Rasch Measurement in Language Education Part 8:

Rasch measurement and inter-rater reliability

James Sick
American Language Institute, Tokyo Center
New York University School of Continuing and Professional Studies

The previous installment of this series dealt with how many-facet Rasch analysis (MFRA) can be used to adjust for differences in rater severity when measures are constructed from subjective judgments. This installment addresses the related issue of rater agreement and inter-rater reliability from the perspective of classical test theory (CTT) and Rasch measurement theory (RMT).

Question:

I recently completed a many-facet Rasch analysis using Facets on of a set of essays marked by 8 teacher raters. Each essay was marked by two raters on a nine-point scale. If ratings differed by more than two points, a third rater was asked to arbitrate. I would like to know the inter-rater reliability of this assessment. Facets Table 7, the Raters Measurement report, indicates a reliability of .99 but states that this is NOT inter-rater reliability. If not inter-rater reliability, what is it? Where in Facets can I find inter-rater reliability? Table 7 also has a column indicating percentages of "Exact Agreement," observed and expected, for each rater. I assume these refer to rater agreement statistics, which I understand, but expected agreements vary amongst raters, ranging from 22 to 45 percent. Why are some raters expected to agree more than others, and for that matter, why are they all not expected to agree 100 percent?

Answer:

In classical test theory, the reliability of tests requiring subjective judgments, such as essays or speaking performances, is generally assessed using agreement ratios or inter-rater reliability. There is no universally agreed index of inter-rater reliability, however. A correlation coefficient can be used to assess the degree to which two raters agree in their rankings. When more than two raters are involved, an average correlation or a correlation adjusted with the Spearman-Brown formula may be used as an index of overall agreement in rankings (see Brown, 1996). I say "rankings" because the correlational approach can obscure systematic differences in rater severity. For example, if one rater awards scores of [5, 4, 3, 2] while another awards scores of [4, 3, 2, 1], the raters will be perfectly correlated, even though they do not agree in their assessments. Another approach is to calculate the proportion of exact or nearly exact agreements amongst raters. Agreement ratios are easy to understand and have practical significance when disagreements above a certain threshold prompt an additional rating. Cohen's Kappa and Fleiss' Kappa are refinements of the agreement approach (see Brown, 2012). Both derive reliability indices from the percentage of exact agreements, adjusting for the probability that agreements can occur by chance. All of these approaches share the common aim of separating the variance due to examinee performance (the true score) from the variance due to the vagaries of subjective judgment (the error).

Before answering your questions, it will be helpful to discuss some philosophical differences between CTT and RMT regarding subjective assessment. In CTT, an essay score is based on rating descriptors and is considered an attribute of the essay. The goal of the raters is to apply the rating rubric with machinelike consistency. Rater disagreement is seen as an indication that one or both raters have not assigned the most appropriate score, and is thus regarded as undesirable. In the CTT approach, test quality is pursued by minimizing disagreements through calibration sessions, fine-tuning the descriptors, and arbitration when large disagreements are identified. Test reliability is estimated by inferring, from their level of agreement, the degree to which raters are identifying the true score.

In contrast, raters in a many-facet Rasch analysis are viewed as independent experts who will sometimes disagree in their assessments. Although calibration sessions and descriptors are considered useful for achieving a shared understanding of the construct, individual essays are regarded as inherently too complex to be consistently matched to a set of descriptors. From the Rasch perspective, essay evaluation is better accomplished by engaging experts with multiple perspectives to respond to the unique characteristics of each individual essay. Scores awarded by independent raters, however, cannot be used "as is" because they are the product of both rater and examinee characteristics. Instead, initial scores are an intermediate step: data that can be used to construct measures of the underlying attribute, in this case English writing ability, that produced the variation in the essays.

In RMT, raters are hypothesized to possess a psychological trait that we label "severity" which leads them to systematically award higher or lower scores than their co-raters. Variation in severity arises from differences in personality, culture, and experience, but most likely reflects the fact that some people are adept at spotting flaws, while others tend to focus on strengths. Both perspectives are valuable, provided that severity is taken into account when constructing measures or making decisions. A many-facets Rasch analysis uses rater disagreements to estimate severity, and adjusts examinee measures accordingly. Rater severity is expected to influence the initial scores in a probabilistic manner consistent with the Rasch model. This is not a given, however. The analyst must verify that the raters fit the Rasch model as part of the process of validating the assessment.

To better address your questions, I've included a Facets raters measurement report from an analysis similar to the one you described (Table 1). This table reports rater severity and fit statistics for ten raters who provided two ratings for 396 essays. These raters varied considerably in severity, ranging from a high of 2.40 logits to a low of -2.60 logits. The reliability index at the bottom of the table indicates the reliability of the severity measures. That is, the degree to which these severity measures would be reproduced if the raters evaluated a similar sample of essays. Reliability is high because there is a lot of variance in rater severity and a large sample of shared ratings (agreement opportunities) from which to estimate it. If you are conducting a study of rater behavior, this reliability will be of interest to you. If your chief interest is examinee performance, however, it is not particularly relevant. Nevertheless, let me point out that a low reliability in the rater measurement report indicates a high level of rater agreement. If raters agreed 100 percent, there would be no variance in severity and the reliability would be zero. If your situation requires high rater agreement for reasons of face validity, a low reliability in the rater report is desirable.

Table 1. ***Raters Measurement Report (Facets Table 7.3.1)***

```
+-----------------------------------------------------------------------------+
|        Model | Infit      Outfit     | Exact Agree. |                       |
|Measure  S.E. | MnSq ZStd  MnSq ZStd  | Obs %  Exp % | Nu Raters             |
+--------------+----------------------+--------------+-----------------------
|   2.40   .19 |  .67 -2.1   .66 -2.2  |  20.0   27.7 |  2 Hayward            |
|   2.20   .20 |  .65 -2.2   .63 -2.2  |  21.1   24.8 |  6 Garland            |
|   2.08   .20 |  .79 -1.2   .75 -1.5  |  30.9   43.8 | 10 Brando             |
|    .70   .20 |  .93  -.3   .99   .0  |  23.0   29.7 |  3 Cruz               |
|   -.41   .21 |  .50 -3.4   .52 -3.1  |  37.3   35.8 |  5 Leigh              |
|   -.84   .18 |  .95  -.2   .94  -.3  |  29.8   42.6 |  1 Grable             |
|  -1.04   .19 | 1.03   .2  1.08   .5  |  17.6   30.3 |  4 Temple             |
|  -1.07   .17 |  .97  -.1  1.02   .1  |  23.7   33.7 |  8 Monroe             |
|  -1.43   .19 | 1.61  3.0  1.53  2.7  |  26.3   41.8 |  9 Rogers             |
|  -2.60   .20 | 1.42  2.1  1.47  2.3  |  19.0   29.2 |  7 Davis              |
+--------------+----------------------+--------------+-----------------------
|    .00   .19 |  .95  -.4   .96  -.4  |              | Mean (Count: 10)      |
|   1.65   .01 |  .33  1.9   .32  1.9  |              | S.D. (Population)     |
|   1.74   .01 |  .34  2.0   .34  2.0  |              | S.D. (Sample)         |
------------------------------------------------------------------------------+
Model, Sample: Separation 8.95  Reliability (not inter-rater) .99
Inter-Rater agreement opportunities: 396
Exact agreements: 97 =  24.5%
Expected:  134.9 =  34.1%
```

As for “where can I find inter-rater reliability in Facets,” that statistic belongs to the realm of CTT, and Facets does not report it. The Rasch equivalent would be the reliability reported in the Examinee Measurement Report, Facets table 7.1.1. This index answers the question “how likely would these measures be reproduced if the examinees produced another set of essays that were evaluated by a similar sample of raters?” I have not included the examinee report here, but for the analysis above, examinee reliability was .91. Let me emphasize that this figure applies to the Rasch measures, as opposed to the raw scores. The rater measurement report indicates that the raw scores were highly influenced by rater differences and are not very dependable as indicators of examinee performance. If inter-rater reliability were calculated, I would expect it to be rather low.

Column 3 of Table 1 reports the ratios of observed and expected exact agreements. The reason why expected agreements are not 100 percent should now, I hope, be clear. The reason expected agreement varies amongst raters is that the probability of agreement is dependent on the relative severity of the co-raters. Hayward and Garland, for example are similar in severity and would be expected to agree with each other frequently. Hayward and Davis, on the other hand, are nearly 5 logits apart and are expected to agree only rarely. Expected agreements for individual raters depend not only on their own severity, but also on the severity of the raters they were paired with.

The validity of the Rasch measures requires that rater severity be consistent in a manner that fits the probabilistic expectations of the Rasch model. Sizeable discrepancies between observed and expected agreements are an indication that raters are not consistently severe or lenient. This is also expressed in the infit and outfit statistics shown in Column 2. Rogers, for example, has an infit mean square of 1.61, indicating that his behavior does not fit the Rasch model very well. With a severity measure of -1.43, he is overall a lenient rater but frequently awards scores more or less severe than expected. This is confirmed in Column 3 where we see that his observed agreements are much less than expected. Leigh has an infit mean square of .50 and more observed agreements than expected. This is an indication that she is not behaving like an independent expert. She may be avoiding disagreement by keeping her scores in a safe middle range, or she may be colluding on her assessments with another rater. Comparing the total exact agreements to the total expected agreements, as shown at the bottom of Table 1, provides a global assessment of rater fit. In this example, observed agreements were less than expected, indicating that raters generally acted independently, but were not always consistent in severity.

To summarize, rater agreement is viewed differently in CTT and RMT. In CTT, disagreement amongst raters is seen as a source of error that must be minimized. Inter-rater reliability indices reflect the CTT approach and are generally not reported by Rasch programs such as Facets. In RMT, rater disagreement is seen as an indication that raters are behaving independently, bringing multiple perspectives and differing expertise to their assessments. In RMT, rater disagreement is used to estimate rater severity, which can then be employed to adjust performance measures for the detrimental effects of rater independence. For adjustments to be valid, however, it must be verified that raters are consistent in severity in a manner that conforms to the expectations of the Rasch model. Rater fit statistics and observed to expected agreement ratios can be used to assess rater fit. If the data reasonably approximate the Rasch model, the reliability reported in the examinee measurement report is the nearest equivalent, in purpose and substance, to inter-rater reliability as employed in CTT.

References

Brown, J. D. (1996). *Testing in language programs.* Saddle River, NJ: Prentice Hall Regents.

Brown, J. D. (2012). Statistics Corner: How do we calculate rater/coder agreement and Cohen's Kappa? *Shiken Research Bulletin, 16*(2), 30-36. Retrieved Dec 2, 2013 from http://teval.jalt.org/sites/teval.jalt.org/files/SRB-16-2-Brown-StatCorner.pdf