine scoring used by Manning. Coupled with the need to score both PI items and UI items, this was found to be a major constraint on test usefulness. Elder and von Randow (2008) addressed scoring practicality in a study of CE as part of a battery of diagnostic tests by using a computer administered test with one redundant word per line of text. The computer administration removed the need for manual scoring and this format meant that every line of text functioned as a PI item, eliminating the need to score UI items. As well as finding an acceptable level of reliability, Elder and von Randow (2008) argued for construct validity for academic placement on the grounds that CE promotes rapid skimming and scanning that is essential for academic reading.

Background to this Study

This study was conducted in the second year of a two-year English program at a private Japanese university. Three class groups taking a course titled *ESP* were included in this study. The university's official objective for ESP classes was to prepare students to take tests, but no specific test was identified as a target, meaning that focusing on any specific test would be of questionable benefit to students desiring to prepare for a different test. As this was my first time to teach this course, I checked other teachers' course outlines on the university website, which showed that *TOEIC* (ETS, 2013) practice textbooks were typical, although not universal. Also, although second-year students could choose from a range of English courses with standardized course titles and official objectives, studying English was compulsory and informal discussions with students suggested confusion about the choice of second-year classes and little awareness of the objectives of the different courses. This made it prudent to assume that most students would be motivated primarily to achieve required course credits, rather than any real need or desire to take English tests. Course content was therefore based on a textbook focusing on general listening and communication, supplemented by reading texts from Quinn et al. (2007). Both the listening and reading texts included comprehension check questions that were generically similar to those found in standardized tests, but most class time was spent on process oriented tasks such as partial dictation and listening cloze tasks in order to encourage fluent auding and rauding rather than explicit linguistic analysis. Thus, instead of practicing for any specific test, classroom tasks focused on language processing that is required for success in any test of listening or reading.

Weekly review tests of listening and vocabulary based on course content accounted for 40% of final grades, with another 30% determined by an official final exam prepared by the classroom teacher but administered by the university. The final 30% of grades was determined by a classroom test administered in the final week of class. This comprised three sections of 10% each: listening, vocabulary, and reading. The listening and vocabulary sections repeated the weekly in-class quizzes, while the reading section comprised new CE tests based on the weekly readings. Although classroom activities had focused on speeded tasks that integrated listening and reading in order to promote rauding, none of the formats used was considered practical for use as a summative assessment. CE was therefore used for the reading test on the grounds of being easy to construct and being process oriented. However, the existing research on CE investigated its use as in proficiency tests, so evidence was required of its validity for the different purpose of a classroom achievement test. Also, not having used this format for summative purposes before, questions arose about practicality and scoring. In particular, machine readable tests were essential for practicality, and investigation of the contribution of PIs and UIs was necessary to provide evidence of validity. This study therefore focuses on the psychometric characteristics of the resulting CE tests and concerns over the effect of the UI items.

Research Questions

1. Does the CE task provide suitable psychometric performance for use as a summative classroom test?
2. Should UI items be included or excluded from scoring for summative purposes?
3. Is formula scoring desirable for summative course grades?

Method

Participants

Participants were drawn from convenience sampling of three class groups taking compulsory English classes at a Japanese university. Of the 124 students enrolled in the course, 108 attended the final exam, 64 males and 44 females.

Design

Test specifications.

Practicality demanded machine readable tests, so a revised pencil-and-paper format was developed. A randomization process was used to select the locations of PIs, in contrast to Elder and von Randow's (2008) format, where every line of text contained a PI. Students were only required to identify lines of text that contained redundant words, not the specific words. These lines of text functioned as PIs while the unchanged lines functioned as UIs. A sample test is shown in the Appendix. Test forms were based on 10 readings from Quinn, et al. (2007) which were familiar to students, as they had been used in class as reading speed practice tasks and tests. These were numbered as readings 11 to 20, but followed a different order than originally used by Quinn, et al. (2007). Redundant words were added at locations determined by random numbers generated by Microsoft Excel. Line spacing between insertions was determined by a random number from 1 to 4, with 1 indicating an insertion on the next line and 4 indicating a jump to the fourth line. The location within each line was then determined by a random number from 1 to 20, with 1 indicating an insertion before the first word of the line, 2 before the second word, and so on. When this second random number exceeded the number of words in the line, counting continued to the next line, meaning that the number of lines between insertions could exceed 4.

Inserted words were sourced from Davies and Gardner's (2010) *Frequency Dictionary of Contemporary American English*, which lists 5000 words in order of frequency. The first insertion, "trim" on line 4 of the example test in the Appendix, was the 5000th word listed, the second insertion, "electronics" on line 7 was the 4999th word, and so on, in ascending order of frequency. However, in cases where an inserted word was not unambiguously incorrect, it was skipped and the next highest frequency word used instead. The second of the 10 texts used words starting from the 4900th most frequent, the third text starting from 4800, and so forth, with the 10th text therefore starting from the 4100th most frequent word. This meant that the inserted words were of much lower frequency than the overall texts. As the unmodified texts had all been provided to students as class handouts, students who had revised the reading material for homework should have been able to achieve a high score just by eliding any unfamiliar vocabulary.

Test administration.

Demonstration and practice tests.

The sample test shown in the Appendix, based on Reading 11, was used as a demonstration test in class one week before the final test. Students were provided with a paper copy of the test while the teacher demonstrated the procedure on a projector. First, students were instructed to read the text quickly and elide any redundant words (i.e. PIs). After finishing the entire text, they were then told to mark the answer bubble in the right-hand margin on any line where they had elided a word, leaving the answer bubble untouched on lines with no elisions. To ensure that all students understood the procedure thoroughly, students were then instructed to work in small groups to compare and discuss their practice test sheets, then to change groups and compare again. Students were allowed to take this demonstration test form home to review. Following the demonstration, a practice test was administered to check how long students took to complete an operational test. Three different readings, based on Readings 12 to 14, were used for practice, with each student administered a single reading. The fastest students completed the test in less than five minutes, with most comfortably finishing in less than 10 minutes.

Operational tests.

The operational test was administered as part of the final classroom test in the last week of class, with the CE section following a listening section and vocabulary section. The listening and vocabulary sections repeated weekly classroom tests in one long test that was expected to take about 60 minutes to complete, leaving about 25 minutes for the CE section. However, the answer sections of the listening and vocabulary sections required scanning lists of words to match sounds and meanings, so students with faster word recognition ability would be left with more time for the CE section. The CE section was therefore expected to reward students with higher reading speed or who had reviewed the reading material for homework.

Each student was administered three of the nine operational readings. Each test booklet included one of the readings used the previous week as practice tests (Readings 12 to 14), plus two new readings based on Readings 15 to 20. To prevent cheating, nine different test booklets were produced, following the pattern shown in Table 1. This administration pattern provided linking between the different readings to ensure that all test forms could be equated using Rasch analysis, despite no student taking all items. Readings 12 to 14, which had been used as practice tests, were included to allow the possibility of using the practice tests for final grades in the eventuality that transportation delays or sickness caused students to miss the final test, although this did not prove to be necessary.

Table 1
Test administration pattern

	Reading								
	12	13	14	15	16	17	18	19	20
Test form 1	*			*					*
Test form 2	*				*			*	
Test form 3	*					*	*		
Test form 4		*		*					*
Test form 5		*			*			*	
Test form 6		*				*	*		
Test form 7			*	*					*
Test form 8			*		*			*	
Test form 9			*			*	*		

Data collection and analysis.

Scoring and coding.

Completed test forms were scanned and scored using Remark Office OMR version 8.4 (Gravic, 2012), which recognized any answer bubbles that had been marked and also read a bar code that identified each test form. Remark Office OMR recorded shaded answer bubbles as “1” and unshaded answer bubbles as

"Error". This data was then exported to Microsoft Excel for reformatting prior to Rasch analysis using Winsteps version 3.91.2 (Linacre, 2017). "Error" codes, indicating an unshaded response bubble, were recoded as "0". As each line of text functioned as an independent item, with lines containing redundant words being PIs and those without being UIs, the Winsteps answer key was coded as “1” for PIs and “0” for UIs. All analyses were conducted using Winsteps’ default estimation settings for the dichotomous Rasch model, with mean item difficulty specified as 0.00 logits.

Rasch analysis and data-model fit.

Rasch analysis was used due to several major benefits over classical test theory (CTT) analysis (Bond & Fox, 2015; Sick, 2008). The most obvious of these is that item difficulty and person ability are both provided as log-odds units, or *logits*, which provide equal-interval measures, unlike the raw percentage scores used in CTT. Logits are probabilistic units, that allow the probability of a person succeeding on an item to be calculated from the difference between the person's ability and item's difficulty. When person ability is precisely equal to item difficulty (i.e. a difference of 0.00 logits), the person has a 50% expectation of success. This increases to approximately 73%, 88%, and 95% when person ability exceeds item difficulty by 1, 2, or 3 logits, respectively, while it falls to approximately 27%, 12%, and 5% when item difficulty exceeds person ability by 1, 2, or 3 logits. Logits thus allow item difficulty and person ability to be directly compared on a shared measurement scale, making it very easy to determine if test items are of suitable difficulty for students.

Another practical benefit of Rasch analysis is that logit measures are sample independent, meaning that logit scores still represent the same measurement scale even for students who took different test forms, provided there is sufficient connectivity within the dataset to link all persons and items. This allows different test forms to be combined into a single analysis and student grades to be equated between different test forms.

A further benefit of Rasch analysis is the provision of diagnostic fit statistics showing how closely the observed data match the theoretical ideals of measurement (Bond & Fox, 2015; Wright & Stone, 1979). Fit statistics are provided as both information weighted *infit* statistics that are the more important indicator of whether measurement quality has been compromised, and unweighted *outfit* statistics that show the effect of outlying responses (i.e. when low-ability persons succeed on difficult items or high-ability persons fail on easy items). Both infit and outfit are provided as standardized values, indicating whether the misfit is statistically significant, and mean-square values, showing the substantive size of the misfit. Mean-square values have an expected value of 1.00, indicating that the data perfectly match Rasch model expectations, with a minimum possible value of 0.00 and no upper limit. Following Linacre’s (2014) guidelines, mean-square values below 1.50 are productive for measurement, values between 1.5 and 2.0 are not degrading, and values greater than 2.0 are degrading for measurement.

A somewhat counter-intuitive concern is the effect of *overfit*, caused by unexpectedly predictable response patterns. In contrast to CTT, where extremely predictable response patterns, manifested by high item discrimination, are seen as a theoretical ideal (J. D. Brown, 2005), overfitting items in Rasch analysis indicate redundancy in the dataset. These excessively predictable response patterns mean that some items or persons contribute less information than modelled, resulting in measurement being muted (Linacre, 2014). Further to this, because the average mean-square value is constrained to a value close to 1.00, overfitting items or persons will entail that other items or persons misfit, as illustrated by Holster and Lake (2016), because unusually predictable items or persons make other items or persons unpredictable in comparison. It is therefore too simplistic to consider misfit as undesirable and overfit as desirable. Instead, Rasch data-model fit statistics provide quality control tools showing the degree of distortion of the measurement rulers and allow diagnosis of the source of the misfit. Due to the small sample size and low-stakes nature of this classroom assessment, mean-square values between 0.50 and 1.50 were considered to be well-fitting, and values above 2.0 to be seriously misfitting, following Linacre's (2014) guidelines.

Formula scores.

Institutional constraints required teachers to retain test papers and records of grade calculations to allow students to contest grades. With the criterion for a passing grade set at 60%, the potential need to explain grade calculations to non-specialists made it preferable to use raw scores for the calculation of grades rather than rescaled logit measures, which would be difficult to explain to non-specialists. Because the number of UIs greatly exceeded the number of PIs, students who simply left all response bubbles blank would be rewarded with a high score if the number of correct UIs and PIs were simply added together. However, if only PIs were scored, then students who simply marked every response bubble would score 100%, making a simple but defensible scoring formula necessary. Formula scores were calculated by subtracting the proportion of incorrectly answered UI items from the proportion of correctly answered PI items, using the equation:

$$F = P - (1 - U) \tag{1}$$

Where: F = Formula score
P = Proportion of PI items answered correctly
U = Proportion of UI items answered correctly
(1 – U) = Proportion of UI items answered incorrectly

Equation 1 can be expressed as the more convenient, but mathematically equivalent:

$$F = P + U - 1 \tag{2}$$

This scoring formula would result in a formula score of 0 for the two limiting cases of extreme caution, where a candidate did not respond to any item (i.e. $P = 0$, $U = 1$), and extreme guessing, where a candidate marked all response bubbles (i.e. $P = 1$, $U = 0$). It would also give an expected average score of zero for uninformed guessing, but with a range of positive and negative scores due to random chance.

Results

Research Question 1: Psychometric Performance

Summary statistics for the analysis of persons and items are respectively given in Table 2 and Table 3, which both show results for the analysis of all items and of the PI items only. Mean item difficulty was anchored at 0.00 logits, following conventional Rasch practice. Logit measures of person ability in Table 2 increased from -0.25 logits for the 110 PI items to 1.44 logits for all 267 items, reflecting that the UIs were substantively easier than the PIs. Table 4 shows the mean difficulty of the UIs and PIs, respectively -1.17 logits and 1.68 logits, a statistically significant difference of 2.85 logits, $t(263) = -30.01$, $p < .001$, $r = .88$. This is a substantively large difference, a person with an expectation of success of 50% on an average PI item having an expectation of success of approximately 95% on an average UI item.

This is easily seen in Figure 1, which maps person ability against item difficulty. The PIs, coded as "P", showed a range of difficulty roughly matching the range of person ability, although there are a considerable number of persons with ability lower than 1 logit, but relatively few items at this level. The UIs, coded as "*", mostly fell far below the ability of any person, suggesting that they contributed little to measurement. This is supported by Table 2, which shows the same reliability coefficient of .89 for both analyses, contrary to the expectation that a test of 267 items would show considerably higher reliability than a test of 110 items (J. D. Brown, 2016). This level of reliability is excellent for a relatively short classroom assessment, corresponding to person separation of 2.85, indicating very high confidence that the test was able to distinguish low-ability students from high-ability students.

Table 2
Person summary statistics

		Count	Total Score	Logit Measure	*SE*	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd
M	All	99.0	72.9	1.44	0.30	0.99	-0.1	1.08	0.0
	PI	36.7	16.4	-0.25	0.41	0.99	0.0	1.00	0.1
SD	All	0.7	10.8	0.97	0.05	0.24	1.5	0.67	1.4
	PI	1.6	9.0	1.29	0.08	0.11	0.7	0.22	0.8
Max.	All	100.0	95.0	3.92	0.53	1.72	4.2	4.39	3.4
	PI	39.0	37.0	3.29	0.75	1.34	2.5	1.77	2.4
Min.	All	98.0	54.0	-0.16	0.26	0.60	-3.3	0.42	-2.9
	PI	34.0	3.0	-2.51	0.34	0.71	-1.8	0.21	-1.9
Separation	All	2.85							
	PI	2.85							
Reliability	All	.89							
	PI	.89							

Note: *n* = 108, All = All items (*k* = 267), PI = Planned items (*k* = 110)

Table 3
Item summary statistics

		Count	Total Score	Logit Measure	*SE*	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd
M	All	36.0	25.4	0.00	0.54	0.99	0.0	1.08	0.1
	PI	36.0	16.1	0.00	0.40	1.00	0.0	1.00	0.0
SD	All	1.1	8.7	1.62	0.21	0.15	0.9	0.75	1.1
	PI	1.1	4.5	0.72	0.03	0.23	1.3	0.38	1.2
Max.	All	37.0	36.0	3.64	1.02	1.66	4.0	6.06	4.2
	PI	37.0	30.0	2.09	0.55	1.73	4.1	2.58	3.9
Min.	All	34.0	5.0	-2.57	0.35	0.58	-3.2	0.33	-2.9
	PI	34.0	5.0	-2.15	0.37	0.54	-3.0	0.44	-2.7
Separation	All	2.57							
	PI	1.38							
Reliability	All	.87							
	PI	.66							

Note: *n* = 108, All = All items (*k* = 267), PI = Planned items (*k* = 110)

Table 4
Mean item difficulty by type

	Count	*M* (Logits)	*SEM*	*SD*	*MDN*	Separation	Reliability
All	267	0.00	0.10	1.62	-.36	2.62	.87
UI	157	-1.17	0.07	0.89	-1.21	0.83	.41
PI	110	1.68	0.06	0.66	1.68	1.40	.66

```
Logits            Persons - Items
    4                   +
                   #    |
                  ##    |
                        |  PP
                  ##    |
                   # T  |
                        |T
                        |  P
    3              #    +  PP
                 ###    |
                 ###    |
               #####    |  PPPPP
               #####    |  PPP
                 ### S  |  PPPP
                   #    |  PPP
                   #    |  PPPPPPPPP
    2              #    +  PPPPPPP
               #####    |  *PPPPPP
                 ###    |  PPPPPPPPP
               #####    |S PPPPPPPPPPPPPP
              ###### M  |  PPPPPPPP
                  ##    |  PPPPPPPP
            ########    |  PP
                  ##    |  **PPPPPPPP
    1          #####    +  *PPPPPP
               #####    |  PPPPP
               #####    |  PP
     ###############    |  P
              ###### S  |  **PP
               #####    |  P
                ####    |  **P
                   #    |  **
    0              #    +M ******
                   #    |  ***
                        |  ****P
                        |  ********
                     T  |  **
                        |  **************
                        |  ****
                        |  *************
   -1                   +  ************
                        |
                        |  *********************
                        |  *****
                        |
                        |S ************
                        |  ******
                        |  *******
   -2                   +
                        |
                        |
                        |  *****************
                        |  *******
                        |  ******
                        |
                        |
   -3                   +  ******************************
Logits            Persons - Items

Key: # = 1 person
     P = 1 planned item
     * = 1 unplanned item
```

Figure 1. Person-item map. Person ability is mapped against item difficulty on a common logit scale shown on the left. A position higher on the scale indicates a higher ability person or a more difficult item.

Also of interest in Table 2 and Table 3 are the outfit mean-square fit statistics. These have an expected mean value of 1.00, a value observed for the PI items for both persons and items. However, when the UI items were included in the analysis of all items, the mean-square outfit rose to a mean value of 1.08 for both persons and items, with standard deviations rising from 0.22 to 0.67 for persons and from 0.38 to 0.75 for items. However, the outfit mean-square value reflects misfitting outlying items, with most UI items far too easy for most students and therefore contributing little information for measurement. Although careless mistakes on the UI items added noise to the dataset, the most important indicator of measurement quality is the information weighted infit mean-square figure. The infit mean-square values reported in Table 2 and Table 3 were acceptable overall, indicating that the noise in the dataset did not substantively affect measurement.

Figure 2 further illustrates this, with the PI items shown in the top panel and the combined analysis of all items shown in the bottom panel. When analyzed alone, two PI items were of serious concern, with mean-square outfit values above Linacre's (2014) threshold of degradation of 2.0, while another eight had values above 1.5. Two misfitting items out of 110 is not cause for serious concern given that this is a fairly low-stakes classroom test and that the sample of persons is fairly small, meaning that item fit statistics are vulnerable to a small number of unexpected responses. Figure 2 also shows that many PI items were overfitting, with mean-square values far below the expected value of 1.00, indicating that some items were extremely predictable, while others were quite unpredictable. When the UI items were added to the analysis, as in the bottom panel of Figure 2, the PI items showed less misfit and overfit than when analyzed in isolation, but many UI items were highly misfitting or overfitting, with 22 having mean-square values exceeding 2.0. As the UI items were extremely easy, many of the most misfitting items had five or fewer incorrect responses, making these items vulnerable to careless answers by only one or two higher ability persons. Because the average mean-square value is constrained to a value close to 1, adding many misfitting UI items to the analysis made the existing PI items appear relatively more consistent, so there are now only four PI items with mean-square values greater than 1.50 compared with 10 when the UI items were excluded. These results indicate that the UI items did not contribute to productive measurement, being too easy for the majority of students and resulting in misfitting outfit mean-square statistics.

However, not scoring the UI items would be potentially problematic in the eventuality of a student contesting grades, so the extent of the distortion of the measurement scale was investigated through anchoring of item difficulty. Persons with mean-square fit statistics exceeding 1.50 were removed from the analysis, leaving 71 of the original 108 persons. The 120 best fitting items from this analysis, with mean square values between 0.80 and 1.20 for infit and 0.70 to 1.30 for outfit, were identified and anchored, meaning that the difficulty of these items was specified at the values observed for the best fitting persons. This procedure locks the measurement scale to the values observed for well-fitting data, allowing the effect of the misfitting data on the measurement scale to be seen by analyzing the entire dataset using the anchored measurement scale. Figure 3 shows that this made effectively no difference to person measures, so subsequent analyses used unanchored values. Thus, the answer to the first research question was that the data provided sufficient psychometric performance for use as a relatively low-stakes component of a classroom test battery, with high Rasch reliability and acceptable infit mean-square statistics.

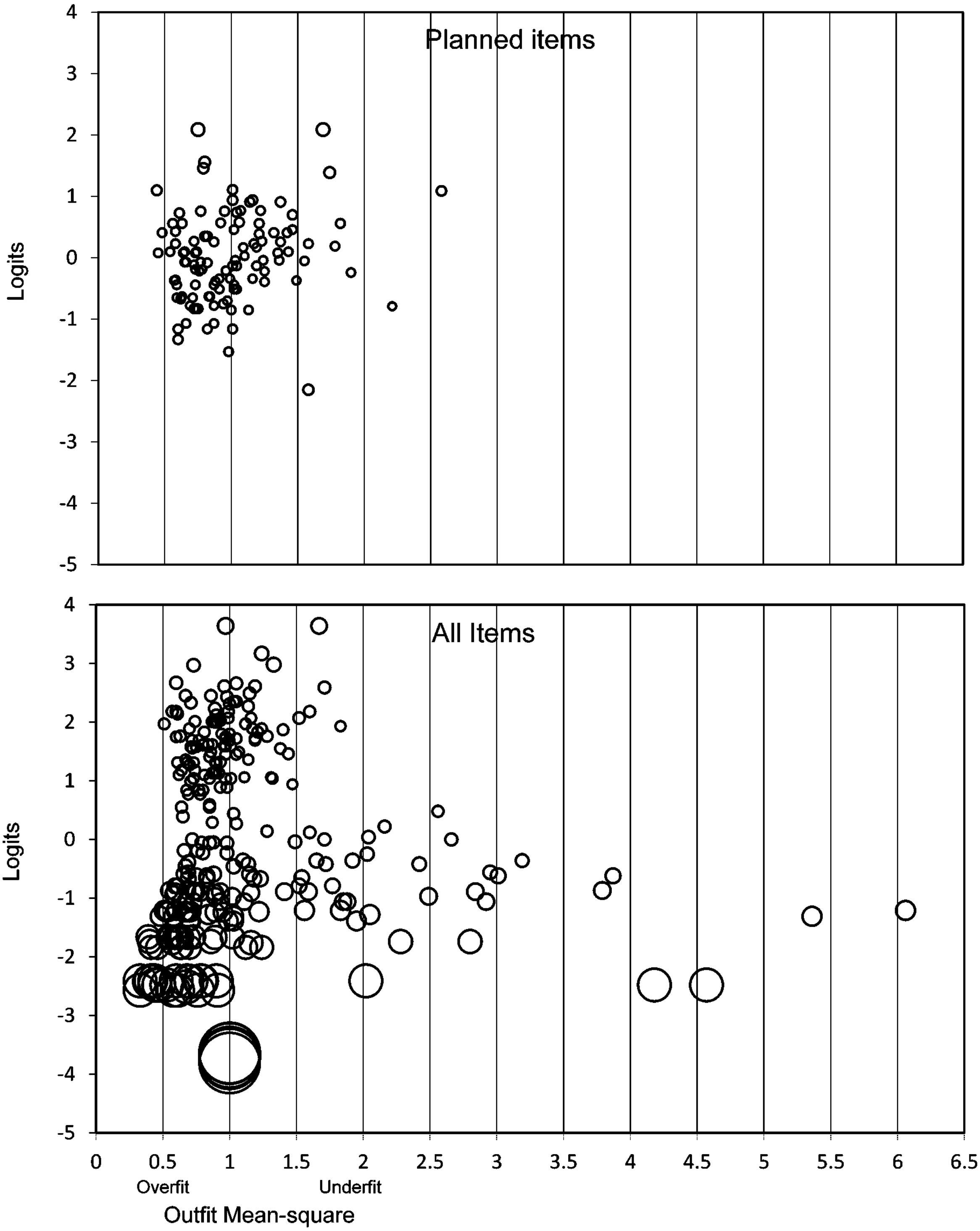

Figure 2. Item mean-square fit. The upper panel shows the planned items analyzed in isolation. The lower panel shows all items analyzed together.

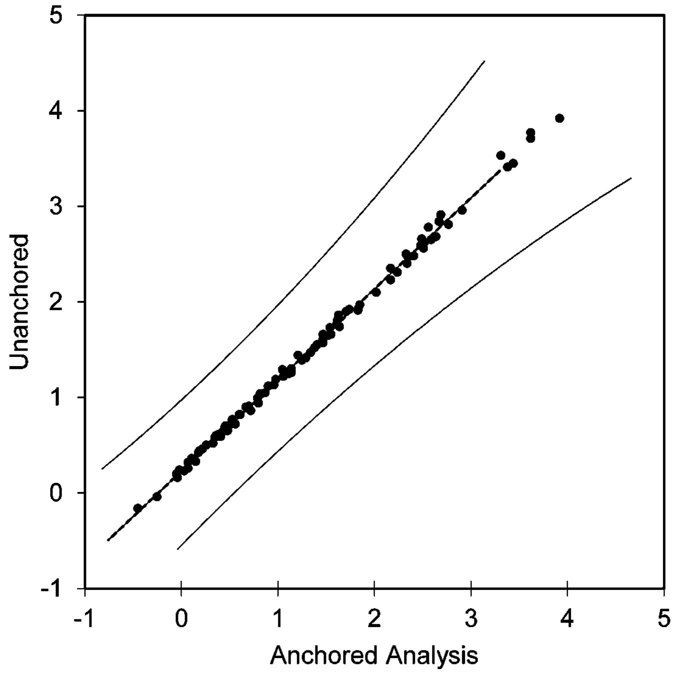

Figure 3. Person ability measures calculated using item difficulties anchored to values generated from best-fitting persons versus unanchored item difficulties.

Research Question 2: The Effect of Unplanned Items

Although the presence of the misfitting UI items did not substantively distort measurement, this still leaves the question of whether it is preferable to include them or exclude them in calculating classroom grades. As it was desirable to include all items for the purposes of assigning grades, it was necessary to check that the presence of so many misfitting items did not disadvantage students. Figure 4 compares the person ability measures from the PI items alone and all items. Five of the 108 persons fell on or slightly outside the 95% confidence margins, a level consistent with chance. Four of these persons did relatively worse when UI items were included, meaning that they are shown above the linear trendline in Figure 2, while one did relatively better, being shown below the trendline. Doing relatively worse when UI items were included indicates that a person falsely identified words as being redundant, which could arise through carelessness, random guessing, or a low-ability person confusing unfamiliar language for an error. Excluding the UI items would therefore have unfairly benefitted the students located above the trendline in Figure 4, evidence for including them.

Doing relatively better when UI items were included indicates that a person was less likely to make a mistake on a UI item than to correctly identify PI items, consistent with excessive caution or very low proficiency. Figure 4 confirms this, with the three persons most advantaged by including the UI items being near the bottom of the proficiency scale. Thus, although a few students did benefit from including all items, the inclusion of all items would not substantively change their course grade. In response to Research Question 2, it was therefore judged to be preferable to include all items for the assignment of class grades as a check against random guessing.

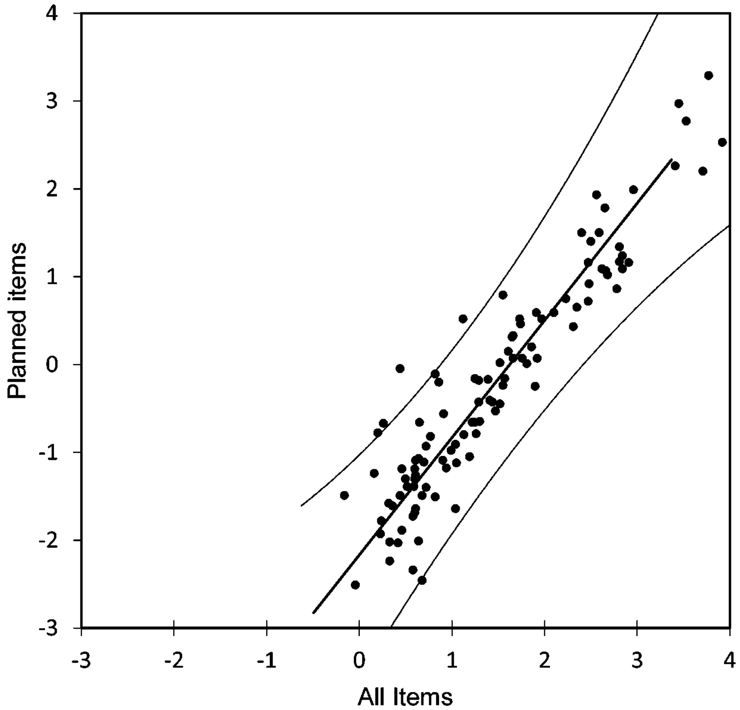

Figure 4. Person ability measures calculated from planned items versus all items. The upper and lower solid lines show the 95% confidence intervals.

Research Question 3: Formula Scores

This left two problems, however. First, UI items were very easy and comprised 187, or 63%, of the of the 297 scored items, so using a raw percentage of correct responses would mean that a passing grade of 60% would be achievable simply through not responding to any item. This made it desirable to use formula scores to adjust for different types of guessing strategies. Second, nine different test forms were used, making it probable that some test forms were more difficult than others. Although Rasch logit measures adjust for different test forms, formula scores do not, so it was necessary to investigate the magnitude of the effect of different test forms. Figure 5 compares the formula scores to raw scores from all items and the PI items alone. Unsurprisingly, the score adjustment made a dramatic difference to low proficiency students when all items were included, as shown in the left panel of Figure 5, where even the lowest students scored about 60% before adjustment, but this dropped to around 0% after adjustment. It is also apparent that the scatterplot more closely followed the linear trendline when all items were included than when only PI items were analyzed, with respectively 98% and 95% of variance explained.

Variation in the difficulty of different test forms is confirmed by Figure 6, showing formula scores versus logit measures, with the same logit measure mapping to formula scores that differ by up 20%. However, the mean standard error of the logit scores shown in Table 2 was 0.30 logits, meaning that the 95% confidence interval of the logit scores covered a range of about 1.20 logits on average, with a larger error for students near the extremes of the score range. Thus, the variation in the difficulty of test forms was generally within the range of measurement error. The CE section of the final test only contributed 10% to overall grades, so the difference between test forms was unlikely to have affected any student by more than 1% to 2%, too small a difference that an unambiguously deserving student would fail. Therefore, formula scores were used for the assignment of grades and students whose final grade fell within two percentage points below a grade boundary were investigated on an ad hoc basis and moved up a grade level if their performance on other components of the overall assessment justified it.

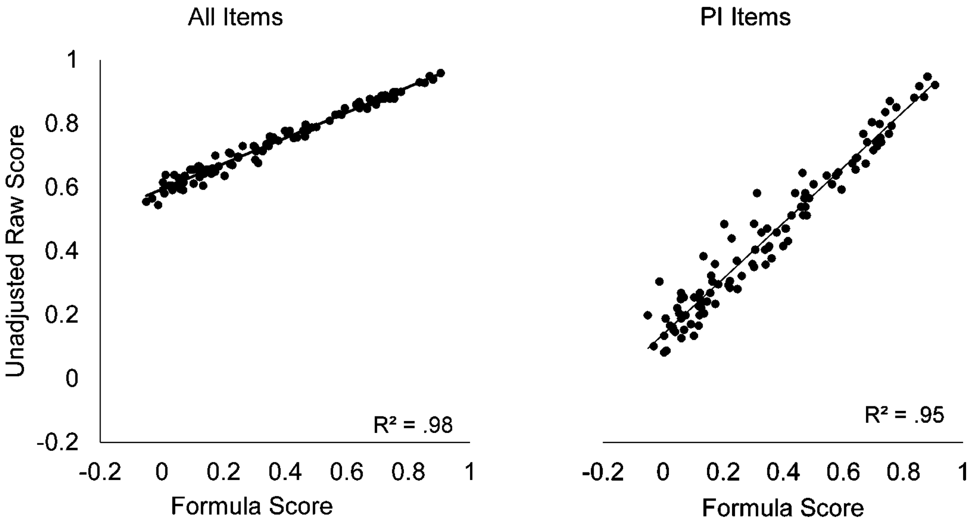

Figure 5. Raw scores versus formula scores. The left-hand panel includes all items. The right-hand panel includes only planned items. Formula scores were calculated by subtracting the proportion of incorrectly answered UI items from the proportion of correctly answered PI items.

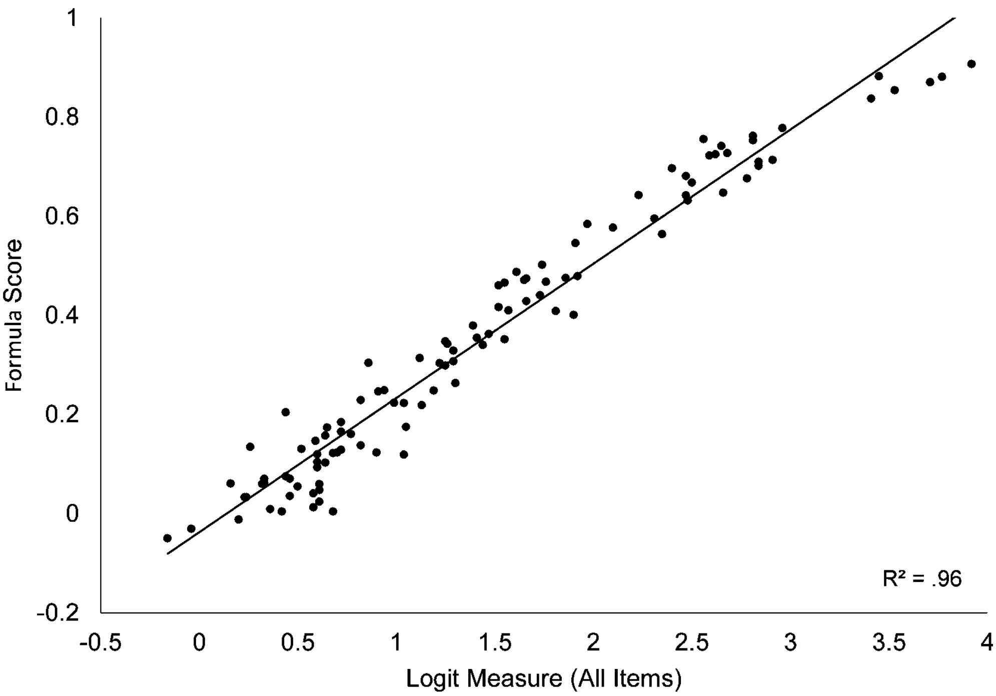

Figure 6. Formula scores versus logit measures.

Discussion and Future Directions

The CE format was found to be an extremely quick and easy format to produce, administer, and process. In this instance, it provided suitable psychometric performance for use in a test battery for assigning classroom grades, with acceptable data-model fit and Rasch reliability of .89 indicating good performance for a classroom assessment. Although the UI items contributed little information to measurement and resulted in misfitting outlying responses, they did not cause substantive harm to measurement. It was necessary to retain the UI items, however, to avoid rewarding very aggressive guessing. Retaining the UI items raised the further difficulty of rewarding excessively cautious students, so course grades had to be either be based on rescaled logit scores or formula scores. The simple formula presented in Equation 2 was found to be effective for this purpose.

The scoring of this format of CE was found to suffer from problems noted by earlier researchers. This test was intended to reward fluent grammatical parsing of simple reading texts, with the identification of randomly inserted words considered a manifestation of this ability. The original intention was that identifying UI items would not be counted towards grades because a passing grade could be achieved by not responding to any items. Rasch analysis confirmed that the UI items were extremely easy to identify, so they contributed little to measurement but resulted in a noisy dataset with a high level of outlying misfit. Although improved psychometric performance was observed when the UI items were removed, the misfit they caused represented only outlying noise rather than a substantive threat to measurement, reflected in acceptable infit statistics. As this study focused on the use of CE for low-stakes classroom use as a component of a varied test battery, the outlying misfit was not considered to be a threat to validity.

A major weakness of this study was that CE was not used as a classroom learning task, but only as a final test, precluding any possibility of a formative benefit. This occurred because this was the first time I had taught this course, so the course objectives and students' motivational orientations were not clearly understood in advance, resulting in production of classroom tasks and assessments on an ad hoc basis. Rather than basing CE on existing reading texts, such as provided by Quinn et al. (2007), it would be preferable to develop texts specifically for use as integrated classroom learning/assessment tasks. This is because the use of CE in this instance was intended as a speeded reading task, so the semantic and syntactic complexity of the texts needed to be constrained to a level that even very low-level students could process without requiring a dictionary or explicit explanation. The presence of idiosyncratic vocabulary or complex grammatical features is likely to confuse students and lead to misidentification of correct language (i.e. UI items) as incorrect. This is potentially a major contributor to the observed pattern of many misfitting UI items. Although these did not substantively affect measurement in this instance, this may indicate that the texts contained features unsuitable for students of this level of proficiency. Thus, future investigations should focus on developing texts specifically intended for use as CE tasks, both for formative classroom practice and for summative course grades. This will require investigation of language features associated with confusing UI items, development of revised specifications for texts and for CE items, and then further rounds of piloting and refinement of item specifications.

References

Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning, 62*(3), 665-703. doi: 10.1111/j.1467-9922.2011.00651.x

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Bowen, J. D. (1978). The identification of irrelevant lexical distraction: An editing task. *TESL Reporter, 12*(1), 1-3, 14-16.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. White Plains: Longman.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill College Press.

Brown, J. D. (2016). *Statistics corner*. Tokyo: JALT Testing and Evaluation Special Interest Group.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84-95.

Carver, R. P. (1993). Merging the simple view of reading with rauding theory. *Journal of Literacy Research, 25*, 439-455. doi: 10.1080/10862969309547829

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, UK: Cambridge University Press.

Davies, A. (1967). The English proficiency of overseas students. *British Journal of Educational Psychology, 37*(2), 165-174. doi: 10.1111/j.2044-8279.1967.tb01925.x

Davies, A. (1975). Two tests of speeded reading. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency*. Arlington, VA: Center for Applied Linguistics.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English*. New York: Routledge.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly, 5*(3), 173-194.

ETS. (2008). The TOEFL® Test - Test of English as a Foreign Language™. Retrieved 28 March, 2008, from http://tinyurl.com/zocgc

ETS. (2013). TOEIC user guide: Listening & reading. Retrieved from http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

Gallagher, K. (2009). *Readicide: how schools are killing reading and what you can do about it*. Portland, ME: Stenhouse.

Gravic. (2012). Remark Office OMR (Version 8.4).

Holster, T. A., & Lake, J. (2016). Guessing and the Rasch model. *Language Assessment Quarterly, 13*(2), 124-141. doi: 10.1080/15434303.2016.1160096

Hoover, W., & Gough, P. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. doi: 10.1007/bf00401799

Hughes, A. (2003). *Testing for language teachers* (2 ed.). Cambridge: Cambridge University Press.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing, 14*(1), 47-84. doi: 10.1177/026553229701400104

Koda, K. (2004). *Insights into second language reading: A cross-linguistic approach*. Cambridge, UK: Cambridge University Press.

Linacre, J. M. (2014). Misfit diagnosis: infit outfit mean-square standardized. Retrieved 22 August, 2014, from http://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2017). Winsteps (Version 3.91.2). Retrieved from http://www.winsteps.com

Manning, W. H. (1987). Development of cloze-elide tests of English as a second language *ETS Research Report Series*. Princeton, NJ: Educational Testing Service.

Maruo, K. (2012). The grammar translation method: The reason for the heavy reliance on it in junior and senior high schools and the implication of this in university English education. *Fukuoka Universtiy Language Education and Research Center: Annual review of language learning and teaching, 10*, 13-18.

Palmer, H. E. (1917). *The scientific study and teaching of languages*. Edinburgh, UK: The Riverside Press.

Quinn, E., Nation, I. S. P., & Millett, S. (2007). *Asian and Pacific speed readings for ESL learners: Twenty passages written at the one thousand word level*. Wellington: English Language Institute at Victoria University of Wellington.

Sakurai, N. (2015). The influence of translation on reading amount, proficiency, and speed in extensive reading. *Reading in a Foreign Language, 27*(1), 96-112.

Sick, J. (2008). Rasch measurement in language education: Part 1. *Shiken: JALT Testing and Evaluation SIG Newletter*, *12*(1), 1-6.

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing, 34*(2), 151-174. doi: doi:10.1177/0265532215623581

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yamashita, J. (2008). Extensive reading and development of different aspects of L2 proficiency. *System, 36*(4), 661-672.

Yamashita, J., & Shiotsu, T. (2017). Comprehension and knowledge components that predict L2 reading: A latent-trait approach. *Applied Linguistics*, *38*, 43-67. doi: 10.1093/applin/amu079

Appendix

Sample Test Paper

11. Fa Hien

Fa Hien was a Buddhist who lived in China more than 1,500 years ago. At that time, many people in China were Buddhists but they did not know much about Buddhism because there were no good books about it in China.

When Fa Hien became a Buddhist, he trim thought a lot about his people. He was unhappy because there were no Buddhist books. So in the year A.D. 399 he went to India to find some books for his people. Two of his friends went with him and they travelled by land.

After travelling for six years electronics through the west part of China, Fa Hien arrived in India. He stayed in India for ten years and lived with Indian Buddhists who taught him about Buddhism. While he was in India, he learned the Sanskrit language and objective he copied several Buddhist books. When he finished copying the books, he left merger India and went to Ceylon by boat.

Fa Hien studied Buddhism in Ceylon too. Buddhism in Ceylon was a little different from Buddhism in India or China. He also bought more books about Buddhism in Ceylon. After staying there for about two years, he decided to go back to bulb China, so he put all his books on a ship and left Ceylon. It was a big ship and there were a lot of people on it.

Everything went well during the first few days of the voyage, but suddenly there was a strong wind and the sky became very dark. Water began to come into the ship through a big hole in the bottom. In order to save the ship everybody had to throw their preach things into the sea. Fa Hien threw all his things away except his books. The weather continued to be bad for several days but one night the ship was carried to an island where the people got off the ship and mended it. They waited on the island until the sea was smooth again. When the wind stopped blowing, they continued their voyage and at last they grandchild arrived at the island of Java in Indonesia.

The ship did not go on to China, so Fa Hien waited in Java for several months before he got on another ship. At that time, it usually took campaign 50 days to sail from Java to China, but after sailing for a month there was another big storm. The people were afraid but Fa Hien did not do anything. He only prayed. People became angry with him and when they passed a small island, they wanted to leave him on the island. When Fa Hien's friends heard about this, they were angry with the people. They said they would tell the king when they arrived in China, if Fa Hien developed was left on the island. Then the other people changed their minds and they did not leave Fa Hien on the island. The storm continued for many weeks but one day they suddenly saw land. They were happy because it was China.

Fa Hien nominate was happy, too, because he still had all his books with him. He went back to his home town and spent the rest of his life writing the books in Chinese and teaching his people about Buddhism. His books are very important in the warehouse growth of Chinese Buddhism.

From Quinn, et al. (2007), reprinted with permission.