Making Better Tests with the Rasch Measurement Model

Omar Karlin, EdD
Assistant Professor, Department of Sociology
Toyo University, Tokyo, Japan

Sayaka Karlin, MSEd
Adjunct Professor, Department of Economics
Toyo Gakuen University, Tokyo, Japan

This study had two aims. The first was to explain the process of using the Rasch measurement model to validate tests in an easy-to-understand way for those unfamiliar with the Rasch measurement model. The second was to validate two final exams with several shared items. The exams were given to two groups of students with slightly differing English listening proficiency. The two exams, a low-advanced and a high-advanced exam, were given to 76 and 45 Japanese university students, respectively. Each exam had 56 questions with 26 shared questions linking the two exams. After conducting a simple Rasch analysis, it was determined that up to 33 questions needed to be modified or deleted from subsequent versions of the exam. The unexpected number of recommended modifications and deletions suggests that, even for experienced teachers, the Rasch measurement model can be of tremendous value by offering greater precision in the assessment of students, as well as greater assistance in the validation of tests.

Literature Review

"Tests do not have reliabilities and validities, only test responses do...test responses are a function not only of the items, tasks, or stimulus conditions but of the persons responding and the context of measurement" (Messick, 1989, p. 14).

Test validity can be defined as how accurately a test measures what it is supposed to measure. Is a listening test actually measuring listening ability? Is an advanced reading test actually measuring advanced reading ability? Are the questions at the appropriate difficulty level for the students? Are the questions worded clearly, or are they confusing students? Teachers need to remember Messick's quote whenever they give their students a test, as it is important to make sure that their test is measuring what it is supposed to be measuring.

One way to assess the validity of a test is to use the Rasch measurement model. While this paper will focus on how language teachers might use the Rasch measurement model, teachers of *any* subject can use the Rasch measurement model to better assess their students and/or validate their tests. The same principles of improved assessment and validation being demonstrated in this paper can be applied to any subject where testing occurs. Traditionally, language teachers have used Classical Test Theory (often referred to as CTT) when making and giving tests (Novick, 1966). With CTT, a person answers questions correctly or incorrectly and gets points for correct answers. While CTT can be easy-to-score, the imprecise nature of the assessment makes it best for low-stakes testing (Nunally, 1978). In contrast, the Rasch measurement model offers teachers several valuable benefits, most importantly, (1) a means of assessing the validity of a test's questions and (2) a more accurate assessment

of the ability of students (Andrich, 1988; Bond & Fox, 2007; Linacre, 1997; McNamara, 2011; Runnels, 2012).

Perhaps a good way to summarize the Rasch measurement model is that it is a method of analyzing response data, in which both the questions on the test (referred to as *items* in this paper) and the people taking the test (referred to as *persons* in this paper) are incorporated into a predictive mathematical model. The Rasch measurement model uses the response data from a test's questions to predict how each person *should* respond to each question. In this process, ordinal data of correct and incorrect responses are converted into interval data (examples of interval data are frequently seen in the physical sciences, such as units of distance, weight, and speed). For example, rather than answers being marked simply as correct or incorrect (ordinal data), the Rasch measurement model is able to assign a specific value to each question, so an easy question might have a difficulty measure of 0.75 logits while a difficult question might have a difficulty measure of 3.40 logits. The conversion of ordinal data into interval data is done for *both* items and persons. Items are given a difficulty measure, which is a number representing the difficulty of a question. This item difficulty can be used to assess the appropriateness of questions. Similarly, persons are given a person ability measure, which is a number representing the ability of people in the construct that is being measured (in the case of this paper, English listening ability for university students in Japan). The Rasch measurement model also produces a slew of other data which indicates how well the real responses matched the model's predicted responses, and this data can be further used to validate a test.

To illustrate the difference between CTT and the Rasch measurement model, imagine a physics test with two questions, "*What is the formula for force?*" and "*How does Einstein's theory of relatively work?*". John answers only the first question correctly and Mary answers both questions correctly. With CTT, John would get a grade of 50% and Mary a grade of 100%. Does this mean that Mary is twice as smart as John? Because John answered a basic question and Mary answered a basic and an advanced question, Mary is probably much smarter than John, but it is difficult to say that she is exactly twice as smart as John. The Rasch measurement model weighs items based on how many people answered the questions correctly, and simultaneously produces difficulty measures for items and person's ability measures for people. These difficulty and ability measures give very precise assessments of where items stand in relation to other items, and where people stand relative to other people (Sadiq, Tirmizi, & Jamil, 2015). In the previous example with John and Mary, the basic question might have a difficulty measure of -0.56 and the advanced question might have a difficult measure of 2.40, while John might have a person ability measure of -0.36 and Mary might have a person ability measure of 2.80. Based on this, the Rasch measurement model offers a much more accurate assessment of an item's real difficulty level or a person's true ability level. This difference in accuracy between CTT and the Rasch measurement model can have real-life consequences for language teachers. In a study by Weaver, Jones, and Bulach (2008), several students entering a university as freshmen were placed in different ability levels depending on whether their placement exam was scored with CTT or with Rasch measurement, illustrating how more precise assessment methods, such as the Rasch measurement model, can lead to better student placement when entering a university.

Another feature of the Rasch measurement model is that it makes it easier for teachers to improve their tests. One way it does this is by putting the difficulty level of the items and the ability level of the persons on a shared scale, so the items and persons can be easily compared, as shown in the *Wright Map* in Figure 1. The Wright Map in Figure 1 includes several x's on the left side of the vertical line which represent the people who took the test. The top x (at 2 logits) represents the person with the highest ability, and the bottom x (at -1 logits) represents the person with the lowest ability. On the right side of the vertical line, numbers from 1-56 represent the questions on the test. The highest item is 20, which was the most difficult question on the test, and the lowest items are 55 and 56, which were the two easiest questions on the test. When a person and an item are perfectly matched, such as the top x and item 36, the person has a 50% chance of answering that question correctly. For the top x, the only item that was above their ability was question 20. Being able to easily see how the people and items match can be useful if teachers want to know if their test was too easy or too difficult. If the test was too easy, the items on the right would be below the persons on the left. If the test was too difficult, the items would be above the persons. This visual inspection is one way that the external validity of a test can be confirmed (Baghaei & Amrahi, 2011).

In the case of Figure 1, items 15, 8, 17, 14, 53, 54, 55, and 56 fell below the person with the lowest ability, with items 14, 53, 54, 55, and 56 far below the lowest person's ability, suggesting that these items should be made more difficult or removed from the test. Related to the visual benefit of seeing how the items and persons match on the logit scale, the Rasch measurement model places items in a hierarchy along the logit scale (from difficult at the top too easy at the bottom) which allows test makers to make *a priori* hypotheses about the difficulty of questions on the test (Beglar, 2010), representing another way to confirm the validity of the test.

Finally, the Rasch measurement model is able to measure unintended constructs within a test. In the earlier example with John and Mary, if a third question was on the test, such as "*What is the composition of water?*", the Rasch measurement model is able to identify this as a chemistry question, and not a physics question (even if the test-maker has not realized this). This is referred to as dimensionality and can be especially useful for teachers and researchers who are making tests and surveys that should focus on one construct. All tests and surveys are multidimensional to some degree (Baghaei & Amrahi, 2011), but the Rasch measurement model can identify exactly how much multidimensionality is present in a test, and it is up to the test-maker to decide if this amount of multidimensionality is tolerable (Baghaei & Amrahi, 2011; Runnels, 2012).

The use of the Rasch measurement model to assess students or validate tests and surveys has become more common in the TESOL field (Baghaei & Amrahi, 2011; Baghaei & Carstensen, 2013; Beglar, 2010; Cox & Clifford, 2014; Huhta, Alanen, Tarnanen, Martin, & Hirvela, 2014; McNamara, 2011; Runnels, 2012; Tiffin-Richards & Pant, 2013; Wu & Dou, 2015). For teachers who want to more accurately assess students or improve the validity of their tests, it is important to understand the basic principles of the Rasch measurement model. This paper will guide readers through the process of making and assessing a test with the Rasch measurement model.

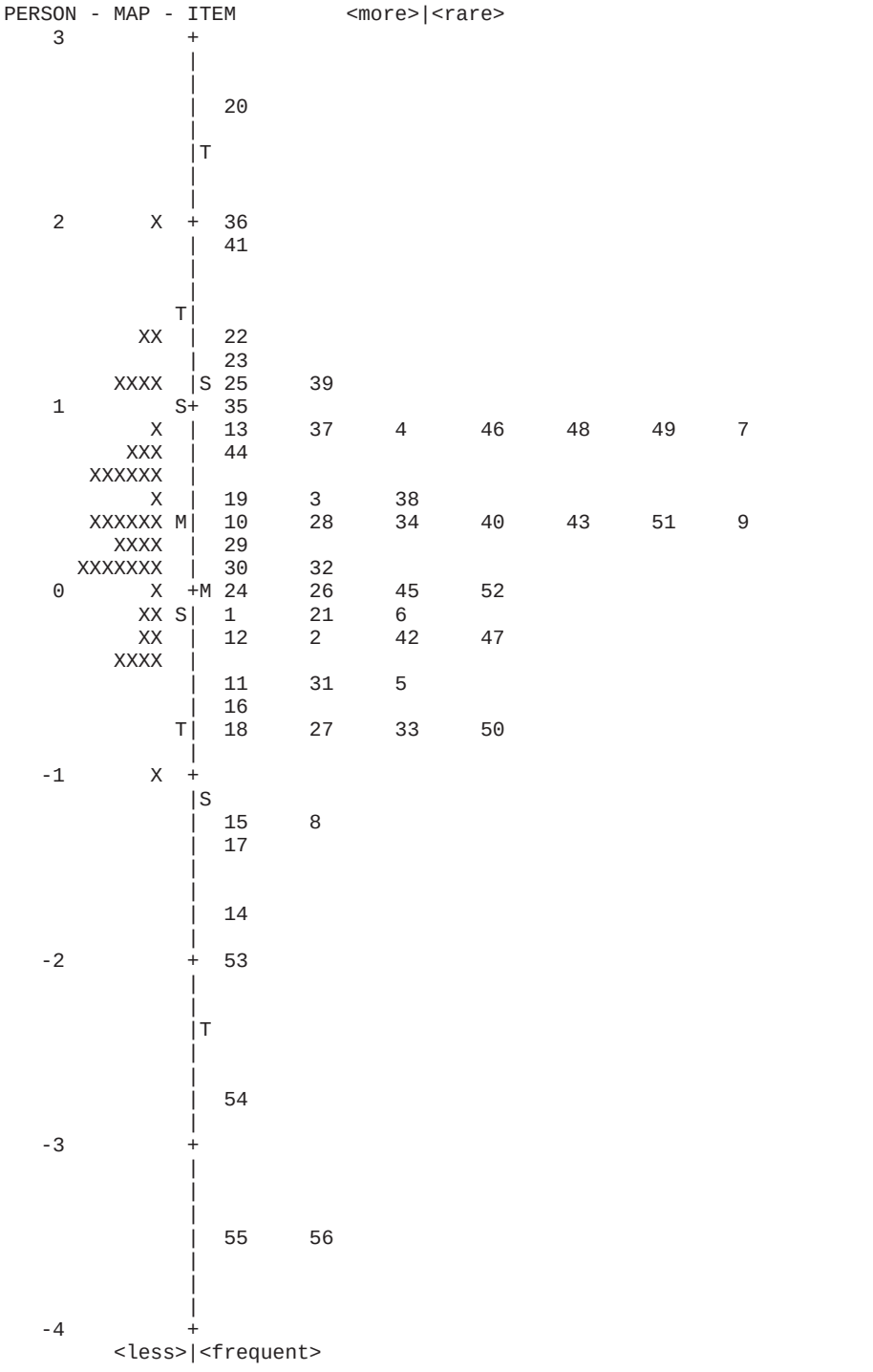

Figure 1. Wright Map for High-Advanced Test

Research Goals

Besides explaining the Rasch measurement model, the goal of this study was to give an example of test creation and assessment. Two separate exams were created for this study, for two groups of advanced students.

Having two levels of students within the advanced level (a high-advanced group and a low-advanced group) created a dilemma in how to fairly assess students. It was necessary to give all students in the advanced level a final exam, but if the exam was too difficult, it would punish the low-advanced group. Conversely, if the exam was too easy, it would not be challenging enough for the high-advanced group. If two distinct exams were created, one for each group, it would lead to distorted grades when comparing the two groups of students. For example, should a student in the low-advanced group who scored a 90% on the easier exam be considered equal to a student in the high-advanced group who scored a 90% on the more difficult exam? How much should the former student's exam score be discounted so a fair comparison could be made with the latter student? Because the Rasch measurement model can collectively assess the relative difficulty of questions on an exam, if the two exams shared several items (illustrated in Figure 2), it would be possible to accurately compare the two groups of students, even if the exams were significantly different in difficulty level (albeit with some shared items).

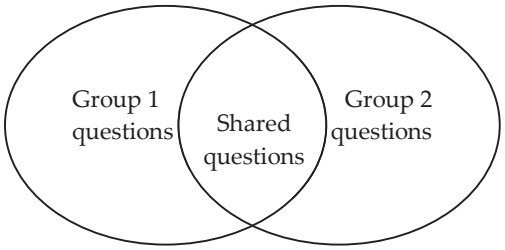

Figure 2. Linking two tests together

When two tests share items, and all items (shared and non-shared) are computed simultaneously, it is known as *concurrent equating method*, one of three ways to link tests (Masters & Keeves, 1999). The concurrent equating method has been shown to have higher consistency and better measurement of items (Baker & Al-Karni, 1991).

After the tests were given, a simple Rasch analysis was conducted on the test data to confirm the validity of the test's questions.

Participants

This research included 121 first-year students in the advanced English level of an intercultural communication program at a large private university in Tokyo. Students were drawn from five different listening classes. Within the advanced level, there were two groups of students: a *low-advanced* and a *high-advanced* group. The *low-advanced* group included 76 students from three classes and had TOEFL iBT scores

roughly in the range of 55-65, while the *high-advanced* group included 45 students (some of whom were returnees) from two classes and had TOEFL iBT scores roughly in the range of 65-80. Because students were in the same level (advanced), they needed to be graded together. However, because there was a significant difference in the ability between the two groups, they could not take the same test (a single test would be too difficult for the low-advanced group, or too easy for the high-advanced group). Using the Rasch measurement model to link two tests with several shared questions would solve this problem.

Instruments

Separate tests were created for the low-advanced and high-advanced groups in a listening course with each test including 56 multiple choice questions. There were 26 questions that were shared between the two tests, and there were 30 questions that were exclusive to each test.

Each test included two vocabulary and seven listening comprehension sections. The questions that were the same on both tests included the two vocabulary sections and two listening comprehension sections, which were based on content from the course textbook. The questions that were exclusive to each test included five listening comprehension sections and were based on content taken from the website www.ted.com.

Procedures

Making level-appropriate tests. The criteria for the tests were that they would take one hour to complete, use some of the textbook's content, test the listening ability of students, and be easy to grade because over 120 students would need to be assessed.

First, because listening passages would need to be included within the test's one-hour time limit, only 25 minutes would be available for answering questions (with 35 minutes for listening passages). It was thought that 56 multiple test questions would be suitable for the test (giving students around 30 seconds to answer each question).

Second, some teachers suggested that a quarter of the questions be vocabulary questions. A quarter of the 56 questions would be around 13-14, leaving approximately 42 for listening comprehension. If 42 questions were reserved for listening comprehension, and seven listening passages would be used in the test, then each listening passage would include six comprehension questions. Ultimately, the test had 56 total questions, of which 14 were vocabulary questions, and 42 were listening comprehension questions.

Third, five-minute listening passages from the website www.ted.com that were the appropriate difficulty level for the low-advanced and high-advanced groups were used in the test. The website at www.ted.com has an extensive library of videos that are available for copyright-free download. Ten listening passages that were roughly five minutes in length were used, with the five that seemed to be easier assigned to the low-advanced test, and the five that seemed to be more difficult assigned to the high-advanced test.

Finally, each of the 56 questions followed a multiple-choice format, which allowed for easy scoring of the test.

Generating data. When using the Rasch measurement model to assess whether the tests were appropriate for each of the groups, it was first necessary to generate data.

To generate data, the test responses must first be entered into a simple text file, and then the text file must be processed with the software *Winsteps 3.68* (Linacre, 2009). An example of a text file with response data is shown in Figure 3.

```
;This is file "G-level Listening Test, Fall 2015-16 (all items)"
TITLE= G-level Listening Test, Fall 2015-16
NI= 86
ITEM1= 1
NAME1= 88
ITEM= ITEM
PERSON= PERSON
CODES= ABCDEFGHIJ
KEY1=
CDEAGBFDBDADBCCBDCABCBCCBBDABACCCCCBAACDABDCCBDACADADCCAADABBCCDBABABACBCDACBDADJBHIEC
&END
;
1 aspirations
2 generate
3 precursor
4 pundit
5 rage
6 revenue
7 well intentioned
8 What was the main theme of this lecture?
9 Which event marked the beginning of mainstream acceptance of hip hop?
END NAMES;
CDEAGBFDBDCCBCDADCABCBCDBDDBBBCBCCCAAACDCBDDCBBAD                    DJBHIEC A1Bob Harris
DFABCGEDBABBCCDBDCBACDCBDBCABACCCBCADABDBBDBBABAC                    DJGHIEC A1 Herman Blume
DFABCGECCBADBCCBDABACCDCDCDABAAACDDBDBCDACBBCDABD                    DABHIEC A1 Ernie McCracken
DFBGCAEDBDAACCCBDCABADCCABCBBACACCADDDAABBBABCDCD                    DJBHIEC A1 Phil Connors
CDEAGBFDABAABCCDDCA          DBCAABADAABDDCDBCAADCCDCDACCDADJBHIEC A3 Peter Venkman
CDEFGBADCBABBCCAACA          ADADCAADDCBBDADBCDADBCBACADCAAGJBHIEC A3 Bob Wiley
CDAEFBGDBCADBCCBDCD          ACDDCAAADDBDDCABDBABBACACBCBDADABHIEC A3 Frank Cross
CDEAGBFDCDABBCCCDAA          AADDABAACDBBBCACBDACBBAAACCBDDFGDHIEC A3 Grimm
```

Figure 3. Example of Winsteps command file

A complete Winsteps manual with dozens of example text files can be downloaded from the Winsteps website as a .pdf file. The example text file in Figure 3 is relatively straightforward and is explained below. A completed text file is referred to as a command file.

Winsteps Command File

At the top of the command file is the name of the text file, followed by the title of the data (neither of these are essential to your analysis). Next are the headings "NI", which indicates the number of items in the test, "ITEM1", which indicates the space where the item responses will begin, and "NAME1" which indicates the space where person names will begin. This is followed by "ITEM", which indicates the term used for the test's questions, "PERSON", which indicates the term used for the people completing the test, and "CODES", which indicates the range of possible answer choices for the test's questions (on the tests in this study, the vocabulary questions had answer options from A-J while the listening comprehension questions had answer options from A-D). This is followed by "KEY1", which indicates the correct answer choices for all of the items on the test (the first 19 answers were for shared questions, the next 30 answers were for the high-advanced test, the next 30 answers were for the low-advanced test, and the final 7 answers were for shared questions), "&END;", which

is necessary code to end this portion of the command file, and, finally, the listing of all of the items.

In the example command file, only the first nine items on the test were listed because listing all 86 items would have required too much space for this article. Of note, spelling does not need to be perfect because these are only labels that will be used in the data output, and as long as the test-maker can identify the item, items do not need to be spelled perfectly (hence the spelling error in item nine). If the test-maker wants, the item can be labelled with a number rather than the full question. When the list of items is finished, "ENDNAMES;" should be included, followed by the specific responses for each student on the test. For example, the first student listed was labelled as "A1 Bob Harris" (a pseudonym). This identified the student as being in class A1 (the high-advanced group) with the name Bob Harris. Bob answered the first 49 items on the test as "C" for item 1, "D" for item 2, "E" for item 3, "A" for item 4, and so on, then did not answer items 50-79 (because these questions were only on the low-advanced test), and then answered items 80-86. The last response was followed by a space, and then the students' identifier (in this case, their class and name). In the example command file, only some students who took the test were listed because listing all 121 students would have required too much space for this article. For an example of a student from the low-advanced group, the fifth student listed was labelled as "A3 Peter Venkman" (a pseudonym). This identified the student as being in class A3 (the low-advanced group) with the name Peter Venkman. Peter answered the first 19 items, then did not answer items 20-49 (because these questions were only on the high-advanced test), and then answered items 50-86.

To run the command file in Winsteps, open Winsteps, go to *File* from the drop-down menu, then select the *Open File* option. Next, a dialog box will open, and then select the command file. Once the command file has been selected, press the *Enter* key twice and Winsteps will generate the Rasch data.

Assessing the Data

When assessing the Rasch data generated by Winsteps, there are several variables that should be examined. An example of the variables produced by Winsteps is shown in Table 1 (see pp. 84-92).

Winsteps allows for the Rasch data to be analyzed in several different ways, such as examining the ability and behaviour of the people who completed the tests or examining the difficulty and reliability of the items on the test. The data shown in Table 1 is an examination of the difficulty and reliability of the items on the test. This data can be obtained by going to the *Output Files* drop-down menu in Winsteps and then choosing the *ITEM File = IFILE* option. Next, a dialog box will open, and the user will be given some choices on how the output should be generated (such as in an Excel file, a text file, or an SPSS file). Unless the user has experience with SPSS, it is probably easiest to choose the Excel file option (a text file will not allow the data to be easily viewed by the user). The Excel output file will include 17 columns of data. Not all of this data is essential for analysis, so only ten columns of data have been included in Table 1.

Table 1

Item Statistics by Measure

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
52	2.43	76	9	0.36	1.06	0.32	1.12	0.49	22b Why does the speaker use the example of the brain producing pain after the body is burned?
20	2.37	45	5	0.48	1.04	0.22	1.49	1.12	20a What was the main theme of this lecture?
62	1.81	76	15	0.29	0.96	-0.19	0.96	-0.12	32b What was the main theme of this lecture?
66	1.81	76	15	0.29	1.02	0.16	1.03	0.21	36b Why does the speaker feel we should change our model"?"
36	1.79	45	8	0.40	1.04	0.26	1.30	0.99	36a According to the speaker, what causes Alzheimer's disease?
55	1.73	76	16	0.29	1.04	0.29	1.08	0.49	25b Which movie does the speaker refer to?
41	1.64	45	9	0.38	1.02	0.15	1.04	0.24	41a In the speaker's story about his own research, what was the problem?
70	1.29	76	22	0.26	1.07	0.65	1.16	1.25	40b According to the speaker, why are governments upset?
71	1.29	76	22	0.26	1.10	0.95	1.12	0.91	41b Which surveillance example was described by the speaker?

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
73	1.22	76	23	0.25	1.06	0.62	1.10	0.82	43b According to the speaker, what is the best way to communicate?
58	1.16	76	24	0.25	1.08	0.87	1.12	1.08	28b According to the speaker, her brother Samuel...
72	1.16	76	24	0.25	1.08	0.86	1.08	0.76	42b What does the speaker suggest for the future?
75	1.16	76	24	0.25	1.00	-0.01	1.01	0.13	45b What is the main problem with using pills?
22	1.13	45	13	0.34	0.95	-0.29	1.08	0.46	22a What was NOT an example of ingenuity by the prisoners?
69	1.09	76	25	0.25	0.97	-0.31	1.00	0.04	39b What are the two main opposing forces identified by the speaker?
23	1.01	45	14	0.33	1.06	0.51	1.11	0.71	23a What is the speaker's reason for many released criminals going back to prison?
63	0.97	76	27	0.24	0.97	-0.32	0.98	-0.14	33b The air inside buildings...
25	0.90	45	15	0.33	1.02	0.23	1.08	0.54	25a Why should society help prisoners more?
39	0.90	45	15	0.33	0.85	-1.19	0.83	-1.16	39a What experience does the speaker describe at the beginning of his lecture?

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
35	0.80	45	16	0.32	1.02	0.24	1.01	0.16	35a What is NOT mentioned as a symptom of Alzheimer's disease?
51	0.80	76	30	0.24	1.06	0.88	1.07	0.87	21b The speaker says that there are three ways to change the brain. What is NOT mentioned?
74	0.80	76	30	0.24	0.98	-0.35	0.97	-0.43	44b What was the main theme of this lecture?
37	0.70	45	17	0.32	1.09	0.87	1.14	1.12	37a According to the speaker, what is the challenge in curing Alzheimer's disease?
46	0.70	45	17	0.32	0.79	-2.12	0.76	-2.12	46a Which negative aspect of meetings is NOT mentioned by the speaker?
48	0.70	45	17	0.32	0.94	-0.60	0.97	-0.21	48a What does the speaker suggest that we do?
19	0.65	121	50	0.19	1.13	2.48	1.15	2.42	19 When mediating, the parties involved must...
49	0.60	45	18	0.32	0.90	-1.08	0.87	-1.20	49a What is NOT mentioned as a way to improve efficiency?

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
77	0.51	76	35	0.24	0.97	-0.51	0.97	-0.55	47b Which example of lasers is NOT mentioned by the speaker?
44	0.50	45	19	0.31	0.94	-0.64	0.93	-0.66	44a What was the main theme of this lecture?
9	0.47	121	55	0.19	0.94	-1.25	0.94	-1.13	9 Which even marked the beginning of mainstream acceptance of hip hop?
50	0.46	76	36	0.24	1.00	-0.07	0.99	-0.17	20b What was the main theme of this lecture?
4	0.40	121	57	0.19	0.90	-2.46	0.89	-2.25	4 pundit
13	0.40	121	57	0.19	1.00	-0.03	0.99	-0.15	13 According to Dr. Lee, hip hop culture has gone beyond the music to focus on a lifestyle which includes...
59	0.35	76	38	0.23	1.04	0.74	1.04	0.73	29b How does the speaker define autism?
10	0.33	121	59	0.19	1.02	0.49	1.02	0.45	10 Which fashion trend was NOT mentioned by Dr. Lee as part of hip hop fashion?
38	0.31	45	21	0.31	0.94	-0.74	0.92	-0.89	38a What was the main theme of this lecture?
28	0.21	45	22	0.31	1.15	1.95	1.18	2.02	28a What is a negative aspect to colonizing Mars?

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
43	0.21	45	22	0.31	0.89	-1.49	0.87	-1.55	43a What does the speaker say is the real challenge?
12	0.12	121	65	0.19	1.06	1.30	1.06	1.26	12 What is NOT mentioned by Dr. Lee when he explains the beginning of hip hop?
34	0.11	45	23	0.31	1.10	1.28	1.11	1.23	34a How much are Alzheimer's disease medical costs expected to increase by 2050?
40	0.11	45	23	0.31	1.05	0.63	1.04	0.48	40a What did the speaker realize after this experience?
61	0.07	76	43	0.24	1.04	0.71	1.04	0.64	31b What is the speaker's attitude towards autism?
29	0.02	45	24	0.31	0.98	-0.18	0.96	-0.37	29a How can we develop our understanding of planetary colonization?
3	-0.02	121	69	0.19	0.87	-2.78	0.86	-2.70	3 precursor
7	-0.02	121	69	0.19	0.90	-2.13	0.91	-1.74	7 well intentioned
82	-0.06	121	70	0.19	1.10	1.90	1.10	1.71	52 Contrived
30	-0.08	45	25	0.31	0.94	-0.73	0.94	-0.56	30a According to the speaker, which idea best represents Fermi's Paradox?

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
32	-0.08	45	25	0.31	1.09	1.20	1.18	1.81	32a What was the main theme of this lecture?
68	-0.10	76	46	0.24	1.00	0.02	1.01	0.20	38b What was the main theme of this lecture?
78	-0.10	76	46	0.24	1.00	-0.03	1.00	-0.01	48b Which process is NOT described as part of the three-headed device"?"
81	-0.13	121	72	0.19	1.04	0.69	1.06	0.97	51 Contingency
76	-0.16	76	47	0.24	0.95	-0.66	0.95	-0.66	46b According to the speakers, where are HIV reservoirs NOT located?
45	-0.17	45	26	0.31	1.02	0.24	1.01	0.10	45a What is the main purpose of the stolen chair example at the beginning of the lecture?
64	-0.22	76	48	0.24	1.01	0.20	1.01	0.15	34b Which activity is NOT mentioned as part of mechanical ventilation?
24	-0.27	45	27	0.31	1.01	0.19	1.03	0.31	24a How many criminals commit a crime within five years of being released?
26	-0.27	45	27	0.31	0.92	-0.88	0.91	-0.75	26a What was the main theme of this lecture?

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
57	-0.28	76	49	0.25	0.98	-0.28	0.96	-0.40	27b According the speaker, her brother Remi...
21	-0.37	45	28	0.32	1.08	0.84	1.09	0.77	21a Which business activity occurring in prison was NOT mentioned by the speaker?
18	-0.39	121	79	0.20	1.04	0.62	1.06	0.74	18 How much does a litigated divorce usually cost?
79	-0.46	76	52	0.25	0.96	-0.35	0.95	-0.41	49b What is the goal of the speaker's plan?
16	-0.47	121	81	0.20	1.05	0.76	1.09	1.03	16 According to Dr. Mayfield, what is the main difference between mediation and litigation?
42	-0.47	45	29	0.32	0.99	-0.07	0.98	-0.12	42a What was the point of the speaker's story about his research?
47	-0.47	45	29	0.32	0.84	-1.57	0.83	-1.24	47a How many views does the speaker's video have?
54	-0.53	76	53	0.26	1.10	0.94	1.13	1.07	24b What was the restriction mentioned by the speaker at the end of the lecture?

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
65	-0.59	76	54	0.26	1.02	0.25	1.02	0.23	35b According to the speaker, where does the healthcare industry rank in energy use?
67	-0.59	76	54	0.26	0.96	-0.29	0.94	-0.44	37b Which government department did the speaker compare hospitals to?
11	-0.67	121	86	0.21	1.01	0.19	1.04	0.38	11 When was the best time for hip hop?
80	-0.67	121	86	0.21	1.04	0.47	1.02	0.23	50 Appalled
31	-0.69	45	31	0.33	1.03	0.27	1.01	0.11	31a How many planets does the speaker say are in our galaxy?
53	-0.73	76	56	0.27	1.00	0.07	0.97	-0.16	23b The speaker mentioned specific research involving the brain. How much was the decrease in pain for the people in the research study?
33	-0.91	45	33	0.35	1.10	0.63	1.12	0.63	33a Which medical problem does the speaker NOT use as an example of research progress?
1	-0.94	121	92	0.22	0.93	-0.60	0.89	-0.81	1 aspirations
6	-1.03	121	94	0.22	0.86	-1.23	0.77	-1.68	6 revenue

Table 1 Cont.

Entry	Measure	Count	Score	Error	IN MSQ	IN ZSTD	OUT MSQ	OUT ZSTD	Item
60	-1.03	76	60	0.29	0.94	-0.30	0.95	-0.22	30b Which area has the speaker NOT learned about through her brothers?
27	-1.04	45	34	0.36	0.95	-0.23	0.87	-0.52	27a What does the Kepler data NOT reveal about a distant planet?
2	-1.24	121	98	0.24	0.87	-0.95	0.76	-1.52	2 generate
5	-1.30	121	99	0.24	0.91	-0.56	0.83	-0.99	5 rage
83	-1.42	121	101	0.25	0.99	0.01	0.91	-0.40	53 Genre
8	-1.55	121	103	0.26	1.04	0.31	1.03	0.21	8 What was the main theme of this lecture?
17	-1.55	121	103	0.26	1.04	0.30	1.13	0.66	17 Which is NOT described as a benefit of mediation?
14	-1.62	121	104	0.27	1.02	0.16	0.96	-0.13	14 What was the main theme of this lecture?
15	-1.62	121	104	0.27	1.00	0.05	0.94	-0.18	15 How does the lecturer initially describe the mediation process?
86	-2.62	121	114	0.39	1.03	0.21	1.12	0.43	56 Wacky
56	-2.63	76	72	0.52	0.99	0.13	0.90	-0.03	26b What was the main theme of this lecture?
84	-2.98	121	116	0.46	0.99	0.11	1.04	0.23	54 Give-and-take
85	-4.63	121	120	1.00	0.96	0.28	0.27	-0.58	55 Trend

The first column is labelled *Entry,* and this represents the order that questions were entered into the command file. Recall that there were 86 total items in the command file, so the first row, labelled *52*, is the 52nd item entered into the command file.

The second column is labelled *Measure,* and this represents the difficulty level of each item. Because this study is attempting to make a more difficult test for the high-advanced group, this column's information is very important. In the first row, the 52nd item entered into the command file, which was question 22 on the low-advanced test, had a difficulty measure of 2.43. This was the highest difficulty measure for all of the items on both tests, which means it was the most difficult question. We can immediately see a problem in that the low-advanced test should not include the most difficult questions. Of the 13 most difficult questions, ten were from the low-advanced test (the numbers accompanied with a *b* in the tenth column *Item* indicate questions on the low-advanced test). When we modify this test, these items should either be made easier, deleted, or switched to the high-advanced test.

The third column is labelled *Count,* and this represents the number of students who answered this item. Items either had *45,* which was the number of students answering high-advanced questions, *76,* which was the number of students answering low-advanced questions, or *121,* which was the number of students answering shared questions.

The fourth column is labelled *Score,* and this represents the total number of students who answered this question correctly. For example, in the first row, the 52nd item, which was question 22 on the low-advanced test, was answered correctly by nine students. Conversely, in the third-last row, the 56nd item, which was question 26 on the low-advanced test, was answered correctly by 72 students. This column gives some indication of the difficulty of each item, however, this variable is not weighted and represents a CTT type of assessment.

The fifth column is labelled *Error* and this represents the accuracy of the difficulty measure variable (which is shown in column two). The greater the error in column five, the less precise the difficulty measure, and high error is usually more evident in items that are either very easy or very difficult (because these items tend to be below or above the ability of most people, and as a result, are more difficult to assess).

The sixth column is labelled *IN MSQ* and represents the infit mean square, which indicates how well the actual responses matched the predicted responses of the Rasch measurement model. Put more simply, the Rasch measurement model can predict how items will be answered based on the answer patterns within the entire group. For example, if person A is answering all items correctly, and item 1 is the easiest item (because everyone is answering it correctly), the Rasch measurement model will predict that person A has a very good chance of answering item 1 correctly. Infit and outfit indicate how closely person A's actual responses match the predicted responses; a value of 1.0 indicates perfect fit (the actual response matches the predicted response). However, if person A unexpectedly answers item 1 incorrectly, this will be represented with higher infit and outfit values. A high infit and/or outfit for a *person* means that the person is answering unpredictably (perhaps because they are cheating, guessing, or having a problem). A high infit or outfit for an *item* means that the item is being answered unpredictably (maybe the question is worded in a confusing way, which is causing students to answer it inconsistently). Basically, the item IN MSQ measures how reliably a question is being answered. If the item IN MSQ is within the recommended range of 0.70 to 1.30 (Bond & Fox, 2007), then it usually indicates that

people understood the item correctly. However, if the item IN MSQ was outside of the recommended range, it usually indicates that something strange was happening when people were answering this item.

The seventh column is labelled *IN ZSTD* and also represents the infit value of the item; however, it is standardized to minimize distortion that could occur because of the sample size. For example, fit problems are sometimes not obvious in the IN MSQ variable when the sample size is very large, while fit problems are always obvious in the IN ZSTD variable. IN ZSTD should fall within the range of -2 to +2 (Baghaei & Amrahi, 2011). If the IN ZSTD falls below this range, it is said to overfit the model, which indicates items that followed the Rasch model predictions too much (i.e. answer patterns were too predictable). If the IN ZSTD is above this range, it is said to underfit the model, which indicates items that did not follow the Rasch model predictions enough. Underfit is regarded as more of a problem than overfit.

The eighth column is labelled *OUT MSQ*, and the ninth column is labelled *OUT ZSTD*. Like infit, outfit gives an indication of how well the actual responses matched the Rasch model's predicted responses. The difference between outfit and infit is that outfit weighs all items equally, whereas infit more heavily weighs nearby items (Sadiq et al., 2015). As a result, researchers tend to prefer infit over outfit because infit is not as vulnerable to skewed data that stems from extreme unpredictability (such as a person with very high ability incorrectly answering a very easy question).

Finally, the tenth column is labelled *Item* and represents the label given to each item in the Winsteps command file. For the two tests in this study, shared items were labelled with a number, low-advanced test items were labelled with a number and a *b* (for example, the item in the first row is *22b* which represents question 22 on the low-advanced test), and high-advanced test items were labelled with a number and an *a*.

Results

To confirm that the tests were set at the appropriate difficulty level, it was necessary to compare the difficulty estimates of the low-advanced test sections to those of the high-advanced test. The average difficulty estimates for each section of each test are shown in Table 2 on p. 95, with higher difficulty estimates indicating more difficult sections, and lower difficulty estimates indicating easier sections.

Difficulty estimates of the shared item sections of vocabulary 1, listening comprehension 1, and listening comprehension 2 were relatively similar, at -0.59, -0.15, and -0.83, respectively. However, the difficulty estimates for the shared item section of vocabulary 2 was much lower at -2.09, indicating that the questions in this section might have been too easy.

Looking at the average difficulty estimates of the low-advanced sections, the listening comprehension 3 (0.69), listening comprehension 5 (0.53), and listening comprehension 6 (0.99) sections were more difficult than all but one of the high-advanced sections (listening comprehension 3 at 0.80). In particular, low-advanced's listening comprehension 6 section was the most difficult section on either test, and would need to be made easier, deleted, or switched to the high-advanced test.

Table 2
Average Item Difficulty by Test Section

Item entry numbers	Type of items	Test section	Average Difficulty Measure
1-7	Shared	vocabulary 1	-0.59
8-13	Shared	listening comprehension 1	-0.15
14-19	Shared	listening comprehension 2	-0.83
20-25	high-advanced	listening comprehension 3	0.80
26-31	high-advanced	listening comprehension 4	-0.31
32-37	high-advanced	listening comprehension 5	0.42
38-43	high-advanced	listening comprehension 6	0.45
44-49	high-advanced	listening comprehension 7	0.31
50-55	low-advanced	listening comprehension 3	0.69
56-61	low-advanced	listening comprehension 4	-0.39
62-67	low-advanced	listening comprehension 5	0.53
68-73	low-advanced	listening comprehension 6	0.99
74-79	low-advanced	listening comprehension 7	0.29
80-86	Shared	vocabulary 2	-2.09

Looking at the difficulty estimates of items on the low-advanced test, items with a difficulty measure of +1.0 were considered as being excessively difficult for that test, and any section with three or more excessively difficult items would need to be switched to the high-advanced test. In the listening comprehension 3 section, item 52 (2.43) and item 55 (1.73) were excessively difficult and would need to be made easier. In the low-advanced listening comprehension 4 section, item 58 (1.16) was excessively difficult and would need to be made easier. In the low-advanced listening comprehension 5 section, item 62 (1.81), item 66 (1.81), and item 63 (0.97) were excessively difficult and the entire section would need to be switched to the high-advanced test. In the low-advanced listening comprehension 6 section, item 70 (1.29), item 71 (1.29), item 73 (1.22), item 72 (1.16), and item 69 (1.09) were excessively difficult,

and the entire section would need to be switched to the high-advanced test. In the low-advanced listening comprehension 7 section, item 75 (1.16) was excessively difficult and would need to be made easier.

Looking at the average difficulty estimates of the high-advanced sections, the listening comprehension 4 (-0.31) section was easy when compared to the other sections. Also, while the listening comprehension 5, listening comprehension 6, and listening comprehension 7 sections were not easy, they were easier than several sections on the low-advanced test, and this would need to be corrected.

Looking at the difficulty estimates of items on the high-advanced test, items with a difficulty measure of -1.0 were categorized as being excessively easy, and any section with three or more excessively easy items would need to be switched to the low-advanced test. In the listening comprehension 3 section, there were no excessively easy items. In the listening comprehension 4 section, item 27 (-1.04) was excessively easy. Despite having only one excessively easy item, the other five items were still easy when compared to items in other sections (as shown in Table 2), and thus, this section should be switched to the low-advanced test. In the listening comprehension 5 section, there were no excessively easy items. In the listening comprehension 6 section, there were no excessively easy items. Finally, in the listening comprehension 7 section, there were no excessively easy items.

A summary of item and section violations of difficulty estimate guidelines is shown in Table 3 on p. 97.

To confirm the validity of the items, one approach (among many) is to look at the fit value for each item and make sure that they fell within the recommended guidelines (0.70 to 1.30 for IN MSQ and OUT MSQ, or -2 to +2 for IN ZSTD and OUT ZSTD).

Looking at the infit of the items, there were no items that violated the guideline for IN MSQ; however, there were several items that violated the guideline for IN ZSTD, specifically, item 3 (-2.78), item 4 (-2.46), item 7 (-2.13), item 19 (2.48), and item 46 (-2.12).

Looking at the outfit of the items, two items violated the guideline for OUT MSQ, specifically, item 20 (1.49) and item 36 (1.30). There were several items that violated the guideline for OUT ZSTD, specifically, item 3 (-2.70), item 4 (-2.25), item 19 (2.42), item 28 (2.02), and item 46 (-2.12).

A summary of item violations of fit guidelines is shown in Table 3.

Discussion

The results of the analysis done on the two tests show why it is important for teachers to check the validity of their tests. Despite having experience in constructing listening exams over several years, the researcher still made several incorrect assumptions about the questions on both tests. The researcher misjudged the difficulty level of seven items, as well as three entire sections (18 items). Combined, this represents 25 out of a possible 86 items, almost a third of all items. Further to this point, the Rasch measurement model indicated that eight items had poor fit, likely indicating poorly-worded questions or answers. The Rasch measurement model identified these problems whereas CTT would not have, which should result in an improved second version of the test.

Table 3
Summary of Item and Section Violations

Item or Section	Violation	Course of Action
Item 84	Too easy	Make more difficult
Item 85	Too easy	Make more difficult
Item 86	Too easy	Make more difficult
Item 52	Too difficult	Make easier
Item 55	Too difficult	Make easier
Item 58	Too difficult	Make easier
Listening comprehension 5 section, low-advanced test	Too difficult	Switch to high-advanced test
Listening comprehension 6 section, low-advanced test	Too difficult	Switch to high-advanced test
Item 75	Too difficult	Make easier
Listening comprehension 4 section, high-advanced test	Too easy	Switch to low-advanced test
Item 3	Overfit the model	Improve wording of item and answers
Item 4	Overfit the model	Improve wording of item and answers
Item 7	Overfit the model	Improve wording of item and answers
Item 19	Underfit the model	Improve wording of item and answers
Item 46	Overfit the model	Improve wording of item and answers
Item 20	Underfit the model	Improve wording of item and answers
Item 36	Underfit the model	Improve wording of item and answers
Item 28	Underfit the model	Improve wording of item and answers

While this study focused on the Rasch data concerning items, the Rasch data concerning persons can also provide valuable insights. The information gleaned from person fit statistics can help teachers identify students who may be answering erratically, either in a way that lowers a student's grade (such as nervousness, carelessness, or lack of focus) or increases a student's grade (such as guessing or

cheating). This information can alert the teacher to a course of action that might be necessary to help the students. Additionally, a teacher might inspect the Wright Map and realize that several items are in the same location along the vertical axis. This would indicate redundant items, and the teacher could delete several extraneous items and still have a valid test. Shorter tests that maintain their validity are more efficient and can free up class time for other activities that help students learn.

Benefits are not limited to teachers. Rasch can benefit learners by placing them in a class that is appropriate to their ability level. As indicated earlier, there is research that has demonstrated that students might be put in a different class based on whether their placement exam was scored with CTT or the Rasch measurement model. Being in a class that is too difficult (or too easy) can have potentially negative effects on a student's confidence, anxiety, and motivation, so it is essential for placement to be as accurate as possible. Additionally, the Rasch measurement model makes it easy to customize tests to a specific ability level, as was illustrated in this article for low-advanced and high-advanced students. Occasionally, schools will create a single standardized exam that every student must take, but this can have a negative effect on lower-proficiency students as their confidence can be damaged when taking a test that is well-beyond their ability. Linking two tests that place all students on the same grading scale can help teachers preserve the confidence of lower-proficiency students by giving them a test in which they can succeed.

The information gleaned from person fit statistics can help teachers identify students who may be answering erratically…

Finally, the research community can benefit from the Rasch measurement model. Many assumptions have been made about how motivation, anxiety, personality, and other affective variables relate to learning. However, if these assumptions are based on surveys and tests that had poor validity, then the conclusions drawn by this research may be false. For example, there has been relatively little research that has shown that personality influences language learning (Dewaele & Furnham, 1999), however if the personality surveys that were used to evaluate students had flawed items (indicated by item fit), or the language tests suffered from multidimensionality (and were not measuring what they were supposed to measure), then it is difficult to believe that personality really has no influence on language learning.

Suffice it to say, teachers, learners, and the research community can all benefit from greater use of the Rasch measurement model in education.

Conclusion

Testing is used in virtually all educational contexts around the world, in both limited (such as a class quiz) and broad ways (such as a common exam for an entire grade of students). With tests occupying such an important role in student assessment, it is essential that teachers ensure that their tests are as well-constructed as possible. When comparing raw scores (CTT) versus the information provided with the Rasch measurement model, there is so much to gain by using a Rasch approach. If it can be agreed upon that the Rasch measurement model provides better and more accurate

information than raw scores, then the only excuse for not using the Rasch measurement model is that the process might be too complicated. Hopefully, this paper has been able to simplify the process so teachers have a better understanding of how to conduct a basic Rasch analysis. The potential benefits of using the Rasch measurement model far outweigh the learning curve associated with the model.

References

Andrich, D. (1988). *Rasch models for measurement*. Newbury, CA: Sage.

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*(5), 1052-1060.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation, 18*(5), 1-13.

Baker, F. B., & Al-Karni,(1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147-162.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101-118.

Bond, T., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cox, T. L., & Clifford, R. (2014). Empirical validation of listening proficiency guidelines. *Foreign Language Annals, 47*(3), 379-403.

Dewaele, J. M., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning, 49*(3), 509-544.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvela, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307-328.

Linacre, J. M. (1997). KR-20/Cronbach alpha or Rasch person reliability: Which tells us the truth? *Rasch Measurement Transactions, 11*, 580-581.

Linacre, J. M. (2009). *Winsteps (Version 3.68)*. Beaverton, OR: Winsteps.com.

Masters, G. L., & Keeves, J. P. (1999). *Advances in Measurement in Educational Research and Assessment*. Amsterdam, Amsterdam: Pearson Education Inc.

McNamara, T. (2011) Applied linguistics and measurement: A dialogue. *Language Testing, 28*(4), 435-440.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Runnels, J. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies, 6*(4), 141-153.

Sadiq, M., Tirmizi, S. H., & Jamil, M. (2015). Using Rasch model for the calibration of test items in mathematics, grade 9. *Journal of Research and Reflection in Education, 9*(2), 82-102.

Tiffin-Richards, S. P., & Pant, H. A. (2013). Setting standards for English foreign language assessment: Methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice, 32*(2), 15-25.

Weaver, C., Jones, A., & Bulach, J. (2008). Comparing placement decisions based on raw test scores and Rasch ability scores. *The Language Teacher, 32*(6), 3-8.

Wu, S., & Dou, T. (2015). Validation of an oral English test based on many-faceted Rasch model. *Journal of Language Teaching and Research, 6*(4), 866-872.

Dr. Omar Karlin is an Assistant Professor at Toyo University, where he teaches English language courses. He obtained his Doctorate in Education in 2015 from Temple University, and his research interests include test construction and validation, the intersection of personality and language proficiency, and teaching listening.

Sayaka Karlin is an Adjunct Professor at Toyo Gakuen University, where she teaches English language courses. She obtained her Master's in Economics from the University of Manchester, and completed her Master's in Education from Temple University in 2017. Her research interests include vocabulary, reading proficiency, and teaching adults.