Multifaceted Rasch Analysis:

Program in Intensive English Writing Placement Exam

David Tasker

Northern Arizona University

Abstract

This research used Many-Facet Rasch Measurement (MFRM) to investigate the functioning of the writing section of a placement test. Data included 97 examinees' scores from the fall 2015 placement exam at the Northern Arizona University Program in Intensive English. Statistical analysis revealed findings central to the validity of the assessment program: the test showed appropriate difficultly and led to a spread of examinee abilities; most raters produced nearly interchangeable, bias-free ratings; and rating scale levels reflected distinct ability levels. However, several critical issues were also revealed: the test identified fewer distinct examinee ability levels than desired; tasks did not cover different difficulty levels; certain scale levels did not contribute meaningfully to scoring; a rater who scored many tasks showed unacceptable leniency and L1 bias; and some examinees may have received similar writing scores despite ability differences. This study also used MFRM-derived inferences to identify validity supporting revisions: it is suggested that rubric wording at wide-score levels be revised, certain scale levels be reconsidered, rater training be targeted toward leniency and bias patterns displayed here, and that MFRM logit scores replace raw scores in placement decisions.

Multifaceted Rasch Analysis:
Program in Intensive English Writing Placement Exam

Background

While different assessment measures are subject to differing confounding influences, writing and speaking assessment are susceptible to a particularly large array of factors. These performance-based assessments are subject to the systematic influences of the raters, tasks, and scales, on which they rely to produce observed scores (McNamara, 1996). The influence of these non-ability factors on raw scores has been well documented in numerous Many-Facet Rasch Measurement (MFRM) studies (e.g., Lynch & McNamara, 1998; Barkaoui, 2013; Bond & Fox, 2007; Engelhard, 1992; McNamara, 1996; etc.). As articulated by McNamara (1996), in this statistical model, each rating is understood to be a function of the interaction between examinee, task, scale, and rater (and criteria, if present). Because “a number of aspects of the examination process contribute to the difficulty that the candidates face in revealing their abilities” (Bond & Fox, 2007, p. 159), all such factors must be investigated and minimized to ensure that scores are valid, particularly when such scores lead to placement decisions that are crucial to the functioning of an English program. MFRM provides a comprehensive method for the identification and analysis of these systematic difficulties examinees face revealing their true abilities. Most importantly, MFRM statistics are sample-independent (Fulcher, 1996), meaning that findings for a particular administration of a test easily generalize to other administrations. This research therefore uses a multifacet Rasch analysis to identify rater-, task-, and scale-related factors confounding the expression of true ability scores; it then uses these MFRM-identified influences on scores to identify and suggest revisions that will further support the reliability, validity, and meaningfulness of placement decisions for a program in intensive English placement test.

Research Questions

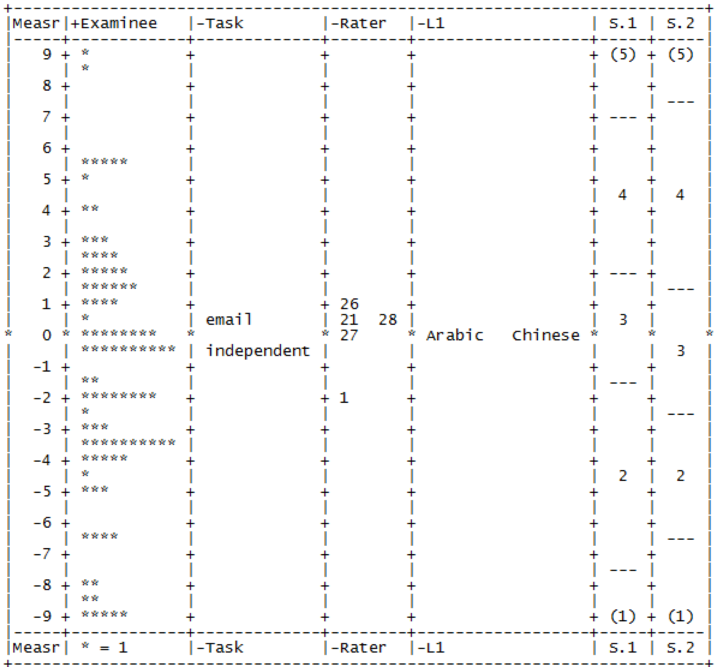
1. Are the abilities of PIE writers properly measured, in a replicable manner?
2. Does writing task difficulty contribute to determining PIE writing level?
3. Do raters use task-specific rating rubrics well to place PIE students into appropriate levels?
4. What rater patterns (including bias) might be affecting raw scores?

Methods

Because data used was de-identified previous to investigation and the research involved no treatment, the study received an Institutional Review Board classification as Not Human Subjects Research. Archived data were collected from the fall 2015 placement test. Initial data included 108 examinees, 5 first languages (L1), 17 raters, and 3 tasks. The initial statistical processing, using a Partial Credit Model in FACETS, revealed two disconnected subsets, which was resolved by eliminating data that failed to meet certain criteria: raters included must have rated 35 or more tasks, and examinees must only come from the two most common L1s (Chinese, and Arabic). Revised data included 97 examinees, 5 raters, 2 L1s, and 2 tasks: the integrated writing task was removed, as it had only been scored by raters who did not perform the threshold number of ratings. The few 0 scores present were also removed. As L1 data was used only to investigate rater-L1 bias interactions, they were treated as a dummy facet. Revised data were statistically processed using a Partial Credit Model in FACETS.

Results and Discussion

For the preliminary discussion of results, the variable map for the facets investigated is presented in Figure 1.



```
+--------------------------------------------------------------------------+
|Measr|+Examinee    |-Task        |-Rater  |-L1               | S.1 | S.2 |
|-----+-------------+-------------+--------+------------------+-----+-----|
|   9 + *           +             +        +                  + (5) + (5) |
|     | *           |             |        |                  |     |     |
|   8 +             +             +        +                  +     +     |
|     |             |             |        |                  |     | --- |
|   7 +             +             +        +                  + --- +     |
|     |             |             |        |                  |     |     |
|   6 +             +             +        +                  +     +     |
|     | *****       |             |        |                  |     |     |
|   5 + *           +             +        +                  +     +     |
|     |             |             |        |                  |  4  |  4  |
|   4 + **          +             +        +                  +     +     |
|     |             |             |        |                  |     |     |
|   3 + ***         +             +        +                  +     +     |
|     | ****        |             |        |                  |     |     |
|   2 + *****       +             +        +                  + --- +     |
|     | ******      |             |        |                  |     | --- |
|   1 + ****        +             + 26     +                  +     +     |
|     | *           | email       | 21  28 |                  |  3  |     |
*   0 * ********    *             * 27     * Arabic   Chinese *     *     *
|     | *********** | independent |        |                  |     |  3  |
|  -1 +             +             +        +                  +     +     |
|     | **          |             |        |                  | --- |     |
|  -2 + ********    +             + 1      +                  +     +     |
|     | *           |             |        |                  |     | --- |
|  -3 + ***         +             +        +                  +     +     |
|     | *********** |             |        |                  |     |     |
|  -4 + *****       +             +        +                  +     +     |
|     | *           |             |        |                  |  2  |  2  |
|  -5 + ***         +             +        +                  +     +     |
|     |             |             |        |                  |     |     |
|  -6 +             +             +        +                  +     +     |
|     | ****        |             |        |                  |     | --- |
|  -7 +             +             +        +                  +     +     |
|     |             |             |        |                  | --- |     |
|  -8 + **          +             +        +                  +     +     |
|     | **          |             |        |                  |     |     |
|  -9 + *****       +             +        +                  + (1) + (1) |
|-----+-------------+-------------+--------+------------------+-----+-----|
|Measr| * = 1       |-Task        |-Rater  |-L1               | S.1 | S.2 |
+--------------------------------------------------------------------------+
```

Figure 1. Variable map for examinee, task, rater, L1, and scale facets.

A few important data trends are apparent in the variable map. Beginning from the left, one can see that examinee ability on this test spreads across approximately 18 logits, a rather wide range, as one might hope for on a placement test. The distribution of this spread is also somewhat desirable, appearing to peak near zero logits and mostly taper off further up and down the scale. At the same time, one can see a few areas of uneven coverage, where major gaps in examinee ability can be found. In the task column, it can be seen that the two tasks appear to have similar, though not exactly equal, difficulties clustering near the zero logit level. In the rater column, four of the five raters cluster with similar though not equal difficulty just above the zero logit range, which is mostly desirable; however, rater 1’s notably different position at approximately -2 logits indicates a greater departure in leniency. In the next column to the right, the two L1s appear to be even because they were used as dummy facets: their positioning is not meaningful. In the final two columns on the right, the horizontal lines indicate the logit levels at which the scales (email, then independent) make separations. Here, one can see somewhat even sizes for most levels, while level 2 on the first scale and level 4 on the second scale cover a wider range of ability/difficulty levels. Also, it is seen that only two examinees' abilities match with scale level 5: this level appears little used.

Relevance to Program in Intensive English and Language Learning

A primary aim of this study was to use MFRM to produce actionable suggestions for improving the validity of the assessment measure. Because certain scale levels absorbed more of a range of scores than desirable, these rubric levels (most importantly, 2 in the email task) can be targeted for revision. A second suggestion is that, because neither scale made much use of level 5, this level can be identified as unnecessary in its current iteration. After combining scale levels 4 and 5 for both scales, a significant Pearson correlation coefficient (.99) revealed that such a combined scale would produce nearly identical rankings, suggesting that scale level 5 can be removed without affecting scores. Thirdly, because examinee overfit suggests equal scores are likely being received despite differences in ability level, and because raters overall were shown to have varying levels of severity, implications regarding raters are easily identified. First, rater training, particularly with a focus on more clearly differentiating ability levels and avoiding leniency or L1 bias, would positively affect the validity of scores produced. While rater training should be maximized, it is clear that confounding rating patterns are still likely to persist, even if self-consistency is supported (Kondo-Brown, 2002; Eckes, 2002). Therefore, it is also suggested MFRM scores, which account for and adjust to the influences of all relevant facets and are expressed on a true interval scale (logits), are used. Such a practice would maximize the accuracy and reliability of scores while respecting real-world constraints to rater training, substantially improve the validity of placement decisions, and facilitate better program functioning through more effective level placement. Each of these results-based suggestions, are of interest for language learning and assessment generally, and the Program in Intensive English writing section of the placement exam in particular.

References

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. Kunnan (Ed.) *The companion to language assessment* (pp. 1301-1322). Cambridge: Wiley.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model (2nd ed.). New York: Routledge.

Eckes, T. (2002). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*, 270-292.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, *13*(1).

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

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