Classical test theory or Rasch: A personal account from a novice user

Jim Smiley
sendaismiley@gmail.com
Tohoku Bunka Gakuen University

Abstract

Educators have utilized Classical Test Theory (CTT) when developing instruments for measuring and assessing pedagogic data. Results derived from standard CTT analysis methods offer valuable insights into the effectiveness of language assessment tools. Tests undergo a series of steps running from initial draft production through test trialing to test revision. Such instruments produced using this method can be shown to be more internally consistent and deliver more valid results than tests that are written *ad hoc* and informed by intuitive rationale. More recently, the Rasch model has gained a following among test developers as an alternative procedure in refining testing vehicles. CTT contrasts with the Rasch model in a number of key areas, differences that, when utilized in the analysis of a test, result in the production of a more internally valid test. This article questions the need for a materials developer to change to Rasch given that the learning curve is steep considering the additional investment of the time necessary to become proficient using Rasch. The conclusion is that Rasch data provides very detailed information that is *sine qua non* for long-term test instrument refinement and materials development, and that CTT data may be enough to begin the test of the test.

Keywords: Rasch modelling, Classical Test Theory, comparisons, test analysis

In this article, I document my ultimately rewarding and educational experience of attempting to broach the topic of Rasch modelling. The learning curve was extremely steep, and even after months of grappling with the core concepts, I can only lay claim to have scratched the surface and to have understood only the merest sampling that the fuller knowledge of Rasch offers. My story may be of interest to those TEVAL members who haven't yet given the Rasch model a go. However for the majority of readers, please let these words act as a guide to the frustration, the angst, the terror that many of your numerical literacy challenged colleagues feel when faced with a bewildering array of figures, of equations, of being asked to grapple with numbers.

I found the introductions to Rasch maddening in their assumptions about the readership of these primers. To qualify this statement, I need to describe my background and then tease out some of the gaps between what was expected of me and what the introductory guides expected. For many years, I have relished opening Microsoft Excel after the exam sessions to collect basic descriptive statistics. After that, I use R to create boxplots, histograms and other visuals to share information among the other language teachers. Usually, I do simple item difficulty and item discrimination analysis to find possible issues in the test construction, to highlight items that are problematic for various reasons. My Excel template has rows for student data, for total correct, for percentage correct on each question option, and so on. I can see at a glance, for example, that Question 7 (a four-option multiple choice question) was answered correctly by 67% of students, 21% choosing option A, 8% choosing B, and 4% choosing D. Both Excel and R give me (albeit with slightly different definitions and therefore results) things like the mean, standard deviation, the quartiles and so on. Books like Brown's (2005) *Testing in Language Programs* and the older tutorial book *Testing for Language Teachers* by Hughes (1989) don't faze me at all. I devoured those texts.

For more complex statistics, I use R. After making sure the conditions are met for the various tests, I generate p-values for t-tests, regression models, chi-squares and so on. I can't describe without referencing a statistics book, for example, what the conditions for ANOVA are, or the exact cut-off value that means I need to use non-parametric tests. And to learn the underlying equations for these tests would require that I study mathematics that I haven't touched for over 30 years! But I had to do that when I opened the first chapter of Bond and Fox's *Applying the Rasch Model* (2007). Without defining key terms, they render their book opaque to the uninitiated. Holster and Lake (2014) presented the basics in a much more readable form, yet by the second page, terms such as *over-fit*, *stochastic*, and *deterministic data* appear. The assumption behind these inclusions must be that the general meanings of the terms outside testing cover the specialist meaning sufficiently. For this reader, I'm afraid that they don't. *Shiken* has a set of resources aimed at introducing Rasch measurement to members (Sick, 2008a, 2008b, 2009, 2010). Once again, the opening pages read as text written for insiders. The manual that came with the Winsteps software needs at least high-school algebra to comprehend. I appreciate the fact that there are concepts, techniques and methods to be learnt. But perhaps there is a cultural gap also at play here, and the in-crowd either not realizing there is or not wanting to overcome the cultural divide.

The version of this article is the result of revisions after two anonymous reviewers commented on an earlier draft. I thank them from the bottom of my heart for their efforts. Both provided copious notes, suggestions for improvements, pointed out errors in my conceptualization of Rasch principles, and generally added significantly to my understanding of Rasch. Reviewers such as them add to the joys of learning. However, this article is still bound to amuse Rasch purists, who will certainly find many misunderstandings remaining. I would highly appreciate those errors to be pointed out and corrected in a later issue of *Shiken*. Perhaps if more novices were to share their stories, TEVAL may become more of a beginner-friendly SIG.

Total test scores from Traditional or Classical Test Theory (CTT) have been described as "simple raw scores" (Holster & Lake, 2014). A test-test taker's final score is obtained through the addition of their correct raw responses. This total and the total of all other test takers in a test session combine to produce data which the test developer uses for analysis. These "group-centred" scores form the basis for statistical analysis and "require the clustering of individuals into discrete categories or populations" (Choppin, 1983). The focus on the group allows for statistics that rely on the nature of that group, not on the specifications of the test instrument itself. For example, using data from a population or a sample, one can easily discover the interquartile ranges, the variance of the mean, whether or not the samples' means are statistically significantly similar or different and so on. With a different sample set, the figures returned deliver another set of statistics. These data provide the test developer with some tools to analyze the validity of the test, but they do not allow for a complete understanding of the validity.

This lack of interface between the test instrument and the resultant raw scores is problematic for test developers. Students are measured on the basis of what may have been a faulty test, yet the absence of technical analytic tools hinders the discovery of a potentially flawed test. Flaws also include reliability issues such as the test actually measuring what it tried to measure (construct validity), and the test question types targeting the skill appropriately (face validity) (Hughes, 1989, pp. 26-27), but a discussion of these is outside the scope of this paper.

Furthermore, test developers need to be able to test their tests independently of ability of the test takers. A stable test returns similar results irrespective of the particular group of students. It behooves the developer to ascertain the reliability of the test and to ensure that the test is able to perform its function. A poor test may be testing irrelevant content, or the manner of the writing may be uncritically biased towards a particular ability level for reasons that are not related to the test but to the quality of the writing. In such cases, the test instrument loses some of its usefulness. A non-test example of an inappropriate instrument would be a measuring jug made of paper used to measure the volume of boiled water. The test instrument, the paper jug, is an inadequate vehicle for its purported task.

CTT theorists have developed methods to overcome these barriers (Brown, 2005). The twin tools of Item Facility (IF) and Item Discrimination (ID) attempt to go beyond the nature of the total score and investigate more detailed relationships between the individual items on a test and the overall scores. IF and ID are relatively easy to understand even for those without a background in statistics. They can be obtained using spreadsheet software, such as Microsoft Excel, with only a minimum amount of preparation when all the raw data is collated. Split-half reliability analysis helps test writers understand the balance of a particular test, where the difficult items are found in a single test. The Rasch model is predicated on the individual at both the level of the test item and the test taker. Various software tools exist that allow detailed analysis of raw data according to the Rasch model. Winsteps (Linacre, 2014) was used here. Using the tables, diagnostic tools, graphs and other functionality available in Winsteps requires at least a solid command of basic statistics and measuring methodology. Its learning curve is steep.

This article attempts to answer the question: is the information provided by Rasch significantly more valuable than CTT given the time required for its study? In other words, is Rasch's payoff enough to justify the time spent? A case study is shown in which a test is subjected to CTT analysis and Rasch analysis. The types of information arising from each analysis are discussed, and the pragmatic decision about the use of CTT and Rasch is given.

CTT provides tools that analyze overall test scores and that aim to judge the whole test holistically. Item facility describes the easiness of any individual test item, item discrimination shows how well an item did in separating the high scorers from the low scorers, and split-half reliability expresses the degree to which subsets of items provide consistent ranking of person ability. Following Brown (2005, p. 66), to calculate item facility (IF) for each item, the total score obtained by each student is divided by the total number of students.

$$IF = \frac{Total\ Correct}{Total\ Number\ of\ Takers}$$

If all test takers got the item correct, IF = 1.0. Correspondingly, if all test takers were mistaken, IF = 0.00. This simple tool can highlight test items that were too difficult or too easy. In Excel, test data can be sorted by IF score. Then the relative number of easy-to-difficult items can be ascertained. Using this, the balance of item difficulty, or facility, can be understood.

Item Discrimination (ID) develops on IF (Brown, 2005, pp. 68-70). A percentage of the examinees is chosen, usually between 25% and 33%. The IF scores of those test takers who scored in the bottom 25% (or 33%) is subtracted from the IF scores of the top 25%.

$$ID = IF(Top\ 25\%) - IF\ (Bottom\ 25\%)$$

Top scorers in a testing group should score higher than low scorers. Test items that distinguish well between these two groups, i.e. when ID < .4 (Brown, 2005, p. 75, citing Ebel, 1979) are stable. If, however, ID < 0:0, lower scorers got the item right more often than higher scorers. When this happens, the item needs to be analyzed to see why this happened.

Split-half reliability (SH) provides an estimate on the overall test reliability (Brown, 2005; Hughes, 1989). Test reliability is a function of both halves of the test resulting in equal scores for each student. On a 100-item test, any individual student can be given two scores:[1] Score 1 comprising the total correct from the odd- numbered questions, and Score 2 comprising the score from the even-numbered questions. If the test is reliable, Score 1 should be similar to Score 2 (Hughes, 1989, p. 32). For example, Student 1 scores

[1] This SH method is the second Hughes (1989) describes and is more robust because his first method of generating score 1 from the first 50 items and score 2 from the latter 50 is problematic for tests whose questions get progressively more difficult deliberately.

38 on the odd-numbered items and 36 on the even-numbered items on a 100-item test. There is not so much discrepancy between these two halves. There is, however, an inconsistency in the results in the test when Student 2 scores 38 on the odd-numbered items and 16 on the even-numbered items.

Split-half reliability speaks more to overall test imbalance than to item or person analysis. Its use as a test of the test is vindicated in that it can point out imbalances in the test design. Brown (2005) also suggested the Cronbach alpha coefficient as another way to calculate reliability, but cautions that "conceptually, the split-half method is the easiest of the internal-consistency procedures to understand" (p. 179).

Winsteps (Linacre, 2014) offers a wide variety of functionality for many different levels of analysis. This section describes five key tools that offer the most immediate benefit to test developers and are the most accessible in that they do not require knowledge of advanced statistics. In a similar way, most users of a software tool such as Adobe Photoshop only use a small subset of that program's functionality. Georg Rasch wanted a method that understood the role and position of the individual within the frame- work of the construct under investigation.[2] "Individual-centred statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated." (Rasch, 1960, p. vii, cited in Choppin, 1983, p.12)

The Rasch model is described mathematically by an equation that predicts the expected score of any individual on any item on the test based on a matrix of item responses from all candidates. This necessitates a complex calculation, and Winsteps performs multiple iterations before settling on the model. The steps below summarize roughly what Winsteps does to calculate data-model fit. These are shown linearly to suit the medium of an article. In real terms, though, the program performs the steps many times, through a number of iterations until it has reached an acceptably low Maximum Score Residual (MSR).

1. Raw score data is read.
2. Each test taker's total score is tallied.
3. Each item's item difficulty level (or item measure) is calculated.
4. Items are placed on a scale of difficulty.
5. Each individual's item-by-item expected score is worked out (i.e. their overall score places them at a particular point on the scale, and this is judged against the difficulty of each item).
6. The score residual is calculated. This shows the difference between the total of the expected scores against the actual scores.
7. The model is refined through a process of updating the subsequent iteration using the information derived from the current one.
8. The iterations stop when the MSR reaches a pre-set level.

Any mathematical model must make assumptions. There is a critical difference in the assumptions underlying IF in CTT and how Rasch works out the difficulty of an item. We have seen earlier how IF is derived in CTT, as a function of the total correct answers divided by the total number of questions. No individual's overall score is factored in. Because of this, CTT requires a further step: the calculation of Item Discrimination (ID). With ID, the test analyst must decide on a percentage of high and low scorers. Using the IF for each group, ID can be established. ID figures are highly useful in discovering faulty test items, but even with IF and ID, it is pragmatically difficult to judge whether an individual test taker got a question right or wrong by luck or by guess.

[2] It is beyond the scope of this article to discuss the mathematical formulae that describe the Rasch model nor develop a discussion into the history and principles behind Rasch's statement. For further reading into the development of the Rasch model, see Bond and Fox (2007) chapters 2 and 3.

A critical difference between CTT and Rasch is that Rasch accounts for the fact that some test takers' responses on test items do not reflect their true ability. An examinee may guess correctly on an item that they have no idea about. Alternately, a high scoring test taker may slip on a relatively easy item and get the item wrong. The concept of the "expected score", then, is crucial to understanding how Rasch decides on the probability of a response type per examinee per item. The sum total of an examinee's correct responses shows that participant's general level. The difficulty of an item is measured against the level of the examinee, and the probability of that person getting that item correct can be understood. In other words, the item difficulties can be used to estimate the probability of any person answering any item correctly.

Two variables are involved, the test taker and the test questions. Let's look at each variable in turn using a test of 10 questions. The test taker can be in one of three states: they are too good for the test, in which case they will score 10/10; they are too bad for the test, scoring 0/10; or they are somewhere in the middle. Both Rasch and CTT effectively ignore those who are too good and those who are too bad. In CTT, test takers are given scores of 100% and 0% respectively. In a sense, CTT does ignore these scores as 0% and 100% are not meaningful beyond a purely ranking measure. Rasch labels these test takers as "extreme" and their data does not contribute to measurement. Another way of expressing these extreme cases is to say that the test does not adequately measure their level. More difficult test items are needed for the high ability examinee, and more simple items needed for the other. The information about the test given by CTT and Rasch analyses can only be effectively utilized when test-test taker scores fall within 1 to 9.

An examinee scored 9/10. They made a mistake on one test item, but overall their final score is as high as possible that is useful for analysis. We need to wonder about the item that was wrong. Did they not know the information in the question, or did they slip up? A test taker who scored 1/10, similarly, may have known the question genuinely or simply guessed. Yet a test taker who scored 9/10 is of a higher level than one who scored 1/10. There may be times in a test when a test taker guessed an answer correctly and other times when they slipped up on a question that is easier than their level. These responses are said to be "unexpected". In order to judge this, we must be able to analyze item facility (IF). The model may be summed up thus: the probability of any student getting any question correct is a result of the difference between item difficulty and person ability. A feature of the Rasch model include is a test taker's total correct score provides rank ordering of ability. In other words, a score of 9/10 indicates higher ability than 1/10, even if there are questions that were answered unexpectedly.

I have selected three tools to demonstrate some of the functionality of Rasch. I believe all three of them to be conceptually simple; they may all be understood without an advanced knowledge of either the underlying mathematical model and they produce values that appear ranked and may be understood as so without losing too much of the inherent subtlety. I have used these to show the power of Rasch quickly and successfully to colleagues far more mathematically challenged than myself.

Point-measure correlation is in some ways similar to ID in CTT. Point-measure correlation refers to the correlation between the difficulty of each individual item and the difficulty of the test as a whole. A value of 1.0 would indicate that all low ability test takers got the item wrong and all high ability test takers got it right, that is it indicates a perfect correlation between the item responses and the estimated Rasch measures of the test takers. A value of zero tells developers that there is no relationship between the particular item's responses and the rest of the test. In other words, whether students got it right or wrong is random. A negative value indicates a flawed test item as the lower scorers got that item correct more often than high scorers. These negative values are more problematic than zero values, and may indicate that the item is flawed in some fundamental way, and that it should be checked to see whether the answer key was wrong, revised, or possibly deleted from the test.

A subtlety that may be missed by novice Rasch users is its apparent ranking method. When classroom teachers see percentage scores, they may interpret them as representing equal intervals on a line from 0 to 100. Yet CTT does not attempt to show the interval between, say, 45% and 46%. Depending on the test, the interval between these two scores may well be virtually nothing, or it may be very wide. Rasch, on the other hand, aims to provide equal interval measures, so Rasch point-measure correlations are based on interval level measures, whereas CTT ID values are not. Teachers may miss this subtlety, but the concept of more difficult and easier items is not challenging.

Table 1 shows typical Winsteps item statistics. Reading from the left, we have the item number, the total score (the number of correct responses), the count of all responses, and the logit measure of item difficulty. No examinee could answer #10 accurately and the estimated measure of 101.87 is thus an extreme score. As mentioned earlier, there is a conceptual gap between these measures which look like percentage figures and the real workings of Rasch. Part of this apparent similarity can be explained by Holster and Lake (2014, p. 140) who suggested setting the mean item difficulty at 50.00 because "figures in the range of 50 to 100 are easier to understand", whereas according to them, researchers "usually set it to 0". But even if these measures are not percentage values, they generally fall within what looks like figures non-Rasch specialist classroom teachers are likely to comprehend. This table is ordered from the most difficult item first then successively adding the easier items. Other orderings are possible (for example, tables ordered by the closest match of the items' measures to the model). The measure values give a ready understandable account of the relative difficulty of each item. The total score figures rise as the measure value falls.

Table 1
Rasch Item Statistics

				Model	Infit		Outfit		Pt measure		Exact Match	
Item	Score	Count	Logits	S.E.	Mnsq	Zstd	Mnsq	Zstd	Corr.	Exp.	Obs%	Exp%
10	0	16	101.87	18.75		MAXIMUM	MEASURE		.00	.00	100	100
8	2	16	77.81	8.67	0.54	-0.8	0.25	-0.4	.68	.46	93.8	89.6
13	3	16	71.33	7.49	0.82	-0.3	0.63	-0.1	.58	.50	87.5	84.8
14	3	16	71.33	7.49	1.04	0.2	1.01	0.3	.46	.50	87.5	84.8
12	4	16	66.24	6.84	1.37	1.1	2.01	1.3	.27	.53	75.0	80.4
15	4	16	66.24	6.84	0.58	-1.3	0.36	-1.0	.76	.53	87.5	80.4
11	7	16	54.23	6.07	1.36	1.2	2.45	2.4	.27	.55	68.8	75.5
1	9	16	47.12	5.99	0.95	-0.1	1.15	0.4	.52	.52	81.3	74.2

Rasch variable maps, or Wright maps, such as shown in Figure 1, plot the test taker and item on a vertical scale according to the test taker's ability and the item's difficulty. High scoring test takers and difficult questions are at the top. Using the visual data, the test developer can a number of kinds of information. Because the data is visual, there is an immediacy to its interpretation that novice users and classroom teachers appreciate. Items that have no corresponding test takers at the top are too difficult and are not useful in segregating populations of higher ability test takers. A few items that are above the level of the examinee group are needed to ensure no ceiling effect. Those items at the bottom are too easy and offer no useful information about the level of the lowest ability test takers. Too many test takers lined up with a single question points to the lack of questions available to discriminate between those test takers. Too many questions for too few test takers indicate that there are too many questions at the same level, again an indication that the test items need to be analyzed for purpose.

Figure 1 uses the same data set as Table 1. Visually, it can be seen that Item 10 is right at the top of the map, and the same downward ordering of the questions' difficulty is observable on the right-hand side. Here we also have student data. As well as the measure of the question item, Rasch also computes a measure for each test taker. These values are positioned on the left-hand side of the map. S14 is the highest ability examinee and S01 the lowest.

```
MEASURE      Person - MAP - Question
                 <more>|<rare>
   90                  +  Q10
                  S14  |T
   80                 T+
                       |  Q8
                       |
   70             S12  +S Q13    Q14
                      S|  Q12    Q15
   60   S03  S06  S15  +
                  S10  |
             S02  S07 M|  Q11
   50                  +M
             S04  S13  |  Q1     Q4
   40             S09  +
        S05  S08  S11 S|  Q3
                       |  Q5     Q9
   30             S01  +S
                       |  Q6
   20                 T+
                       |  Q2
                       |T
   10                  +  Q7
                 <less>|<frequent>
```

Figure 1. Person-item map showing student ability on the left and item difficulty on the right.

The table of distractor frequencies, shown in Table 2, is arguably the single most useful tool which can be interpreted without too much conceptual difficulty for the novice Rasch user. This table shows the number of test takers that selected each particular option for every question. Also, the average ability of test takers for each option is shown. Together, these provide a highly useful tool for the refinement of a test vehicle. Misfitting items are marked with an asterisk. These are items where the correct option was selected more by lower ability level (on average) than not. I use 4-option multiple choice items as an example in Table 2. In the first part of the table, Winsteps shows the item number, the options (here A = 1, B = 2 and so on), and a score value, which is the correct answer. The 1 is always at the bottom of the set. Next to these values next is the data count in both raw figures and percentages of the total test takers. Item 10 was answered correctly by no examinee. Option A (i.e. data code 1) was selected by students whose measured ability averaged 55.04. Option B by students at 47.34, and Option C by examinees at 51.23. The spread of the selection is reasonable. No single distractor monopolized the selection. This can be contrasted by looking at Option B in item 8. Only one examinee chose that and their measured ability level was low.

Item 14 highlights a potential problem in the test. The ability of those examinees who answered correctly as 65.62. Yet a higher ability test taker (at 69.73) chose another answer. The asterisk provides an immediate clue to this problem. In this case, only one higher level test taker made an error, and it is likely that this was simply a slip. But, if there were many higher ability examinees choosing the wrong answer, that is a serious indication that the item needs to be investigated.

Table 2
Distractor Frequencies

ITEM NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN	OUTFIT MNSQ	PTMA CORR.
Q10	2	0	4	25	47.34	5.91		-.16
	3	0	7	44	51.23	5.36		-.01
	1	0	5	31	55.04	9.28		.16
Q8	2	0	1	6	36.46		0.10	-.26
	1	0	7	44	48.08	4.35	0.60	-.20
	4	0	6	38	49.09	5.27	0.70	-.12
	3	1	2	13	77.79	8.06	0.20	.68
Q13	2	0	3	19	43.68	7.23	0.50	-.25
	3	0	4	25	47.33	6.68	0.70	-.16
	1	0	6	38	49.09	5.27	0.80	-.12
	4	1	3	19	69.41	9.59	0.60	.58
Q14	4	0	8	50	43.92	3.56	0.50	-.51
	1	0	4	25	51.29	7.26	1.00	-.01
	2	0	1	6	69.73		3.60	.32
	3	1	3	19	65.62*	11.19	1.10	.46

Method

Participants, Materials, and Procedure

Case study data are taken from a test written to supplement the author's English language textbook *Nursing Care* (Smiley & Masui, 2013). This textbook is designed for students on a nursing course at the university level studying English as a part of their curriculum. The prior English language level assumed at the start of the course is roughly between Grade 3 and Pre-Grade 2 *Eiken*. The *Monkagakusho,* the Japanese Ministry of Education, states that the target finishing level of high-school pupils should be *Eiken* Pre-Grade 2 (MEXT 2013), so this book is considered suitable for the university English course. The test, comprising 50 multiple-choice items, assesses Units 1 to 6 of the book, and there is a further Test B for units 7 to 12. This case study looks only at the first test. They are considered criterion referenced tests (CRT) (Hughes, 1989) as students have finished the relevant units before taking each test. However, there is a degree of norm-referenced type material present. Students at university exhibit a large range in English proficiency, so a published textbook for this level contains material many students have not yet mastered. Ideally, a CRT only assesses elements that were new to students at the start of the course, but in this case because many students did not have a Grade 3 ability prior to the start of the course, a significant amount of the erstwhile assumed language and the technical nursing content were new.

Results

CTT Results

As shown in Table 3, the test produced an average score in the 50% to 60% range. CRTs may be expected to return higher scores if the language is known prior to the start of the course and the test vehicle assesses only the new content. As mentioned above, however, there is a sizeable number of students who have not attained a proficiency level of *Eiken* Pre-2nd Grade. Their task throughout the course will be to simultaneously learn the test content and develop their basic language proficiency. With this taken into consideration, an average of 52.2% may be considered acceptable.

Table 4 shows those items that have the top five and the bottom five IF scores. Items 24 and 15 are above .85 which indicates that they are easy. Item 37 was only answered correctly by 8% of test takers. This item needs to be investigated. Items 29 to 26 all return a score under .2, and they may also be too difficult. IF shows the test developer that there are certainly three items that require thought and perhaps alteration and five or six others that need further analysis before their place in the test is assured.

Table 3
Nursing Care Test Summary Statistics

	mean	SD	Max	Min
Score (Max.50)	26.1	6.8	41	12

Table 4
Item Facility Values

	Top 5					Bottom 5				
Item number	24	15	18	40	9	26	47	44	29	37
Item facility	.87	.86	.8	.8	.79	.19	.17	.14	.1	.08

Theoretically, every student has access to all the information that will be in a criterion referenced test during the course duration. Learning objectives are specified prior to the teaching term and learning actions are chosen to allow for maximal retention of those objectives. A textbook is a set of learning objectives that contain learning actions. Therefore a test that is wholly based on a textbook must be defined as a criterion referenced test. Higher IF scores may be expected than from a norm-referenced test where the items may be drawn from language or content elements examinees have yet to encounter. Conversely, very low IF scores may be indicative of a number of serious issues in the class: students may be unmotivated to learn the material in the textbook, the assumed starting level of the student body may be too high, the class content may have focused on segments of the book that were not targeted in the test. The test items themselves may be too obscure, in that they test too narrow areas of the book, topics or language that appears only once.

Question 29 highlights another test writing difficulty. Only 10% of takers got this item correct. The question's distractors A, B, and C, are all possible answers. The correct answer (D) reads, "All of the above". Examinees were not accustomed to this question type, and it only comes once in the test, so they could not train themselves to expect this type. Upon investigation, Question 37 throws up another issue in test validity. Many tests use the form:

Q37: It is important to be __________ to new patients.
a) helpful b) helping c) helped d) helper

This kind of item seems intuitively useful to many teachers. All verb forms and the noun form "helper" may be in the category of assumed knowledge. Yet, something inhibited examinees from answering correctly. Anecdotally, because the subject matter is sensitive in our institution, I can report a heightened discussion over the use of this type of item when a post-test study revealed that a similar IF score was returned in our entrance exam. Perhaps the control examinees have over verb conjugation is not strong enough to merit a test item that focusses only on that. Discrete point testing may be less valid as a measure of holistic ability than is believed at my institution. Question 44, similarly, offers a counter-intuitive response, this time on the discrete testing of a noun item.

Q44: Aerobics is a good way of keeping __________.
a) exercise b) fit c) health d) lifestyle

IF is seriously limited in its ability to show the distractor selection ratio. That 14% of test takers chose B is known. The ratio chosen for the others is necessary before any assessment can be made.

ID values are summarized in Table 5. Brown (2005) provided guidelines on item discrimination as to which items do a good job in discriminating between the high and the low scores. An ID score of .40 and above indicates a solid item. Scores of between .39 and .30 are considered good. Items whose scores are between .29 and .20 need some alteration. This change depends on whether the item should be made more difficult or easier. The judgement for this action is based on the numbers of test takers scoring accurately in each high or low group. The ID score itself does not give information directly; the writer needs to look at the precise details of the responses for that item. Scores below 0.20 do little to differentiate between the higher and the lower groups. In this test, one item (Question 27) had a negative correlation score. This means that the lower group students scored more highly than the higher group. This item needs to be changed.

> *Q27: Why did Sara stand on some scales?*
> *a) to let the nurse measure her weight*
> *b) to let the nurse measure her body height*
> *c) to measure her weight*
> *d) to measure her body height*

Table 5
Item Discrimination

Discrimination	>.40	>.30	>.20	<.20	<.00
No. of Items	24	8	6	12	1

This question is one of three that follow a short paragraph-length reading. Even without the accompanying text, proficient users of English will be able to eliminate distractors B and D. The answer comes down to the distinction between the passive "having her weight measured" or the active "measuring her (own) weight". The text reads ". . . and the nurse measured her weight". Is this distinction too fine to be useful at this level, or is there something about the distractors that added some complexity to the question. Again, CTT does not offer ready tools to find this out.[3]

Split-half reliability is shown in Table 6. Items were split in two ways: between the first half of the test, assessing listening, and the second half, assessing reading, and between odd-numbered and even-numbered items. Both analyses returned a correlation coefficient of .68, indicating modest reliability. The average scores show that the listening section was statistically significantly easier than the reading section. A paired-sample *t*-test returned values of $t = 14:45$, $df = 142$, $p < .01$. The odd-even split half figures show a slightly less extreme imbalance, and the total scores are reversed.

Table 6
Split-half Reliability

	Mean Score	Correlation
Items 1-25	59.7%	.68
Items 26-50	42.1%	
Odd items	48.1%	.68
Even items	56.1%	

At many points in the analysis, using CTT tools generated more questions than answers. IF information did highlight those areas of ease and difficulty, but without ready access to the distractor selection ratios,

[3] In my pre-Rasch Excel days, I often generated this information in Excel using COUNTIF(cellrange=1), COUNTIF(cellrange=2), and so on. But manually preparing these sheets was time-consuming.

further analysis must necessarily be limited. ID is a useful tool to check if the test items inadvertently contain biases towards lower ability level test takers. Those items that fail the ID test can be analyzed further, but the same limitation applies here as to IF. Split-half reliability talks about the test as a whole, so offers very little to help the writer revise the test.

Rasch Results

Winsteps' summary statistics provide the same basic figures as can be output by Excel; the mean, Standard Deviation, Maximum and Minimum raw scores. Winsteps' Rasch summary statistics, shown in Table 7 and Table 8, provide information about both persons and items, including the logit measures. Winsteps models the persons and items as it works out the precise relationship between these, but models do not return a perfect match with real world data, so the summary statistics indicate the degree to which the data fit the model. A novice user will not know the acceptable range of values for infit and outfit. Taking Holster and Lake (2014) as a guide, the person infit and standard deviation are acceptable at 0.99 and 0.15 respectively. The corresponding outfit values also seem to be acceptable. The item infit and outfit values are similar to that of the person's, suggesting that the model is a satisfactory match to the data. One reviewer pointed out that values +/-0.30 are acceptable revealing that the maximum item infit of 1.25 is good, but the cut-off points of 1.30 and 0.70 are reached in the maximum item infit and outfit, where 1.30 can 1.38 can be seen. These values are the result of the data not matching the model, i.e. when a lower ability student got a difficult item correct. Being summary statistics, the information speaks to the test as a whole. Also shown are the Rasch reliability of separation estimates for the test and Cronbach's alpha, analogous to the split-half reliability shown in Table 6. The Rasch person reliability and Cronbach's alpha are considerably higher than the split-half reliability because they are calculated from the entire 50 items, rather than the 25 items used to calculate split-half reliability.

Table 7
Summary Statistics for Persons

	Total			Model	Infit		Outfit	
	Score	Count	Measure	Error	Mnsq	Zstd	Mnsq	Zstd
Mean	26.1	49.8	50.94	3.30	0.99	0.0	1.02	0.1
S.D.	6.8	0.8	7.14	0.16	0.15	1.0	0.24	1.1
Max.	41.0	50.0	68.69	4.07	1.43	3.2	1.70	3.2
Min.	12.0	43.0	35.63	3.17	0.69	-2.7	0.53	-2.4

```
Real Rmse    3.39 True Sd    6.29   Separation  1.85   Person Reliability  .77
Model Rmse   3.31 True Sd    6.33   Separation  1.92   Person Reliability  .79
S.E. of Person Mean = 0.60
```

```
Notes: 143 persons,  50 items
Person raw score-to-measure correlation = 1.00
Cronbach alpha (KR-20) = .79
```

Table 8
Summary Statistics for Items

	Total			Model	Infit		Outfit	
	Score	Count	Measure	Error	Mnsq	Zstd	Mnsq	Zstd
Mean	74.5	142.5	50.00	1.99	1.00	-0.1	1.02	0.0
S.D.	30.2	0.7	11.11	0.28	0.09	1.3	0.16	1.4
Max.	125.0	143.0	76.79	3.08	1.30	4.1	1.38	3.8
Min.	12.0	141.0	29.75	1.77	0.83	-2.5	0.72	-2.3

```
Real Rmse   2.05 True Sd   10.92  Separation  5.33  Item   Reliability  .97
Model Rmse  2.01 True Sd   10.92  Separation  5.43  Item   Reliability  .97
S.E. Of Item Mean = 1.59
```

```
Notes: 50 items, 143 persons
```

Figure 2 shows the Wright map comparing persons and items. No extreme items or persons were present in this test. Questions 37and 29 are the most difficult with scaled scores of about 75. The summary statistics tell us that the max person was 68.69, and this can be seen on the map. This is analogous to the IF information delivered earlier, and a similar investigation into the possible causes of the item's difficulty may be conducted. Generating the variable map took two mouse clicks. The same cannot be said for creating the IF table. IF informs about the relative numbers of test takers getting the item correct, and the variable map gives an indication of the distance between the upper (and lower) reaches of the items and the examinees. Having a test taker overall range outside that of the items would be highly suggestive of a test that did not accommodate all of the examinees' ability levels. These two tools offer similar information, and together their power contributes more to an understanding of the test.

```
MEASURE             Person + Item
                    <more> | <rare>
   77                      +  Q37
   75                      +  Q29
   72                      +T
   70                    1 +  Q44
   67                    0 +  Q26 Q47
   65                  000 T+  Q21 Q39
   63          00000000011 +  Q33
   60                00001  +S Q31 Q35 Q46
   58          00000000011 S+  Q20 Q27 Q42
   55   00000000000000111  +
   53    0000000000111111  +  Q23 Q25 Q32 Q36 Q41
   51 0000000000000011111 M+M Q14 Q16 Q49 Q5
   48       0000000000011  +  Q1 Q12 Q30 Q4  Q8
   46      00000000111111  +  Q10 Q17 Q2  Q3  Q48 Q7
   43     000000011111111 S+  Q28 Q45
   41            00000011  +  Q19 Q22 Q34 Q38 Q43
   39                0011  +S Q11 Q50 Q6
   36               01111 T+  Q13 Q18 Q9
   34                      +  Q40
   31                      +  Q15
   29                      +  Q24
                    <less> | <frequent>
```

Figure 2.
Variable map showing the distribution of persons and items.

Figure 2 shows the expected bell-curve like histogram for both items and persons. A classroom teacher may feel satisfied with this distribution. However this fails to appreciate a main purpose of a well-designed test which is to discriminate between different ability levels of test taker, so a flatter distribution of item difficulty would suggest a better discriminatory instrument than a bell curve. With all histograms, the bucket size has an important bearing on its shape. For example, Figure 3 shows a zoomed-in view of the gap around the 55 level. Using this information, analyzing questions 27 and 42 against questions 23 and 25 may allow for more precisely targeted questions around those levels to be developed.

```
MEASURE                                         Person + Item
                                                <more> | <rare>
  57                           12  27  56  89  91 110  +  Q27  Q42
  56                                   25  64  92 100  +
  55  1  5  6  19  32  46  49  57  60  88  96 116 128  +
  54                2  36  37  74  86  95 107 120 138  +  Q23  Q25
                                                <less> | <frequent>
```

Figure 3.
Magnified variable map showing items and persons between 54 and 57 scaled points.

Why is question 23 easier than question 42? Perhaps this is impossible to answer definitively, but the process of trying is valuable.

> *There is an underlined word in each sentence. Choose the best meaning from the options.*
>
> *Q23: The doctor was worried about John's <u>diet</u>.*
> *a) John is trying to lose weight*
> *b) what John eats on special days*
> *c) what John eats usually*
> *d) John wants to become smaller*
>
> *Q42: General hospitals have many departments __________ are very big.*
> *a) and   b) too   c) though   d) even*

The textbook glosses the term *diet* in Japanese, and students who remember that definition are likely to select option C. Conversely, there are no direct grammar directions in the textbook, and students have no practice of conjunctions or non-repetition of the subject after a conjunction when there is no comma. Q42 may be challenging from the perspective of a Japanese learner of English through L1 interference as subjects are typically not be repeated. Japanese is a theme-rheme structured language, and syntax such as *Hospitals have many departments too very big* is acceptable. In this interpretation, the emphasising function of *even* may be placed directly after the *departments* to provide the rheme comment on the *hospital*. Or the grammatical potential for complexity may be immaterial if the difficulty is due to the focus on the discrete item which is either known or unknown.

Items 35 and 50 also show an interesting result. Both questions test knowledge of discrete vocabulary items. *Annual* and *updates* are glossed in the textbook and are recycled throughout the unit in which they appear. Both sentences are in the active voice and contain nothing out-of-the ordinary in terms of object and adverbial clauses. Intuitively, I would have estimated *annual* to be the more challenging term especially as *update* is a commonly used word in Japanese that has a very similar semantic scope to the English. Very little separates them in terms of perceived difficulty, yet Item 50 is measured at 39 and Item 35 at 60.

> *Q50: John came to the clinic for his __________ health check up.*
> *a) by year   b) year   c) annually   d) annual*
>
> *Q35: Nurses give patients' families __________ on their health.*
> *a) new   b) updates   c) tests   d) conditions*

The five items with the poorest point-measure correlations are shown in Table 9. One item, Item 27 has a negative value. This is the same item that was discovered by ID and discussed above. The rest are under .10. In the whole table, only 15 items have a value of .40 or above, the cut-off figure Holster and Lake (2014) recommended as showing that an item is functioning well. These point-measure values do not indicate that the test as a whole is performing well as an instrument that differentiates between different ability levels of students. The CTT ID values pointed to 12 questionable items, but Rasch highlighted 35 items that require attention.

So far, the tools have foregrounded items that deserve further investigation. At each juncture, the information regarding the ratio of selection of the distractors was missing. As a result, the test developer can focus the attention on the where but not precisely on the how. The distractor frequencies, shown in Table 10 fill in this missing piece. Winsteps has ordered this table according to the degree to which the model predicted the responses. Those items that functioned less well are at the top as can be seen with the outfit mean-square value in the third rightmost column. To a test designer, however, there are two other columns that hold very valuable information. The data counts and the average ability columns show how many test takers selected each question option and what the overall level of those test takers is. These figures provide a means by which the developer can see exactly how well the test items discriminated the various levels of test taker. An asterisk next to a value indicates that the average level of test taker getting the item correct is lower than the average of another option chosen. Ideally, high scorers select the correct option and lower scorers select the other options. This happened nine times in this test.

Table 9
Point-measure Correlations for the Five Poorest Performing Items

	Total	Total		Model	Infit		Outfit		Point-measure	
Item	Score	Count	Measure	S.E.	Mnsq	Zstd	Mnsq	Zstd	Corr.	Exp.
27	51	142	57.39	1.84	1.30	4.1	1.37	3.8	-.09	.31
33	36	143	63.01	2.01	1.16	1.6	1.38	2.5	.02	.29
26	28	143	66.52	2.19	1.11	0.9	1.36	1.9	.05	.26
44	20	143	70.84	2.48	1.11	0.7	1.28	1.2	.05	.23
37	12	142	76.79	3.08	1.07	0.4	1.24	0.8	.06	.19
47	25	143	68.01	2.28	1.12	0.9	1.27	1.4	.06	.25

Table 10
Distractor Option Frequencies for the Five Poorest Performing Items

			Data		Average	S.E.	Outfit	Point-M
Item	Code	Score	Count	Percent	Ability	Mean	MnSq	Corr.
33 A	1	0	1	1%	38.08	N.A.	0.2	-.15
	3	0	1	1%	46.63	N.A.	0.6	-.05
	2	0	105	73%	51.00	0.67	1.1	.01
	4	1	36	25%	51.24	1.33	1.5	.02
27 B	2	0	6	4%	45.37	2.69	0.6	-.16
	3	0	82	58%	51.76	0.82	1.3	.13
	4	0	3	2%	55.20	2.30	1.5	.09
	1	1	51	36%	50.03*	0.94	1.4	-.09
	MISSING	***	1	1%	50.37	N.A.		-.01
26 C	1	0	3	2%	43.55	1.05	0.4	-.15
	4	0	14	10%	43.72	1.74	0.5	-.33
	2	0	98	69%	51.99	0.64	1.2	.22
	3	1	28	20%	51.66*	1.56	1.4	.05
44 D	1	0	7	5%	44.87	3.38	0.7	-.19
	4	0	12	8%	46.94	1.51	0.6	-.17
	3	0	104	73%	51.63	0.70	1.2	.16
	2	1	20	14%	51.84	1.43	1.3	.05
37 G	4	0	5	4%	47.50	3.72	-0.7	.09
	3	0	27	19%	49.32	1.38	0.9	-.11
	2	0	98	69%	51.38	0.73	1.1	.10
	1	1	12	8%	52.25	1.72	1.3	.06
	MISSING	***	1	1%	52.76	0.02		

Looking at Question 27 again:

> *Q27: Why did Sara stand on some scales?*
> *a) to let the nurse measure her weight*
> *b) to let the nurse measure her body height*
> *c) to measure her weight*
> *d) to measure her body height*

The correct response of #1 was chosen 36% of the time by students who averaged 50.03. Distractor #3 was chosen by 58% of the examinees whose average ability on the test was 51.76. The absolute difference between the levels is only 1.03, so perhaps these students can be judged at a roughly similar level. Distractor #4 was chosen by students of level 55.20, but as the number of students was only three, the possibility that these three students simply slipped up on that item seems likely. Option #2 was selected by 58% of examinees, or 22% more than those who answered correctly. Their average ability was 1.73 points higher. Again, more high ability level examinees answered wrongly. There is very little difference in the wording of both options, the question targets a vocabulary item or phrase. One solution springs to mind. In Japan, some scales have the dual purpose of weighing the body and measuring the height. It is possible that cultural knowledge interfered with examinees ability to separate the meanings in options A and B. These are nursing students under discussion, and even though the non-specialist view of *scales* may be similar in Japanese and English, there remains the possibility that the specialist understanding is different. This can be readily checked by questioning a native Japanese speaker about the semantic space for *scales*. If a discrepancy does exist, future editions of the textbook may need to incorporate it as a teaching point.

Looking at Item 37 again:

> *Q37: It is important to be __________ to new patients.*
> *a) helpful   b) helping   c) helped   d) helper*

On this item, higher ability examinees answered correctly, but the degree of discrimination between those and the others is very narrow, measured at 0.87. The existence of difficult questions in a test does not detract from its validity, but such a fine line is perhaps troubling. The same pattern can be observed for Item 44:

> *Q44: Aerobics is a good way of keeping ___________.*
> *a) exercise   b) fit   c) health   d) lifestyle*

The term *fit* is explained in the textbook, and the adjective-noun distinction *healthy- health* is practiced in the workbook. Options #1 and #4 were dismissed by test takers. These options need to be reworked to allow for a better spread of answer-distractor options. Few examinees selecting an option indicates that that option is not working usefully towards any target the question may have. More usefully, the usages of *healthy* and *fit* may become a teaching point in an updated revision of the textbook.

The absolute measured difference between the 73% of examinees who selected the wrong option and those 14% who answered correctly was 0.21 scaled points. This lack of clear discrimination between levels brings the quality of the test into question. The differentiation between distractors needs to be clearly demarcated, especially that between the correct responses and the others. In this test, most of the items are separated by only a few percentage points.

Conclusions

Both CTT and Rasch indicated some weak items in the test. In the Rasch analysis, Item 27 produced a negative correlation in ID and PT measure values. CTT's IF and ID values identified a number of items that produced questionable figures. None of the IF figures, though, were sufficient to uncategorically eliminate any item. IF provided a clue as to where the problem items were. One by one, an analysis of each item was necessary. ID suggested that there were 12 weakly discriminating items. Winstep's point-measure correlations pointed to over 30 items.

I can prepare the worksheets for IF, ID and split half for a data set of 140+ examinees on 50 questions in about an hour if my template files are available. From scratch, the process would take upwards of two hours. The same amount of data can be used to set up a Winsteps analysis in a few minutes. From then, each analysis requires only a two mouse clicks. As a classroom teacher, the amount of time saved by using Winsteps is considerable. As a materials developer, the readily digestible information is invaluable. However, CTT is conceptually straightforward, while Rasch is not.

In summary, CTT's IF and ID are a good place to begin the analysis of the test. They can indicate potential problems. The key word here is "potential". The analysis needs to go back to the raw data in Excel and hunt for more detailed information. Oftentimes, the trail goes cold as, for example, to discover the exact ID relationships that go beyond the top and the bottom 25% are simply not there. At other times, the search leads back to the original test paper for a study of the actual language in the paper. This is not a bad action, of course, and in all analyses need to end up with the test paper in hand. However, the better the quality of the numerical data, the less the analyst needs to be concerned with items that are not problematic, and the more they can focus on the real issues in the test. In this paper, I have only scratched the surface of what Rasch can do. Its true power lies far outside my current reach. My lack of experience will be clear to specialists reading this; they will have constantly scratched their heads wondering "why didn't he write about this or that?" However, I hope that they may reflect on the gap between their expert position and my own and come forward to help make Rasch more accessible to many who are presently unaware of its might. Rasch provides highly detailed and compelling tools for the analyst. The learning curve, though, is steep.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum.

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

Choppin, B. H. (1983). The Rasch model for item analysis. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Holster, T. A., & Lake, J. W. (2014). How high can they jump: An introduction to Rasch measurement. *文藝と思想 (Bungei to Shisou: The Bulletin of Fukuoka Women's University International College of Arts and Sciences), 78*, 19-45.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Linacre, J. M. (2014). Winsteps (Version 3.81.0). Retrieved from http://www.winsteps.com

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

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

Sick, J. (2008b). Rasch measurement in language education: Part 2. *Shiken: JALT Testing and Evaluation SIG Newletter, 12*(2), 26-31.

Sick, J. (2009). Rasch measurement in language education: Part 3. *Shiken: JALT Testing and Evaluation SIG Newletter, 13*(1), 4-10.

Sick, J. (2010). Rasch measurement in language education Part 5: Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing and Evaluation SIG Newletter, 14*(2), 23-29.

Smiley, J., & Masui, M. (2013). *Nursing Care*. Brighton: Perceptia Press.