From raw scores to Rasch in the classroom

Trevor A. Holster[1] and J. W. Lake[2]
trevholster@gmail.com
1. Fukuoka University
2. Fukuoka Jogakuin University

Abstract

Smiley's experience reported in this issue of *Shiken* is probably quite typical of moving from traditional analysis to Rasch analysis. Traditional analysis, exemplified by Brown's (2005) *Testing in Language Programs*, provides statistics such as item facility values (IF) and item discrimination (ID) which will identify most of the same problematic items as Rasch analysis, and it's unlikely that classroom grades would change to any substantive degree between the two for a thoughtfully developed test. Rasch analysis provides benefits beyond analogues of traditional item analysis, however, and this paper argues that two important practical benefits are the variable map, or Wright map, which provides a quick visual summary comparing students with instructional features, and data-model fit statistics which provide for diagnosis and identification of students requiring remedial instruction. This study illustrates the potential of these for curriculum planning and classroom diagnosis through analysis of the vocabulary section of an academic English placement test.

Keywords: Diagnostic assessment, Rasch, item analysis, vocabulary testing

As Smiley reports in this issue of *Shiken*, traditional item analysis includes item facility values (IF), which rank item difficulty by the proportion of correct responses, and item discrimination (ID), which shows whether high ability persons scored higher overall on an item than low ability persons, a simple assumption being that higher ID is generally better. Rasch software reports several statistics regarding item performance, including point-measure correlation and infit and outfit statistics. As Linacre (2012) explains, the point-measure correlation is closely related to the point-biserial correlation that can be used for the same purpose as ID (Brown, 2005, p. 70), so the closest analogue of ID is the point-measure correlation. Rasch fit statistics are based on a different conception of discrimination, however, and this is fundamental to understanding the differences between the Rasch model and traditional analysis. In traditional ID analysis, we assume that higher scoring students answer correctly more often on all items than lower scoring students, so ID allows us to identify items that behaved unexpectedly. We need some difficult items, i.e. with low IF values, to target high ability students, and these will have high IDs. We also need some easy items, i.e. with high IF values, to target low ability students, and these will have much lower IDs or correlations because many low ability students will answer them correctly. Although negative ID values indicate problematic items, a good test will have items with a range of IF and ID values, so higher ID alone does not automatically indicate a better item. The Rasch model shares the expectation that high ability students will succeed more than low ability students on all items and that point-measure correlations will vary for effective items but should always be positive, but Rasch data-model fit is calculated by comparing the observed discrimination of items, which are never equal, with a theoretical ideal in which all items have equal discrimination (see Sick, 2010, for discussion of Rasch model assumptions). However, Rasch discrimination is very different from the traditional ID value, so traditional analysis has no direct analogue to Rasch fit statistics.

The left hand panel of Figure 1 illustrates this key feature of the Rasch model, showing item characteristic curves (ICCs) for three items of different difficulty. The vertical axis shows the probability of success of a person on an item, ranging from a lower limit of 0.00 to an upper limit of 1.00. The horizontal axis shows person ability in log-odds units, or "logits" (Bond & Fox, 2007). When item difficulty and person ability are perfectly matched, the person has a 50% chance of success, giving odds of 50/50, or 1/1. The natural logarithm of 1/1 equals 0, so an expectation of success of 50% means a difference between item difficulty and person ability of 0.00 logits. In Rasch analysis, there is no absolute zero point indicating

zero ability, so 0.00 logits is just an arbitrary point that, by convention, indicates the mean difficulty of the sample of items. In Figure 1, therefore, we would expect 50% of students with ability of 0.00 logits to succeed on an item of average difficulty and 50% to fail. If the same group of students took an easier item, with difficulty of -1.00 logit, we would expect about 73% to pass and about 27% to fail, i.e. odds of about 73/27, because person ability is about 1 logit higher than the item difficulty and the natural logarithm of 73/27 is roughly 1. For a more difficult item of 1 logit difficulty, the probability of success falls to about 27% because odds of 27/73 corresponds to a logit difference of about -1.

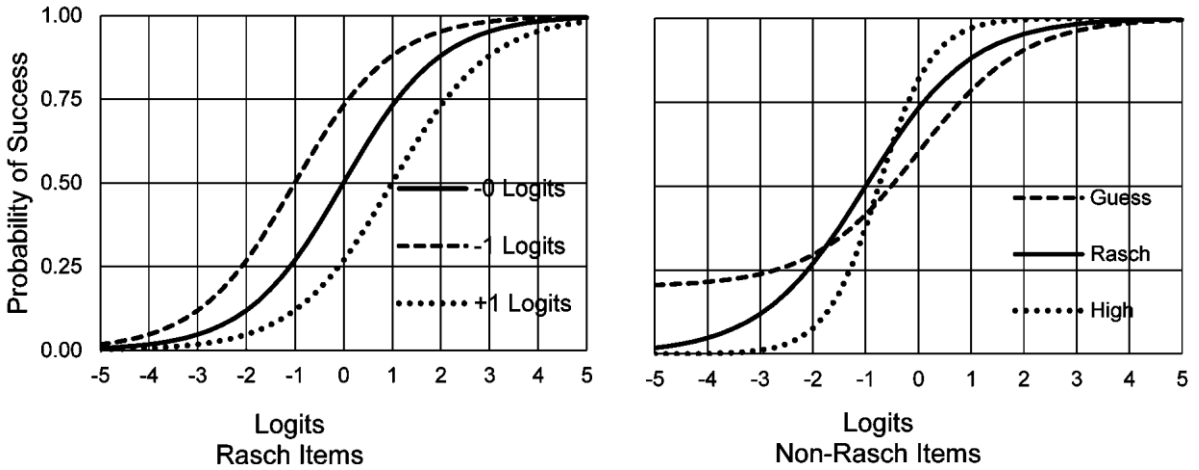

Figure 1. Rasch and non-Rasch item characteristic curves for items of different difficulty levels. The vertical axis shows the probability for three items of different difficulty. The Rasch model assumes parallel curves, but non-Rasch models allow non-parallel curves.

So far this is consistent with the commonsensical expectation that success on items will correlate with person ability, but what is conceptually important about the Rasch ICCs in Figure 1 is that they are parallel, i.e. that the slope of each curve is the same at each point on the vertical axis. The difference in difficulty between the successive items is 1.00 logits at every probability level. In other words, the relative difficulty of the items is theorized to be invariant regardless of the ability of the persons taking the test. Similarly, the relative ability of the persons is theorized to be invariant regardless of the set of items used in the test. The Rasch model thus assumes a stable hierarchy of person ability that does not vary for different samples of items, and a stable hierarchy of item difficulty that does not vary for different samples of persons. This theoretical ideal is only possible if ICCs are parallel (Engelhard, 2013), and, as item discrimination in the Rasch model is simply the slope of the ICC at the 50% expectation of success level, the ideal of invariant measurement is only possible if all items have identical discrimination.

However, real data sets never perfectly fit the idealized Rasch model, in fact, they often misfit quite dramatically. The right-hand panel of Figure 1 shows response patterns that illustrate items that misfit Rasch assumptions. One item, *Rasch* follows a Rasch ICC, but another item, *High*, has a much steeper slope, i.e. higher discrimination. The problem this causes is that low ability students have a higher probability of success on *Rasch* than on *High*, i.e. *Rasch* is easier than *High*, but for high-ability persons, the hierarchy is reversed and *Rasch* is more difficult. This is another way of saying that students seem to have followed different trajectories of acquisition for these two items. In the case of a classroom test where test items are based on course content, if a large number of items misfitted in this way, we might want to investigate to see whether our curriculum is mixing different types of knowledge and skills that should be assessed separately. Another source of misfit is shown by the item *Guess*. In this case, the

expectation of success does not approach the lower limit of zero assumed by the Rasch model, so even very low ability persons still have about a 20% chance of success. This is the type of pattern we might see in a situation where guessing is possible, such as a multiple-choice test with five answer choices or in a constructed response question that gives clues to the answer. An important point about these three items is that the major problem for the Rasch model is mixing items with ICCs that diverge too much from parallel trajectories, so items that function well in one test may misfit if used in a test that measures a different type of knowledge.

Rasch analysis provides a simple diagnostic tool to identify items or persons that violate the model's assumptions in the form of mean-square fit statistics. Fit statistics are generated from the patterns of the differences between observed responses and statistically expected responses, known as score residuals. In a dichotomously scored test, observed responses can only have values of 0 or 1, while expected responses, or probability of success on items, can take any value between the asymptotes of 0.00 and 1.00, so observed values and expected values can never be exactly equal. When person ability and item difficulty are perfectly matched, the probability of success equals .50, so the residual is 0.50 for a correct answer and -0.50 for an incorrect answer. Small residuals will occur when high ability persons succeed on easy items or low ability persons fail on difficult items, while large residuals will occur when low ability persons succeed on difficult items or high ability persons fail on easy items. Across the entire data set, these values are expected to follow a chi-square distribution, and the mean-square fit statistics provide a confirmatory analysis to see whether the observed data fit the modelled distribution.

The mean-square statistic has an expected value of 1.00, indicating patterns of responses that perfectly match the Rasch model, with a lower limit of zero and no upper limit. Mean-square values below 1.00 indicate responses that are more predictable than expected, called overfit, while mean-square values greater than 1.00 indicate less predictable responses, called misfit (or underfit). For the three items in the right hand panel of Figure 1, *Rasch* would show perfect data-model fit, but *High* would overfit the model and *Guess* would misfit the model. In the real world, some items and persons will inevitably be more consistent than average and some will be less consistent, so aiming for perfect data-model fit is not the objective. Rather, we need to investigate whether the misfit is severe enough to threaten the interpretations we wish to make of the test scores and whether there are systematic patterns of misfit that indicate sampling problems with either items or persons.

While much of the published research on language testing is from the perspective of large-scale standardized proficiency tests, where practicality and reliability are paramount concerns, a materials writer or textbook planner who wishes to integrate assessments into a course of study may be more concerned with criterion referencing student ability against instructional content or in diagnosing students or instructional items that follow unusual developmental trajectories. Rasch analysis provides useful tools for this, so the purpose of this paper is to demonstrate its benefits within instructional programs. This study was conducted as part of a curriculum development project for reading classes in a newly established Academic English Program (AEP) at a Japanese public woman's university. In 2011, the first year of the program, detailed goals and objectives were not available and different teachers used different reading textbooks and instructional approaches. Students' TOEFL score trajectories diverged from the assumptions of the university and prefecture, so textbook selection was reviewed and a placement test developed for 2012, with intended secondary uses as a diagnostic and achievement test. The 2012 test form had three sections of 50 items each and this was revised in 2013 to five sections of 40 items each, including content derived from the assigned textbook series, *Reading Explorer* (Douglas & McIntyre, 2009).

Each of the five different levels in the *Reading Explorer* series comprised 24 reading passages followed by five comprehension check questions intended to prepare students for tests such as the TOEFL. Each

reading passage targeted 10 academic words for explicit instruction, but it was apparent that many students needed to study non-academic words in the textbook as well, so supplementary vocabulary instruction was required and a vocabulary test was needed to determine appropriate vocabulary for students of different proficiency. Given the TOEFL orientation of the program, Davies and Gardner's *A Frequency Dictionary of Contemporary American English* (2010) was adopted as the basis of the vocabulary section of the placement test, on the assumption that higher frequency words are generally more important to learn and more likely to be integrated into long-term knowledge because they will tend to be encountered more frequently in authentic use. Thus, Section 1 of both test forms aimed to measure vocabulary knowledge at different levels of word frequency. Each frequency band of 1000 words from Davies and Gardner (2010) was tested by 10 items, with the expectation that the average difficulty of items would increase as word frequency decreased, allowing the lexical burden of reading passages to be estimated for students at different TOEFL levels, providing evidence to guide textbook selection for classes of different levels.

RQ1. Did the difficulty of items in the vocabulary section follow the hypothesized hierarchy based on word frequency?

RQ2. Did students demonstrate good data-model fit, indicating that students from different high-schools followed similar trajectories of vocabulary learning?

Method

Participants

All participants were female Japanese university students enrolled in an academic English program at a public Japanese women's university. Placement tests were administered in April 2012 and April 2013 to assign students to both academic English classes and first-year seminar classes conducted in Japanese. The 2012 cohort had 249 students and the 2013 cohort had 243 students, for a total of 492 students.

Instrument

The vocabulary section of the test comprised 50 items in 2012 but was reduced to 40 items in 2013. Although the VST (Beglar, 2010; Nation & Beglar, 2007) was considered as a source of vocabulary items, the frequency lists provided by Davies and Gardner (2010) were considered to more relevant to the AEP's academic focus so a new test was developed using a synonym matching format instead of the definitions used in the VST. As the students were enrolled in an academic English program, knowledge of very high frequency vocabulary was assumed, so each item stem used a word taken from the first 500 words listed by Davies and Gardner (2010), with the correct answer, the key, being synonymous with this. The distractors were of similar frequency to the key and of the same part of speech, so the difficulty of items was hypothesized to result from the frequency of the key and distractors. A sample item is shown below:

With
A) Ago B) Least C) Enough D) Already E) Together

For the 2012 test, 10 items were sampled from each of the 1000 word frequency bands in Davies and Gardner's 5000 word list, giving 50 items in total. Analysis of the 2012 results showed that the items from the first and second 1000 frequency bands (1K and 2K) were too easy for most students, so these were replaced with 10 academic items derived from the reading textbook series for the 2013 test, leaving 40 items. This analysis therefore includes 60 items, with 1K and 2K items used only in 2012, academic items used only in 2013, and 3K, 4K, and 5K items linking the two subsets of data.

Procedure

Tests were administered on the first day of semester, supervised by AEP teachers. Administrative constraints dictated a two-hour time limit for the placement test, raising concern over speededness. Observation during test administration showed that most students finished all sections within the allotted time, and that speededness did not affect the vocabulary section, which was the first section of both test forms. Therefore, missing responses were coded as incorrect. All analyses were conducted using Winsteps (Linacre, 2010). Following each test administration test forms were scanned and processed using Remark Office OMR version 8.4 (Gravic, 2012), response data exported to Microsoft Excel 2010 (Microsoft, 2010), and then imported into Winsteps as a plain text file.

Results

Figure 2 shows the variable map, or Wright map, with mean item difficulty set as 400 and 1 logit scaled to 50 points, to give an approximation to the TOEFL scale. The vertical scale thus allows visual comparison of person ability and item difficulty because both are measured in the same units. Persons are shown in the left column, with item distribution shown in the second column, and items displayed by frequency band in the rightmost six columns. Items are labeled by frequency level, with targeted academic vocabulary labeled as "AW". Most persons fall within the TOEFL 400 to 500 range, consistent with students being unable to read unsimplified texts upon entry to the AEP. Students with TOEFL levels of 400 would have a 50% expectation of answering an average item, while students at the 500 level would have an 88% expectation of success on an average item. A general trend can be seen for high-frequency items to be easy and academic items to be difficult, but the pattern is not strongly deterministic, with one very easy 5K item and two very difficult 2K items. This is supported by Table 1, showing mean item difficulty by frequency band. This must be interpreted very cautiously because 10 items per frequency band is insufficient for definitive results, but Table 1 shows the expected trend of mean item difficulty increasing as vocabulary decreases. Although adjacent frequency bands aren't clearly separated, with some 3K words easier than most 1K and 2K words, the evidence supports the view that Davies and Gardner (2010) provide a useful classroom guide for prioritizing vocabulary items.

However, a curriculum aims to match students with language features of appropriate difficulty, and this is where the benefits of Rasch measurement become apparent. Although raw scores can rank-order person ability and item difficulty, they do not place person ability and item difficulty on a shared measurement scale. Also, rank-ordering using raw scores requires that all persons take the same set of items and that all items are taken by the same set of persons unless equating procedures are used, greatly complicating matters when different test forms are used, as in this study. Rasch analysis provides comparison of both persons and items on the same measurement scale even when different test forms are used, as long as the test forms have a subset of 10 or more common items that can be employed to statistically link the forms, so the Wright map shown in Figure 2 allows curriculum planners and classroom teachers to quickly see the relative ability of each student compared to instructional items. We can see that very few persons were below 400, but 1K and 2K words mostly fell below this, while academic words were mostly above the average person ability of about 470. Therefore, it seems reasonable to focus vocabulary instruction on 3K and 4K words for most students, while reviewing and consolidating 1K and 2K words with the lowest group and introducing academic words with the upper levels. In this way, Rasch analysis allows curriculum planners and materials writers to visually compare the levels of students and items to check that instruction is appropriately targeted.

However, as Figure 1 showed, the hierarchy of item difficulty of the Wright map assumes adequate data-model fit. The "pathway" maps produced by Winsteps provide a simple visual tool to investigate this. Figure 3 shows the pathway maps for items, shown in the left-hand panels, and persons, shown in the

right-hand panels. The vertical scale shows item difficulty and person ability. Each circle represents one item or one person, with the size of the circle representing the measurement error. If two circles overlap on the vertical scale, we do not have 95% confidence that they are different in difficulty or ability. The horizontal scale shows mean-square fit statistics, which can range from zero to infinity. A value of 1.00 indicates that the randomness in the data matches the expectations of the Rasch model, while values below this indicate unexpectedly predictable data and values higher than 1.00 indicate noisy data. Linacre suggests a rule-of-thumb that mean-square statistics between 0.5 and 1.5 are productive of measurement. However, two sets of mean-square statistics are produced, information weighted infit statistics, shown in the upper panels, and unweighted outfit statistics, shown in the lower panels. The information weighting of the infit statistic emphasizes responses where person ability and item difficulty are well matched because this is where information is maximized, so this statistic is the crucial indicator of whether the instrument supports measurement. The outfit statistic, generated from unweighted responses, shows the effect of outlying responses, such as when low ability persons succeed on difficult items or high-ability persons fail on easy items. Comparison of the patterns of infit and outfit thus gives important diagnostic information about where unexpected responses are occurring.

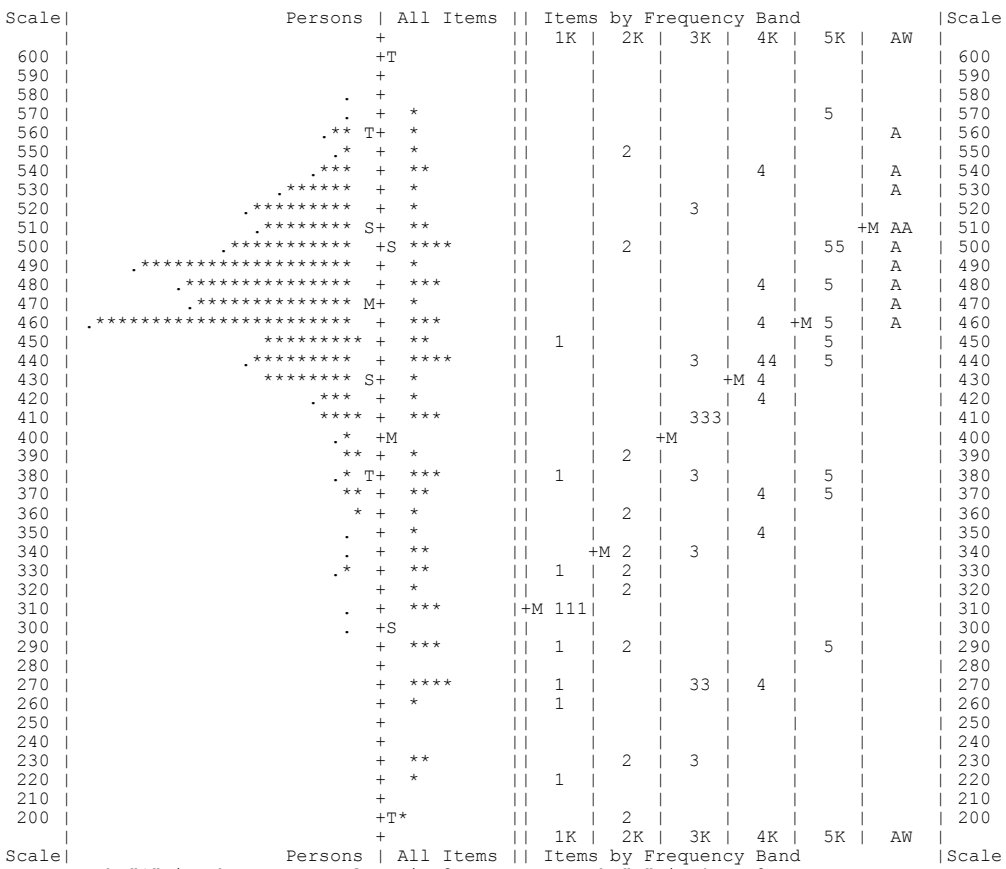

Figure 2. Person-item map showing person ability and item difficulty scaled to approximate TOEFL level. Items are identified by frequency band, with academic items identified as "A". "M" indicates the mean difficulty for all items, and the median for each frequency band.

Table 1
Item Difficulty by Frequency Band

		Item Difficulty			
Item Level	Count	Mean	Median	S.D.	S.E. Mean
1K	10	311.74	307.31	61.36	20.45
2K	10	351.23	337.51	102.94	34.31
3K	10	368.69	396.05	83.45	27.82
4K	10	418.75	432.46	71.51	23.84
5K	10	442.98	458.18	75.22	25.07
AW	10	506.62	505.19	30.13	10.04
All	60	400.00	412.22	98.08	12.77

Note: Subtotal reliability =.85
Scale: Mean item difficulty = 400, 1 logit = 50

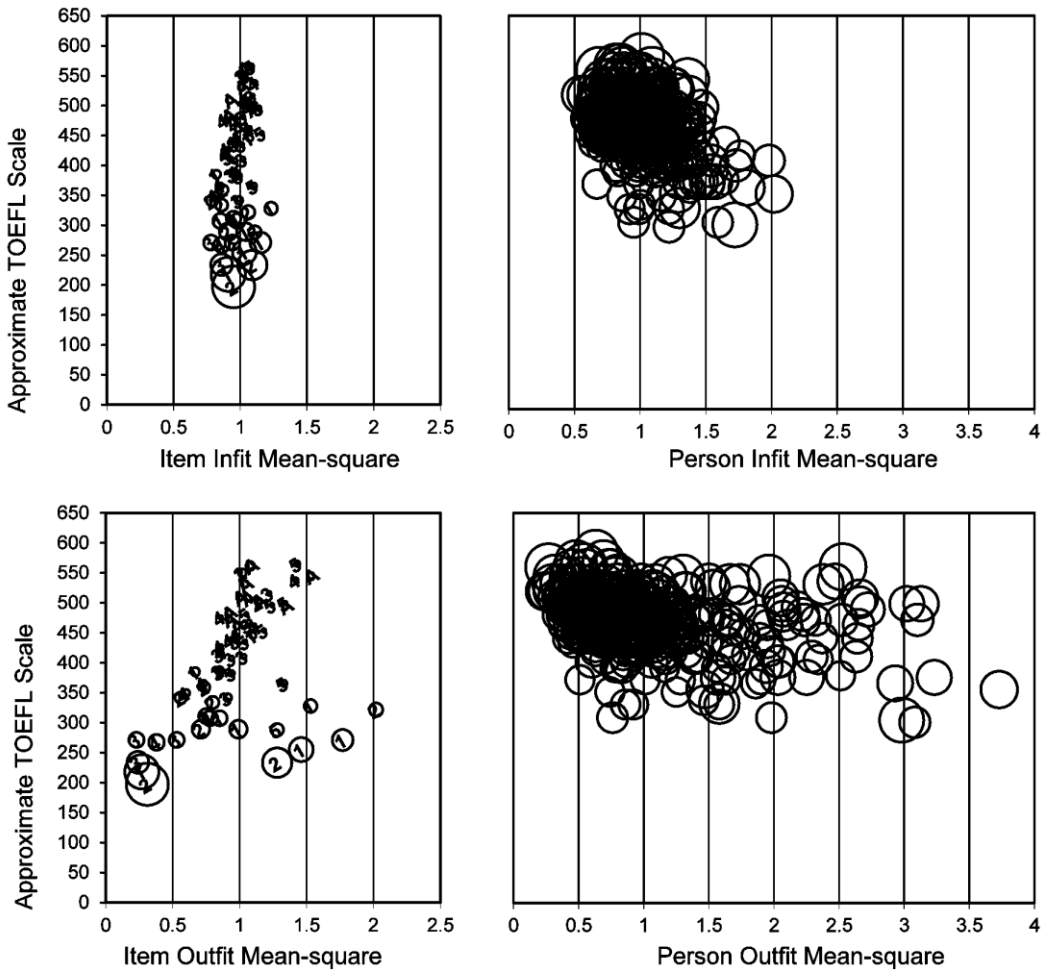

Figure 3. Pathway maps showing data-model fit. The vertical axis shows an approximate TOEFL scale. Each bubble shows a single item or person, with the size of the bubble indicating an approximate 95% confidence band of difficulty or ability. The horizontal axis shows mean-square fit statistics, with 1.5 being a rule-of-thumb threshold for concern.

In the case of the items, the infit statistics are extremely good, but the outfit statistics show several misfitting items that did not follow the parallel acquisition trajectories assumed by the Rasch model. The person statistics show a more worrisome pattern, with the infit statistics showing a number of misfitting persons and the outfit statistics showing many. Thus, many students are not displaying parallel trajectories of vocabulary acquisition, suggesting idiosyncratic exposure to English vocabulary at high-school or from studying for university entrance exams. Although the item statistics indicate a relatively stable hierarchy of item difficulty, the evidence points to many students having idiosyncratic vocabulary knowledge. This suggests the need for remedial instruction for higher ability students who incorrectly answered easy items, and thus might struggle with high-frequency vocabulary despite having considerable knowledge of academic vocabulary. Winsteps provides an accessible solution to this in the form of Kidmaps.

Figure 4 shows the Kidmap for one student. The central vertical scale shows item ability, with the student's ability estimated as 499 plus or minus 22 and the horizontal bars showing the 95% confidence band. The left-hand side of the map identifies the items that were answered correctly, while the right-hand side identifies the items answered incorrectly, so "35.1" indicates Item 35 was given a score of 1, while "33.0" indicates that Item 33 was given a score of 0. The items in the lower right quadrant show unexpected failures so this student should revise Items 33, 9, 27, 50, 45, and 43. What is notable is that this student is above average in ability but has followed an acquisition trajectory that diverges from the overall group, so the remediation is targeting easy items that were incorrect. In contrast to remediation aimed at helping low ability students close the gap to average students, this remediation targets higher ability students with idiosyncratic knowledge in order to bring them in line with the sequencing assumed by the curriculum planner. In contrast, the upper right quadrant shows expected failures, so this provides a sequence for non-remedial instruction of items above the student's current level.

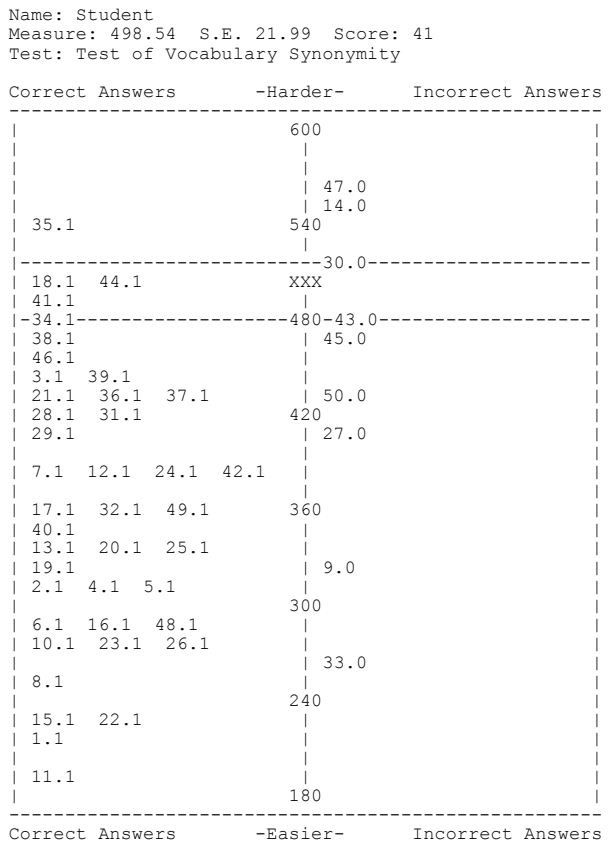

Figure 4. Diagnostic Kidmap for a single student showing correct and incorrect responses by item difficulty. Logit measures are shown on the vertical scale. The upper left quadrant shows items with unexpected success, indicating items requiring investigation, and the lower right quadrant shows items with unexpected failure, indicating items requiring remedial instruction.

Discussion and Conclusions

This study aimed to demonstrate how Rasch analysis can be of practical value to curriculum planners, materials writers, and classroom teachers using data from the vocabulary section of an academic English placement test. Although traditional analysis of raw scores can rank-order item difficulty and person ability, and techniques are available to criterion reference person ability to language features, Rasch analysis provides an extremely practical solution to these, while Rasch data-model fit provides a simple conceptual framework for diagnostic assessment. The key theoretical assumption of the Rasch model is that all items and persons follow parallel developmental trajectories and mean-square fit statistics provide an indication of the magnitude of deviations from this idealization. In this study, items showed acceptable data-model fit, supporting the existence of a stable hierarchy of item difficulty. The Wright map is emblematic of Rasch analysis and visually confirmed the hypothesized trend for item difficulty to increase as vocabulary frequency decreased. This provided a practical guide to inform teachers about vocabulary that is likely to cause difficulty for students of different levels. However, many students misfitted the Rasch model, suggesting idiosyncratic trajectories of vocabulary acquisition in high-school English classes and supporting the need for remedial instruction for high-level students with misfitting response patterns. The Kidmap produced by Winsteps provided an individualized diagnostic report to identify test

items requiring remediation. These findings illustrate that Rasch analysis has benefits for language programs beyond the identification of misbehaving items, providing insights into the behavior of individual students that are conceptually simple enough for non-specialists to interpret.

Acknowledgments

This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research Grant #25370643. We are also very grateful to Jim Sick for his valuable comments and feedback.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101-118. doi: 10.1177/0265532209340194

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.

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

Douglas, N., & McIntyre, P. (2009). *Reading explorer*. Boston: Heinle.

Engelhard, G. (2013). *Invariant measurement*. New York: Routledge.

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

Linacre, J. M. (2010). Winsteps (Version 3.70.02). Retrieved from http://www.winsteps.com

Linacre, J. M. (2012). A User's Guide to Winsteps, Minstep Rasch-Model Computer Programs Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Microsoft. (2010). Excel (Version 2010).

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13.

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.