［論文］

# Creation and Validation of a Reading Comprehension Test

Robert Rowland

## 解読テストの作成と妥当性の検証

**要旨**

本論文は英語による読解力を図るテストの作成と妥当性の検証について論じる。まず、高校一年生のための解読テストの作成にあたって研究結果に基づいたデザインについて説明する。そしてテストの実行から集めたデータをラッシュモデルで分析し、妥当性の検証結果と次の実行にあったての改善すべきポイントを話す。研究の結果は、次のとおり。１）ラッシュモデル分析は作成されたテストの高い妥当性を示す２）テスト実行以前の異なった学習環境や試験者の性別はテスト結果に著しい影響をもたらさなかった３）単純な訂正でテストを改善することができる 。

Academic achievement is often measured by tests. Large–scale assessments determine the futures of millions of young people every year. Validity theory, or the measure of how well these tests measure what they are supposed to measure, has become essential to ensuring test–takers and the people who infer meaning from their scores can trust these large–scale instruments (Fulcher & Davidson, 2007). The majority of academic assessments every year, however, are small–scale classroom measurements performed on a daily basis. Validity of these instruments, though not nearly as life changing, is equally important in determining the value of education. Fortunately, the same conceptual models and statistical tools used to measure validity of larger tests can also be used for classroom measurement. This study examines the construction and validity of a high school reading comprehension test in terms of Messick's (1989) concept of construct validity, measured by the Rasch model.

## Literature Review

### *The mechanism of reading*

Research into how the skill of reading is learned over the last 40 years has revealed that it is a complex and highly nuanced process. Goodman (1970) revolutionized the way reading pedagogy was understood by proposing the idea that the action of reading stems from two different types of linguistic processing. Bottom–up processing is the understanding of the physical form of language on the page. This includes not only letters and punctuation, but also a wide variety of other cues that readers must recognize in order to process the fundamental elements of meaning, such as morphemes, words, phrases, grammatical patterns, and discourse patterns (Grabe, 2009). Readers also engage in concept driven processing of a text, which is referred to as top–down processing. Top–down processing is when we draw on our own experience and ability to infer meaning to decipher the meaning of a text. Nuttall (1996) wrote that teaching reading with a balanced focus on bottom–up and top–down processing enables readers to effectively dissect a text by alternately guessing meaning (top–down) and then checking linguistic detail to confirm or revise guesses (bottom–up).

To guide students through the complex activity of reading, researchers have recommended that students be encouraged to use strategies for reading comprehension. Eskey (2005) suggested that strategies for reading can be divided into several categories: pre-reading, while reading and post-reading. Teaching strategies for each stage of the reading process scaffolds reading and gives learners a concrete set of tasks to complete while reading, the completion of which may give them a motivation boost. Grabe (2004) said that teaching learners to use a combination of different strategies simultaneously while reading will boost their overall reading comprehension. Brown (2007) provided an extensive list of microskills (bottom-up) and macroskills (top-down) and suggests that these skills be taught iteratively and in balance to ensure that students become well-rounded strategic readers.

### *Assessing reading skills*

Grabe (2000) wrote that although our understanding of the processes of reading has advanced dramatically in the past few decades, our approach to the assessment of reading has not. He argued that our concern with the validity and reliability of reading assessment has outweighed the desire to discover and test new methods of testing reading. Alderson (2000) countered this argument by asserting that in assessment, there are few if any alternatives to a high level of concern with validity and reliability. Thus, the traditional, straightforward measures of reading comprehension, questions about main ideas and details in passages, have been and will continue to be acceptable methods of reading assessment.

Assessment of reading skills, on other hand, is an area on which there is little agreement in the literature. In fact, there is no small amount of controversy about how many different reading skills exist, if indeed they do exist. Rost (1993) said that there may only be one general reading skill which he called "general reading comprehension." Carroll (1993) argued the existence of 4 discernable reading skills: general reading comprehension, special reading comprehension, reading decoding and reading speed. Other studies have found numbers of skills ranging from 10 (Drum et al., 1981) to as many as 22 (Pollitt et al., 1985). While there is little agreement about the divisibility of reading skills, the fact that some process happens in the brain of a reader that leads to comprehension of a text seems to be commonly accepted. After extensively reviewing the literature on reading skills, Weir & Porter (1996) suggested that, judging from the major trends in research, there are only 3 distinct operations in reading:

1. Skimming: going through a text quickly
2. Reading carefully to understand main ideas and important details
3. Using knowledge of more specifically linguistic contributory skills, syntactic structure, discourse markers, lexical and or grammatical cohesion, and lexis.

### *Designing reading tests*

When designing assessment instruments for reading, Alderson (2000) suggested that test design decisions affect the difficulty of a test. Among these decisions are whether or not the learner can refer to a text when answering questions; whether one or many texts are used; what types of questions are used and the format of answers.

The effects of the ability to refer to texts when answering questions has been discussed at length in the literature. Davey and Lasasso (1984) found that students who could refer back to a text while testing received higher scores than those that did not. Studies by Alderson and Urquhart (1985) and Brown (1984)

found that students with background knowledge relevant to test texts outperformed those without background knowledge, suggesting that the inability to refer to a text while answering may present an unfair disadvantage to learners unfamiliar with themes in the text. Johnston (1984) found that the ability to refer directly to the text while answering questions forced students to rely on reading comprehension as opposed to background knowledge. He also found that taking away the text and testing for reading comprehension may be more of a test of memory than comprehension.

The number of texts and each text's length also has an effect on the difficulty of the test. Alderson and Urquhart (1985) argued that on tests that have one long text, as opposed to many shorter ones, students able to read faster will have a distinct advantage. Therefore, for tests of reading comprehension, it may be better to use what Alderson (2000) called 'Testlets.' Testlets are short texts with comprehension questions relevant only to the given text. As long as test design controls for text interdependence on the test, that is, as long as no one answer influences an answer for a different testlet, these types of items can control for the reading speed advantage. They also help control for advantage to students with relevant background knowledge as long as there are a variety of themes within the texts.

Pearson and Johnson (1978) said that there are three distinct types of questions typically used on comprehension tests. The first and easiest are textually explicit questions. In these questions, both information from the question and the answer are both found in the same place in the text, usually in the same sentence. The second type of question is textually implicit questions. Unlike textually explicit questions, the answers to this type must be combined across several sentences of the text, raising the difficulty. The most difficult type of questions are Script–based questions. These demand that readers combine both information from the text and from their own background knowledge to answer questions. Alderson (2000) questioned, however, whether this last type of question can be classified as a reading comprehension question, as it relies on information outside of the text to be answered.

Alderson (2000) pointed out that there are a wide variety of ways to assess reading comprehension which can be divided into two categories based on response type. Open response pattern questions require the test taker to write out an answer. Often, there is more than one possible response to these questions. Cloze questions, short response and summary writing questions are examples of open response questions. The advantage of this question type is that it allows for partial credit when a student shows sufficient understanding of a text and question, but does not provide a fully correct response. The disadvantage with these questions is the lack of predictability of answers and consequent time demand during marking. The second category of questions is closed response type questions. These questions require students to make a choice from a series of choice options. Multiple choice, true–false and ordering questions are all examples of closed response types. The advantage of these question types is that they are dichotomous, that is, they are either right or wrong with no partial acceptability. This means that they are very easy to mark. The disadvantage of this question type is that, especially in the case of multiple choice and ordering questions, the multitude of answer options may confuse students. Also, it is nearly impossible to determine student answer motivation.

Regardless of methodology of test design, the assessment of the quality of a test is largely determined by a multitude of factors unique to a given testing situation. The examination of how different elements of a test instrument and testing situation affect the outcomes and decisions made about the outcomes is a heavily researched topic known as test validity. Though a well–established and deeply discussed topic in the literature, there are different interpretations of validity.

### *Validity*

Validity is widely considered to be the most important element of any assessment. It has been defined and redefined many times. Cronbach and Meehl (1955) proposed that validity be defined by the evidence collected in demonstration of validity. They described criterion–oriented validity, content validity and construct validity as three distinct aspects on which one must gather information to argue the validity of a test. Then, Messick (1989) took Cronbach and Meehl's idea of construct validity placed it at the center of a framework of validity that is still the most popular to this day (Fulcher & Davidson, 2007).

Messick (1989) proposed the idea of a construct–centered approach to validity. This approach is contrasted with a task centered approach, in that it puts the constructs to be measured, rather than the tasks that measure them, at the center of the development of methods of measurement and scoring. In Messick's construct validity there are 6 facets: content, substantive, structural, generalizability, external, and consequential.

The content facet contains content relevance, representativeness and technical quality. Content relevance and representativeness dictate that the items on the test and the cognitive processes that they require should be directly related to the construct being assessed. Technical quality refers to the aspects of the test construction, such as the reading level of texts and items. The substantive facet refers to how evidence that examinees are actually engaged in the cognitive processes the items and tasks were designed to induce is found. Contained in the structural facet are the aspects of the test involved in implementation and scoring, such as whether or not the time constraints are appropriate for the test. Generalization refers to whether or not test interpretations and measurements are relevant to other tasks that aren't included in the test but are considered to measure the same construct. External validity looks at how the measurements are affected by other behaviors that are part of the test and otherwise. A score should only be affected by expected interactions with the construct. Consequential validity looks at what test scores and their interpretations mean for actions beyond the test. If an examination of a test fulfills all of these facets across all test takers, we can assume that it has construct validity.

### *Measuring construct validity with the Rasch Model*

The Rasch Model is a commonly used tool to determine the validity of an instrument. Rasch analysis uses a unit of measurement called logits to measure ability and item difficulty independently of test items and test takers (Rasch, 1966). The data from these analyses can be used to determine the conditions and specifications that a successful measurement tool should have (Runnels, 2012). Rasch analysis has been widely used to validate measurement instruments for tests of English in Japan (Beglar, 2010; Beglar and Hunt, 1999; Runnels, 2012). Researchers have been able to show a clear connection between the results of Rasch analysis and facets of Messick's (1989) construct validity. To confirm content relevance, the fit statistics can be examined to determine 1) if the test is targeting the intended construct and 2) if there are any misfitting items (Runnels, 2012). Content representativeness can be determined by examining person–item maps. If items and people are regularly distributed, we can assume that 1) the construct has been thoroughly assessed and 2) an appropriate range of difficulties for the population were represented (Baghaei, 2008). To examine content technical quality, the fit statistics can be examined to see if there are misfitting items. For example, if test takers identified by the model as low ability are answering high difficulty questions correctly while high ability students are not, there may be an issue with the technical quality of the test (Runnels, 2012).

The substantive facet of construct validity can also be confirmed by examining the fit statistics. If test takers are performing as predicted by the model, one could argue that they are engaged in the mental processes intended to be assessed by the items (Runnels, 2012). The structural facet can be determined by examining the multiple choice question distractor analysis and the conditions of the test implementation. If distractors are distracting test takers meaningfully and consistently and the amount of time allotted to finish the test is appropriate, one could make an argument that the test is structurally valid (Wolfe & Smith, 2007).

To examine the generalization facet, one would need to take one of two approaches. One approach would be to administer the test multiple times with different populations. If variance is within an acceptable range across different populations, scores may be generalizable (Linacre, 2007). By correlating the results of two different tests, one could also examine the external facet (Runnels, 2012). The other approach would be to divide the test items into two groups and compare the fit statistics of each group. If the two groups of items are built from the same specifications and yield acceptable fit data, one could make an argument that test scores might be generalizable (Linacre, 2007).

The Rasch model does not provide data to confirm the consequential aspect of construct validity. However, if the threshold for fit statistics is appropriate for the assessment context, and the stakeholders are satisfied with the interpretations of the scores of the exams, actions taken as the result of scores could be considered well justified.

### *Research questions*

In light of the implications of the reviewed literature, the following research questions were set for this study:

1) Does the developed instrument have an acceptable level of construct validity?
2) Are there any significant differences across groups of students?
3) How can this data be used to improve the instrument for future measurements?

## Methods

### *Participants*

The participants in this study were 75 10th–grade students at a private high school in Tokyo, divided into two classes. There were a total of 54 girls and 21 boys evenly divided across both classes. We will call them class A and class H. Each of these classes is divided into two sections. The researcher and another teacher were responsible for 1 section of each class. We will call these teachers R and S. All of these students were English majors and had been selected from a group of 350 students as high academic achievers based on their academic record from junior high school and results from the school's entrance examination. They had received 6 hours of instruction in English per week in developing reading, writing and speaking skills from a native English teacher for a year. The reading curriculum had 3 different streams. The first stream was a general reading comprehension curriculum that focused on reading skill building and focused on the main skills outlined by Weir & Porter (1996). The second stream was a timed–reading curriculum for building reading fluency. The final stream was an extensive reading curriculum. The students had received an additional 4 hours of instruction a week from Japanese English teachers in grammar, vocabulary and intensive reading skills, for a total of 10 hours of English lessons a week for 1 academic year. The program in which they studied was highly competitive and the lowest achievers were expected to be removed from the

program at the end of the year. As a result, students were highly motivated to perform well in their classes from the beginning of the year. Surveys given by the school to determine attitudes towards English language education indicated that these students were highly interested in achieving mastery of the language. All other students from the same academic year outside of this program surveyed reported that developing communication skills was of high interest and mastery to be out of reach.

### *Instrument*

The instrument used in this study was a reading comprehension test. It was composed of two unseen reading texts and 17 multiple–choice comprehension questions. The texts had 11 and 6 questions respectively. The students had access to the texts while answering questions. There were two types of questions on the test: textually explicit and textually implicit. 7 questions were textually explicit and 10 questions were textually implicit. All questions were written to minimize the effect prior knowledge on test performance. The topics for the texts were chosen to maximize learner familiarity with the content. The topic of the first text was a 5–paragraph essay on the causes and effects of child labor. The students had recently completed a 3–week unit exploring the child labor situation in a specific third–world country, so it was assumed that test takers would all have adequate background knowledge of the topic. The topic of the second text was a 5–paragraph essay about the appeal of Hawaii from a Japanese perspective. Again, based on discussions throughout the year with students about vacations and the high–level of general awareness of Hawaii within the Japanese population, it was assumed that students would have adequate background knowledge to comprehend the text.

The vocabulary of each text was profiled and modified so that approximately 90% percent of vocabulary was within the GSL 2000 level. Vocabulary from AWL and off–list vocabulary was examined to gauge learner familiarity. These words were then kept or changed to arrive at approximately 95% expected known–vocabulary on the test. Readability in terms of Flesch–Kincade reading ease scores and Flesch–Kincade grade levels were also measured and found to be within the range of other texts used in reading comprehension exercises in lessons throughout the semester. Vocabulary and readability information of both readings is summarized in Table 1 and Table .

Table 1

*Vocabulary profile and readability of each reading*

| | Reading 1 | | Reading 2 | |
|---|---|---|---|---|
| List | % | Cumulative % | % | Cumulative % |
| GSL 1000 | 88.0 | 88.0 | 75.9 | 75.9 |
| GSL 2000 | 2.1 | 90.1 | 8.4 | 84.4 |
| AWL | 6.7 | 96.9 | 2.3 | 86.7 |
| Off-list | 3.0 | 100 | 13.2 | 100 |
| Word count | 430 | | 522 | |
| Reading ease | 65.4 | | 52.8 | |
| Grade level | 7.7 | | 9.1 | |

### *Procedure*

The instrument was created by the researcher (R) and teacher S. Following completion, the test was reviewed by another teacher familiar with the program. This stage in test development is called alpha–testing and Fulcher and Davidson (2007) argue that it is a valuable practice because it allows test creators to gather information about the usefulness of test items prior to implementation. The alpha–tester suggested some structural improvements to the test design. These minor changes were made and the final draft was prepared the day before implementation. The test was given in the students' homeroom in a 50–minute test period. The test was proctored by the researcher. In addition to the reading comprehension test described in this paper, students also took a vocabulary test during this test period. The vocabulary test was designed to take around 15 minutes and was administered prior to the reading comprehension test. All students were able to finish the test within the allotted time period. The tests were then collected, graded, and scores were input into a spreadsheet for analysis.

### *Analysis*

Scores from the test were analyzed using MINISTEPS® Rasch software version 3.81.0 (Linacre, 2007), a free–trial software package of WINSTEPS®. Data from this analysis were examined in a similar fashion to the study by Runnels (2012). Item difficulty and person ability were measured in logits. Item strata were calculated to determine how many distinct difficulty levels existed in the items. Smith (2001) suggests that a minimum of 2 strata are necessary for items to be considered representative of the target domain. A person–item map was generated. The information on this map indicates the spread of item difficulty and the spread of person ability on a single scale and provide information for almost all of the facets of Messick's (1989) construct validity. Infit and outfit statistics were also calculated. Infit statistics can be examined to find response patterns that don't fit the model in terms of ability. Outfit statistics indicate inexplicable behavior such as guessing and random mistakes (Linacre, 2007). Fit statistics have two values: mean–square values (MNSQ) and z–standardized scores (ZSTD). For low–stakes multiple–choice tests, acceptable fit ranges are between .7 and 1.3 MNSQ, and –2.0 to 2.0 ZSTD (Linacre, 2007).

## Results and discussion

The research questions for this study were as follows:

1) Does the developed instrument have an acceptable level of construct validity?
2) Are there any significant differences across groups of students?
3) How can this data be used to improve the instrument for future measurements?

To answer the first research question, the data will be examined with relation to each facet of Messick's (1989) construct validity. Table 2 shows a summary of output for all items. These data indicate the degree to which the test has content relevance, representativeness, technical quality, and fulfills the substantive facet. The overall test average was 79.8% and Rasch reliability rating for items was .89. Program expectations dictated that an average of 80% across all 4 would indicate satisfactory learning throughout the course, so this test also has strong consequential validity, as results are clearly indicative of learning resulting from the reading curriculum. There were 2.78 strata of item separation, so although the test items did not cover a broad measure of difficulty, there was sufficient separation to infer that there was adequate representativeness of the main construct (Smith, 2001). Also, the MNSQ and ZSTD values were all within

acceptable ranges. In light of this data, and the fact that all test takers were able to finish the test within the prescribed constraints, this instrument had good content relevance, had no glaring technical problems, and fulfilled the substantive and structural facets of Messick's (1989) construct validity. To examine representativeness, we must look at the person–item map. At the same time, we can address the answer to research question number two, mainly whether or not there were performance differences across different cross–sections of test–takers.

Table 2

*MINISTEPS® Summary of output for all test items*

| | | | | | Infit | | Outfit | |
|---|---|---|---|---|---|---|---|---|
| | Total score | Count | Measure | Model error | MNSQ | ZSTD | MNSQ | ZSTD |
| *Mean* | 59.4 | 75 | -2 | 0.5 | 1.00 | 0.2 | 0.85 | 0 |
| *SD* | 6.9 | 0 | 2.0 | 0.4 | 0.07 | 0.4 | 0.28 | 0.5 |

Separation = 2.78; Item reliability = .89

Figure 1 shows the person–item map. Items can be seen on the right of the map and increase in difficulty from bottom to top. Persons [Rasch term] can be seen on the left of the map and increase in ability from bottom to top.

Questions are coded first with the reading in which they appear (text 1 = R1, text 2 = R2) followed by the number in the order of questions for each text. Test–takers are coded as their class (either A or H), followed by their teacher (either R or S) and their gender (either [B]oy or [G]irl). One item (R1Q5) falls well above the majority of test–takers and the rest fall either at or below the average ability level. 3 items fall well below all test–takers (R1Q3, R1Q4 and R1Q2). All but two students are above 0 logits. This fact, combined with the fact that the majority of items fall below the ability level of the majority of students indicates that the items on this test were easy for the majority of students. Furthermore, the test items demonstrate poor representativeness of varying levels of difficulty. However, this test was a summative assessment of skills that all students were expected to have mastered by the time of assessment. Therefore, the poor representativeness of item difficulty could be an indication of a high level of ability that is not represented by a person–item comparison alone.

There is no significant difference between the abilities of students of different classes, teachers or gender. This was a relief to the teachers because class H has considerably fewer contact hours in the months leading up to this test. In addition, from observations in daily lessons, both teachers feared that the boys would under perform compared to the girls.

These data, however, indicate that this was not the case, as there is a similar spread of abilities across both gender populations. It seems that the answer to the second research question is that the majority of items on the test accurately represented the material taught by each teacher and learned by each class.

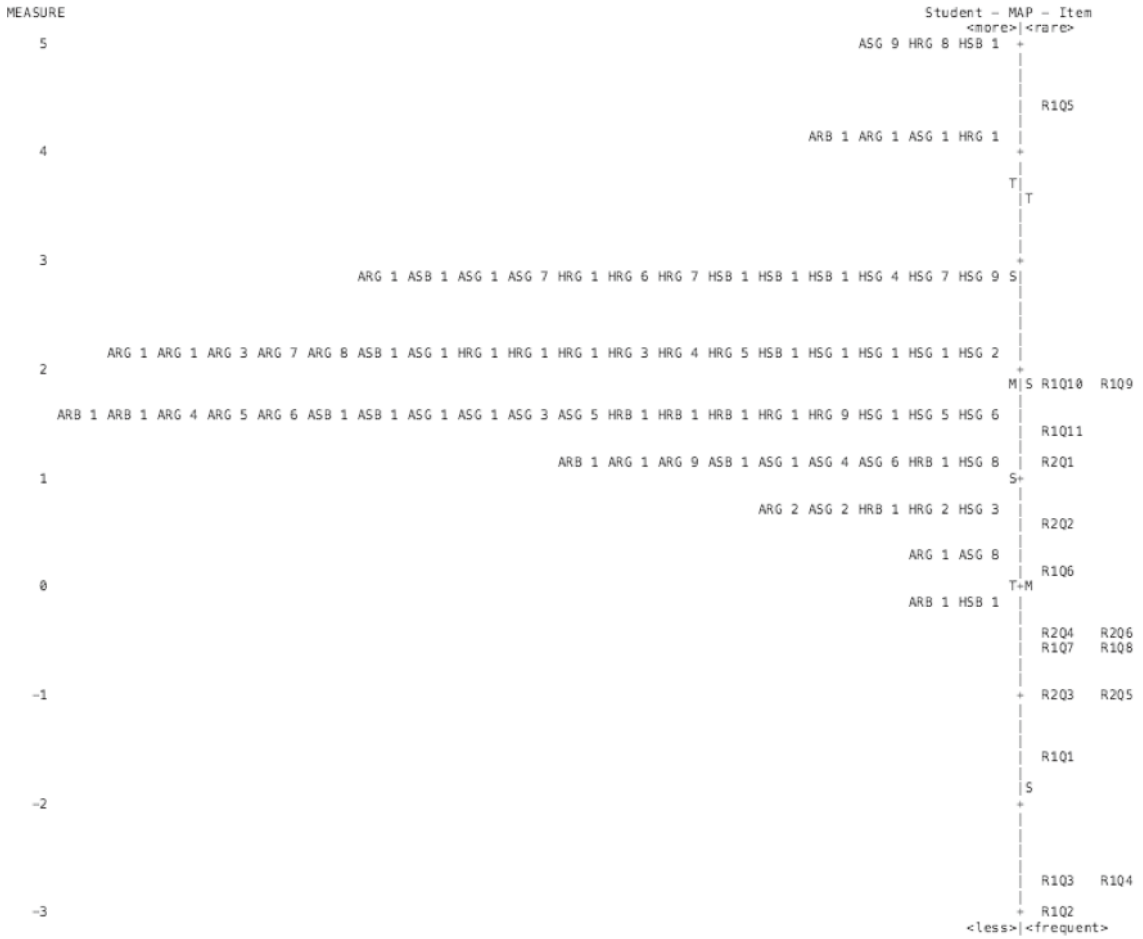


Figure 1

*Rasch person – item map for all items and measures*

By examining how item difficulty and person ability interact, we can also address research question number three by determining how the test could be improved for future use. From these data it is evident that one item (R1Q5) was much more difficult than the others. Only 10 test–takers out of 75 answered this item correctly. The infit statistics for this item, however, are within the acceptable range. In fact, all infit statistics are within the acceptable range in both MNSQ and ZSTD. Several items, however, are outside of the acceptable range in outfit. 1 item was just above acceptable MNSQ range (R2Q2). 2 items were just below acceptable MNSQ range (R2Q6 and R2Q4). Two more items were well below the acceptable MNSQ range (R1Q3 and R1Q4). As this test was a low–stakes classroom test, these items do not adversely affect the consequential validity. Furthermore, the two items that fell well below the fit range were also very low in difficulty. It was the hope of the designers that most items would be easy for students, so these items are acceptable. Also, none of the point measure correlations were negative, which indicates that all of the items are measuring the construct similarly and were construct relevant (Runnels, 2012). Still, to understand why they might not have fit the model, a distractor analysis was run on all 6 items.

Table 3 shows the item fit statistics for all test items. The left–most column shows the items. The next two columns show the total number of correct answers (Total Score) and the total number of times answered

(Total Count) per item. The next column shows the difficulty of each item (Measure). All items are in descending order of difficulty. The columns that follow the measure are the standard error for difficulty measure, MNSQ and ZSTD for infit and outfit, the point measure correlation and the observed and expected correlation for all items.

Table 3

*Item statistics for all items in order of Rasch measure*

| | | | | | INFIT | | OUTFIT | | PT-MEA. | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Total Score | Total Count | Measure | Model S.E. | MNSQ | ZSTD | MNSQ | ZSTD | Corr | Exp | Exact obs% | Match exp% |
| R1Q5 | 10 | 75 | 4.5 | .4 | .95 | -.1 | .79 | -.3 | e .55 | .51 | 90.3 | 90.2 |
| R1Q10 | 39 | 75 | 1.9 | .3 | 1.00 | .0 | 1.03 | .3 | G .41 | .42 | 70.8 | 66.2 |
| R1Q9 | 40 | 75 | 1.8 | .3 | 1.07 | .8 | 1.05 | .5 | D .37 | .41 | 58.3 | 66.2 |
| R1Q11 | 45 | 75 | 1.5 | .3 | 1.12 | 1.3 | 1.09 | .7 | C .31 | .39 | 61.1 | 66.7 |
| R2Q1 | 51 | 75 | 1.1 | .3 | .94 | -.5 | .89 | -.6 | d .41 | .36 | 73.6 | 70.6 |
| R2Q2 | 58 | 75 | .5 | .3 | 1.05 | .4 | 1.31 | 1.2 | A .27 | .32 | 76.4 | 77.6 |
| R1Q6 | 62 | 75 | .1 | .3 | 1.01 | .1 | .98 | .0 | H .28 | .29 | 79.2 | 82.3 |
| R2Q4 | 67 | 75 | -.5 | .4 | .91 | -.2 | .63 | -.8 | b .33 | .24 | 88.9 | 88.9 |
| R2Q6 | 67 | 75 | -.5 | .4 | .90 | -.3 | .68 | -.6 | a .33 | .24 | 88.9 | 88.9 |
| R1Q7 | 68 | 75 | -.6 | .4 | 1.11 | .5 | 1.28 | .7 | B .13 | .22 | 90.3 | 90.3 |
| R1Q8 | 68 | 75 | -.6 | .4 | .98 | .0 | .72 | -.5 | f .27 | .22 | 90.3 | 90.3 |
| R2Q3 | 70 | 75 | -1.0 | .5 | 1.04 | .2 | .76 | -.2 | F .19 | .19 | 93.1 | 93.0 |
| R2Q5 | 70 | 75 | -1.0 | .5 | 1.00 | .1 | .84 | -.1 | h .20 | .19 | 93.1 | 93.0 |
| R1Q1 | 72 | 75 | -1.6 | .6 | 1.06 | .3 | .92 | .2 | E .11 | .15 | 95.8 | 95.8 |
| R1Q3 | 74 | 75 | -2.7 | 1.0 | .99 | .3 | .44 | -.1 | g .14 | .09 | 98.6 | 98.6 |
| R1Q4 | 74 | 75 | -2.7 | 1.0 | .92 | .2 | .21 | -.5 | c .21 | .09 | 98.6 | 98.6 |

Table 4

*Distractor Analysis of Misfit Items*

| Item | Data Code | Score value | Data count | % | Average ability | S.E. Mean | Outf MNSQ | PTMA Corr |
|---|---|---|---|---|---|---|---|---|
| R2Q2 | c | 0 | 13 | 17 | 1.34 | .20 | .9 | -.28 |
| | b | 0 | 4 | 5 | 1.86 | .76 | 3.2 | -.04 |
| | a | 1 | 58 | 77 | 2.22 | .16 | 1.0 | .27 |
| R1Q3 | b | 0 | 1 | 1 | .68 | | .4 | -.14 |
| | d | 1 | 74 | 99 | 2.07 | .14 | 1.0 | .14 |
| R1Q5 | c | 0 | 30 | 40 | 1.69 | .15 | .8 | -.25 |
| | d | 0 | 31 | 41 | 1.83 | .17 | 1.1 | -.16 |
| | b | 0 | 4 | 5 | 2.38 | .32 | 1.3 | .07 |
| | a | 1 | 10 | 13 | 3.70 | .51 | .8 | .55 |
| R1Q4 | a | 0 | 1 | 1 | -.11 | | .2 | -.21 |
| | d | 1 | 74 | 99 | 2.08 | .14 | 1.0 | 21 |
| R2Q4 | b | 0 | 1 | 1 | -.11 | | .2 | -.21 |
| | a | 0 | 5 | 7 | 1.06 | .31 | .7 | -.21 |
| | c | 0 | 2 | 3 | 1.11 | .00 | .6 | -.13 |
| | d | 1 | 67 | 89 | 2.18 | .14 | .9 | .33 |
| R2Q6 | c | 0 | 2 | 3 | -.11 | .00 | .2 | -.30 |
| | b | 0 | 5 | 7 | 1.23 | .26 | .8 | -.19 |
| | a | 0 | 1 | 1 | 1.59 | | 1.0 | -.05 |
| | d | 1 | 67 | 89 | 2.18 | .14 | .9 | .33 |

Table 4 shows the results of the distractor analysis of the 6 items that were outside of acceptable range for outfit statistics. The left–most row shows the items, followed by answer choices (answer code), indicator of the correct response (score value), the number of responses at each answer choice (data count) and the percentage of total responses each answer choice represents. The next columns show the average ability of each choice listed in logits, the standard error of measure, outfit and point measure correlation statistics. The data here show that R1Q3 and R1Q4 were easy questions that all but one student answered correctly. That student was a high ability student and this contributed to overfit. Items R2Q2, R2Q4 and R2Q6 distractors also show a high number of correct answers and it is likely that the distractors are distracting takers in a meaningful way. Only R1Q5 shows that the distractors were too distracting and possibly prevented students from answering correctly. This item should be rewritten to adjust for this phenomenon.

## Conclusion

The reading comprehension test examined in this study was determined to have an acceptable level of construct validity for its context. Furthermore, though there were smaller sections of male and female students taught by two different teachers, no significant differences in the performance of different student groups were found. With minor changes to a small number of test items, this test could easily be improved.

There were some limitations to this study. First of all, because of time constraints, differential item functioning was not examined. This could have provided information about the external validity of the instrument. Also, because there was only one population of test-takers and only one instance of the test, it is difficult to make a strong argument for the generalizability of results of this analysis.

To further develop this instrument, item R1Q5 should be rewritten. In addition, a number of more difficult questions should be added to the test to better distinguish test taker ability strata. Also, the properly functioning questions could be used to create test specifications. With test specifications, it would be possible to develop new items for this test. Data from a larger number of items would help further substantialize validity arguments. In addition, different instruments could be created from these specifications and their performance could be compared with this instrument to collect more information for validity arguments.

### References


Alderson, J.C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J.C., and and Urquhart, A.H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing 2* (2), 192-204.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction*, 22(1), 1145-1146.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27, 101-118.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, 16(2), 131-162.

Brown, H.D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. Englewood Cliffs, NJ: Prentice Hall Regents

Brown, J.D. (1984). A norm-referenced engineering reading test. In A.K. Pugh and J.M. Ulijn (eds.), *Reading for professional purposes*. London: Heinemann Educational Books.

Carroll, J.B. (1993). *Human cognitive abilities*. Cambridge: Cambridge University Press.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52, 281-302.

Davey, B., and Lasasso, C. (1984). The interaction of reader and task factors in the assessment of reading comprehension. *Experimental Education* 52, 199-206.

Drum, P.A., Calfee, R.C., and Cook, L.K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly* 16, 486-514.

Eskey, D. (2005). Reading in a second language. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. (pp. 567-579). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Goodman, K. (1970). Reading: A psycholinguistic guessing game. In H. Singer & R.B. Ruddell (Eds.), *Theoretical models and processes of reading* (pp. 497-508). Newark, DE: International Reading Association.

Grabe, W. (2000). Developments in reading research and their implications for computer-adaptive reading assessment. In M. Chalhoub-Deville (ed.), *Reading acquisition*. Hillsdale, NJ: L. Erlbaum.

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics*, 24, 44-69.