Evaluation of a Streaming Instrument

Hiroshima Bunkyo Women's University
Judith Runnels

Abstract

Throughout the last decade, the Rasch model has frequently been used as a tool to evaluate assessments. Rasch analysis provides estimates of item-difficulty and person-ability that are not dependent on raw scores. It is therefore particularly suitable for estimating ability from a test for the purposes of streaming students into levelled classes. The analysis can also be used to evaluate the assessment by highlighting items that are causing unexpected response patterns. The current results demonstrate that a multiple-choice test of English designed to stream students acted as a fairly effective tool for distinguishing the abilities of the population of test takers. However, several items appear to require modifications prior to future administrations of the test, with the hope that the streaming process will be rendered more precise. The suitability of using Rasch-based methods to evaluate streaming tools is discussed.

Introduction

Recently, Rasch analysis has replaced classical test theory analyses of pedagogical assessment because of its differing theoretical stance and practical applications. Compared to deterministic methods, Rasch analysis is based on a prescriptive model, meaning that conditions for the data to meet are prescribed, rather than being pre-determined (Salzberger, 2010). In other words, instead of rather than building

a model from a set of data, the Rasch model specifies criteria that a successful measurement tool should meet (Bond & Fox, 2007). Whereas in classical test theory raw test scores are used as estimates of ability, in Rasch-based methods, the focus is on the pattern of item responses. The advantage of this is that the assumptions about or estimates of student abilities are not dependent on a single test, or a specific number of items from that test (Wright, 1999). Essentially, Rasch analysis takes individual differences in ability into account while measuring the difficulty of items and vice versa (Rasch, 1966). Specifically, in the Simple Rasch Model the probabilities of providing correct responses to items of specific estimated difficulties by persons of estimated abilities are calculated (Wright & Stone, 1979). This information is then interval-scaled, producing an illustration (graphical or numerical) of the relationship between test-taker and item scores - one which cannot exist using raw scores alone (Wu & Adams, 2007). Rasch analysis is also frequently used to provide validity evidence for pedagogical assessments and in particular for multiple choice tests (Runnels, 2011; Beglar, 2010; Baghaei, 2008; Meara & Buxton, 1987). The results can be used to highlight test items which are causing some degree of unexpected response patterns among test takers (Edwards & Alcock, 2010). Knowing this information allows test developers to create arguably more reliable assessments and provides useful feedback for administrations to use when placing students in classes according to ability, otherwise known as streaming (Tyndall & Mann, 1996).

The current study was designed to analyze the effectiveness of a proficiency test developed to determine the top 30% of test-takers. All test-takers are members of a first-year cohort of about 300 students entering a private university in Japan. The goal of the test was to identify approximately the top 90 students, who would then be streamed into 3 or 4 different English classes. For in-house assessments designed

to meet specific requirements, it is imperative to check that there are no inherently problematic items which may be affecting test-takers' responses, especially if the goal of the test is to distinguish ability for streaming (Jackson *et al*., 2002). Several aspects of Rasch analysis are used to identify items which might confound the process of student ranking, thus affecting the efficacy of the test as a streaming tool. Specifically, point-measure correlations are used as a measure of how strongly the item is measuring the direction of the construct. Any items causing high ability students to respond incorrectly when low ability students are responding correctly are likely to exhibit negative point-measure correlations and may skew the streaming process (Boone & Scantlebury, 2006). Fit statistics provide another way to check the item's relevance to the intended construct, by highlighting any misfitting items which are possibly representing a different construct (Smith & Suh, 2003). Fit statistics also examine how a test-taker's response patterns match those predicted by the model. Person-item maps (graphical representations of item difficulty and person ability measures) illustrate gaps in abilities or difficulties, suggesting that some domain of the construct has not been assessed (Baghaei, 2008). This can also be examined with item strata, which identify statistically distinct difficulty levels, thus ensuring a range of item-difficulties have been included (Wright & Masters, 2002).

For the current analysis, an item-person map, point-measure correlations, item and person strata and fit statistics will be analyzed with regards to evaluating the effectivness of a newly designed assessment as a streaming instrument. The ultimate goal of the analysis is to determine if the test is able to distinguish a spread of student abilities which can ultimately be used to create levelled classes. Furthermore, the results of the Rasch analysis will be used to highlight potentially problematic items, which can subsequently be used to improve the assessment as a streaming instrument

prior to future administrations of the test.

Method

Participants

306 incoming students of Hiroshima Bunkyo Women's University in Hiroshima City, Japan, took the test in 2012 prior to the start of the academic year. All examinees had completed six years of required English classes at junior and senior high school.

Instrument

The test was a 64 item multiple choice test of English covering topics that students would be studying during their upcoming year of English study. The test lasted 95 minutes and was administered during a specially scheduled time prior to commencement of classes. The test results had no bearing on grades and was solely employed to help with dividing the student population into levelled classes, with the upper 30% of abilities being the primary targets of the streaming process.

Procedures

Data were analyzed using WINSTEPS® Rasch software version 3.66.0 (Linacre, 2008). To determine item difficulty and person ability, the Rasch measure was calculated for all items and test-takers. The Rasch measure is the probability of a person correctly responding to a given item and is related to their ability and the difficulty of the item (Rasch, 1960). It is calculated with the following formula:

$$\log[\, pni/(1 - pni)] = Bn - Di \quad (1)$$

where *Bnis* the ability of a person *n* and *Di* is the difficulty of item *i*.

To identify statistically distinct item difficulty or person ability levels, item and person strata are calculated (Wright & Masters, 2002). Smith (2001) requires a minimum of two difficulty levels in order to be able to deem items representative

of the assessed content. The following formula is used to calculate item and person strata (Beglar, 2010):

$$\text{Item strata} = (4G_{item} + 1/3)$$

where *Gitem* is the Rasch item seperation value (derived by dividing the item standard deviations by the average measurement error).

A person-item map which graphically illustrates the relationship between person-abilities and item-difficulties is examined to explore the spread of items, the spread of abilities, as well as gaps or overlaps in item difficulty and ability (Stelmack *et al*., 2004). The item measure correlations are calculated by a fairly complex formula involving the predictability of data, item targeting for the abilities of the sample of test-takers, as well as the distribution of the person measures (Linacre, 2004). Item correlations close to zero suggest that the item may be measuring the test construct differently (Wolfe & Smith, 2007). Negative item measure correlations suggest that the item is opposing the direction of measurement and are therefore flagged for further investigation (Bond, 2003).

Finally, to check for any items that cause unexpected response patterns, fit statistics are calculated. Infit statistics reflect response patterns where the test is targeting ability, while outfit statistics highlight unexpected responses (Linacre, 2007). Infit and outfit statistics are manifested in mean-square values' (MNSQ) size and z-standardized scores (ZSTD) which indicate the significance of the misfit. According to Linacre (2007), acceptable values for low-stakes multiple choice tests for MNSQs range from 0.7 to 1.3 and -2.0-2.0 for ZSTDs.

Results & Discussion

Table 1 illustrates some summary statistics for items and test-takers. The overall mean score on the test was 55.7% (*SD* = 12.6), cronbach's alpha for items was

0.99 whereas for person measures, the reliability was somewhat lower, at 0.81. The low mean score (approximately 56%) is not of major concern, since this test had zero impact on students' grades- it was solely used as a streaming instrument. The mean Rasch measures are provided in the MEASURE column (Table 1). The item separation is also provided. to accept the test as representative of the assessed construct, there should be at least two levels (Smith, 2001). An item strata of 8.71 illustrates a wide range of item-difficulty. On the other hand, there is a comparatively lower separation value for person abilities at 2.09, although this is enough to accept that there are statistically distinct ability levels within the student population.

Table 1 Summary Statistics

	TOTAL SCORE	COUNT	MEASURE
MEAN	35.7	64.0	53.59
S.D.	7.9	.0	7.31
Person SEPARATION	2.09	Person RELIABILITY	.81
Item SEPARATION	8.71	Item RELIABILITY	.99

Figure 1 shows the test's variable map which illustrates a reasonable range of item-difficulty and test-taker ability. The most difficult item (I0061) is shown at the top of the figure on the right of the y-axis and the most-abled student is the highest on the left-hand side. In Figure 1, it can be seen that there is a general pattern of item difficulties spanning within and slightly beyond the abilities of the test-taking population. Several items fall outside the abilities of the test-takers, moreso on the low end of the spectrum than the high end. At the top of the y-axis, it can be seen that there are two items that are well beyond the abilities of test-takers. If there were more test takers at the difficult end of this spectrum (in the 70-100 logit range), then more items would be required to ensure measurement of all abilties was being

covered. However, these two items in particular (items 61 & 11) certainly require further investigation, since they are not contributing to the precision of measurement by being so far beyond the abilities. It could either mean that there is some inherent property of these items that is either confusing or misleading, that the topic of the item is unfamiliar to test-takers, or simply that the difficulty truly does go beyond the English abilities of the student population. Nevertheless, these items should not necessarily be removed from the assessment in order to prevent a ceiling effect (Stelmack *et al*., 2004): they should instead be adjusted with the goal of bringing them to only slightly above student abilities, rather than over 20 logits beyond. A greater number of items around the 65 to 80 logit range would prevent a ceiling effect and result in a more effective measurement tool.

Despite the large gaps at the difficult end of the spectrum, there are no major gaps across the remaining item difficulties. The abilities of students however, are clustered slightly more tightly than difficulties. If the ultimate goal of this test was to stream all students into levelled classes, then the measurement tool may have to be adjusted due to the abundance of test-takers in the 50 to 60 logit range of ability. However, since the goal was to highlight the 30% of most abled students, this test appears to have been relatively successful since approximately a third of students responded correctly to items ranging between 60 and 70 difficulty logits, and this range is spanned by 13 items. The test could be further improved by increasing the difficulty of some of the items that fall well below the abilities of all students. In the case of items 41, 44, 46, 49, 55, these items could potentially be eliminated from the assessment since there are no participants at those ability levels: they thus, do not contribute to the precision of measurement. Ultimately, there are no major gaps across neither item difficulties nor abilities, thus suggesting that the test distinguished the range of student abilities effectively enough for its purposes of separating out

the top third of students. With the exception of the seven aforementioned items, the variable map suggests that the test is reasonably well-targeted for this group of test takers.

Table 2 shows fit statistics, point-measure correlations and Rasch measures for the aforementioned 7 items as well as an additional two misfitting items described hereafter. No items exhibited misfitting infit. There are five items with misfitting outfit, as follows: 11, 25, 51, 61 and 41 - the former four at the difficult end of the spectrum, while the latter item is the easiest item on the test (Table 2). Of these five, only the difficult items (11, 25, 51 and 61), have a ZSTD outside of the acceptable range, in addition to extremely low (even negative, in the case of items 51 and 61) point measure correlations. Item 61 is the most difficult item (99.49 logits) and it is likely that test-takers who endorsed this item (4 out of 306) did so by random guessing. The remaining four items also caused some sort of unexpected response patterning, likely due to test-takers for which the model predicts would respond incorrectly, in fact gave successful endorsement and vice versa. Further investigations of these items are required with modification or elimination from future test administrations being possible options. Nonetheless, the feedback from these analyses can be used constructively: while there are nine items which require revisiting, the remaining items appear to measure the construct in a consistent direction.

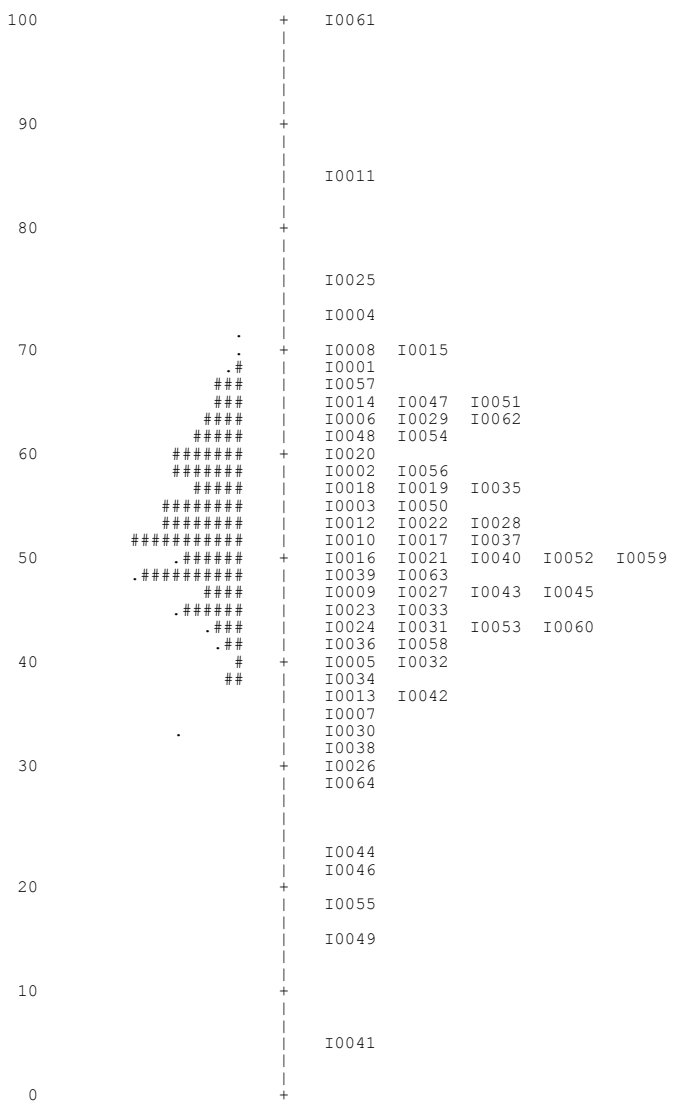

Figure 1. Variable map for person ability and item difficulty parameters. Items referred to in the text have been highlighted.

Table 2 Rasch analysis output for Rasch measures, fit statistics and point measure correlations for items flagged for further analysis.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.
61	4	306	99.49	1.03	.2	2.93	2.3	-.08	.08
11	15	306	85.64	1.00	.1	1.94	2.5	.05	.16
25	38	306	75.05	1.11	.9	1.67	3.3	.00	.23
51	81	306	64.95	1.23	3.3	1.41	4.1	.03	.30
44	290	306	22.34	.97	.1	.83	.5	.21	.15
46	292	306	20.89	1.00	.1	.88	-.3	.16	.14
55	295	306	18.32	.97	.0	.78	-.5	.19	.13
49	298	306	14.97	.98	.1	.74	-.5	.16	.11
41	303	306	4.88	.94	.1	.24	-1.5	.24	.07

Conclusion

The current article used Rasch analysis to evaluate a multiple choice English test. The test exhibited very few problematic items and was able to distinguish student ability effectively enough to use the information for streaming the most abled students into levelled classes. Furthermore, the results offered some useable feedback in terms of identifying items which may require some modification before future administrations. The current study, while attempting to evaluate a test as a measurement tool for ability, did not include such analyses as unexpected response or item distractor analyses. In the current analysis, no test of unidimensionality was performed and while the infit statistics provide evidence towards this test being an appropriate subject for Rasch analysis, further measures should certainly be taken. Additionally, determining if there are any items that are causing unexpected response patterns either across groups or across sections of the test (differential item or test functioning) should also be included. Weaver (2007) deems differential item

functioning an imperative check, especially when the results of a test are used for the streaming of students in different majors of study. Ultimately, the results of a Rasch analysis provided useable information which allowed not only for separating students by ability but also contributed to the future processes of development, modification and monitoring of in-house designed pedagogical assessment.

References

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

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.

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

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

Bond, T.G. (2003). Validity and assessment: a Rasch measurement perspective. *Metodologia de las Ciencias del Comportamiento* 5(2), 179-194.

Boone, W., & Scantlebury, K. (2006). The role of Rasch analysis in science education utilizing multiple choice tests. *Science Education.* 90, 253-269.

Edwards, A., & Alcock L. (2010). Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics Applications,* 29(4), 165-175.

Jackson, T.R., Draugalis, J.R., Slack, M.K., Zachry, W.M., & D'Agostino, J. (2002). Validation of Authentic Performance Assessment: A Process Suited for Rasch Modeling, *American Journal of Pharmaceutical Education,* 66, 233-243.

Linacre, J. M. (2004). Test validity and Rasch measurement: construct, content, etc. *Rasch Measurement Transactions,* 18(1), 970-971.

Linacre, J. M. (2007). A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs. Chicago, IL: winsteps.com.

Linacre, J.M. (2008). *A User's Guide to Winsteps/Ministeps, Rasch Model Computer Programs.*

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing,* 4, 142–154.

McNamara, T.F. (1996). *Measuring second language performance.* New York: Longman.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology,* 19(1), 49-57.

Runnels, J. (2011). Evaluation of an Achievement Vocabulary Test Using Rasch Analysis. *Studies in Linguistics and Language Teaching,* 22, 165-185.

Salzberger, T. (2010). Does the Rasch Model Convert an Ordinal Scale into an Interval Scale? *Rasch Measurement Transactions,* 24(2), 1273-1281.

Smith, E.V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement,* 2(3), 281-311.

Smith, R.M., & Suh, K.K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates, *Journal of Applied Measurement,* 4(2), 153-163.

Stelmack, J., Szlyk, J.P., Stelmack, T., Babcok-Parziale, J., Demers-Turco, P., Williams, R.T., Massof, R.W. (2004). Use of Rasch person-item map in exploratory data analysis: A clinical perspective, *Journal of Rehabilitation Research & Development,* 41(2), 233-242.

Tyndall, B., & Mann, D. (1996). Validation of a New Holistic Rating Scale Using Rasch Multi-faceted Analysis. In Cumming, A & Berwick, R.(Eds.), *Validation in Language Testing* (pp. 39-57). Clevedon: Cromwell Press.

Weaver, C. (2007). A Rasch-based evaluation of the presence of item bias in a placement examination designed for an EFL reading program. *Second Language Acquisition- Theory and Pedagogry: Proceedings of the 6th Annual JALT Pan-SIG Conference,* 84-96.

Wolfe, E.W., & Smith, E.V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement,* 8(2), 204-234.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement.* Mahwah, NJ: Lawrence Erlbaum.

Wright, B.D., & Masters, G.N. (2002). Number of person or item strata, *Rasch Measurement Transactions,* 16, 888.

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

Wu, M. & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions.