Rasch Measurement in Language Education Part 7:

Judging plans and disjoint subsets

James Sick
International Christian University, Tokyo

Previous installments of this series have provided an overview of Rasch measurement theory, reviewed the differences among the various Rasch models, and discussed the assumptions and requirements that underlie Rasch measurement theory (RMT). In this installment, I will address a practical problem that can occur when using many-facet Rasch analysis (MFRA). MFRA is often used to adjust for differences in rater severity or other factors when measures are constructed from subjective judgments. Readers unfamiliar with MFRA and the differences among the Rasch family of models might wish to review Part 3 in this series.

Question:

My institution recently held a student speech contest with 9 teachers serving as volunteer judges. The 51 student participants were assigned to 3 rooms where a three-judge panel rated each speech for content, language, and presentation. When all speeches were completed, the scores were compiled and the three highest scoring students received a prize.

Now that the contest has finished, I am analyzing the results with MFRA with the aim of improving the judging process in future contests. When I run the analysis using Facets (Linacre, 2012a), it runs but returns the message "warning – there may be 3 disjoint subsets." Could you explain what this means and what, if anything, I should do about it?

Answer:

With some follow-up communication, we determined that the 3 judges assigned to each room did not rotate. That is, three judges rated all speeches in room 1, three different judges rated all speeches in room 2, and yet another panel of 3 judges rated the speeches in room 3. These are the 3 disjoint subsets. We can estimate the relative severity of judges within rooms by examining how they rated the same speeches. However, we cannot make similar comparisons with judges in other rooms because they rated no speeches in common.

We also confirmed that the mean scores awarded in each room differed: room 3 had a mean substantially lower than rooms 1 and 2. Was this because the speeches delivered in room 3 were of lower quality? Possibly. However, it is equally feasible that the speeches in that room were as good as the others, but the judging panel was more severe in how they interpreted and applied the judging criteria. Perhaps the three panels calibrated their scores independently at the start of the sessions. Alternatively, perhaps one judge on panel 3 had substantially more demanding standards, bringing down the average score for that room. In fact, a casual inspection of the raw scores indicated that one judge in room 3 awarded fewer points in total than any of the other 8 judges, lending support to that possibility. At any rate, because the lower scores in room 3 could feasibly be due to either judge differences or speech differences, we cannot fairly compare speech scores across rooms.

The MFRA that you conducted with Facets has used the raw scores from all performances to construct a single, logit-delineated scale, and placed both judges and participants along it. Logit measures for participants indicate the quality of their speeches. Logit measures for judges indicate their severity in applying the rating scale. The participant measures have been automatically adjusted for judge severity

by adding or subtracting an amount equal to the average severity of the judges who provided the scores. Unfortunately, in your analysis this accomplishes very little. Because all participants in a room were scored by the same three judges, the severity adjustment in any particular room will be the same for all. Moreover, differences in severity among judges in different rooms, even though Facets has estimated them, are not dependable. Facets employs a procedure called maximum likelihood estimation to locate the combination of rater and participant measures that is consistent with the data and best fits the Rasch model. However, this "best fit" solution is neither predictable nor transparent when there are disjoint subsets. The final estimate could be attributing room differences to performances, to judges, or to any additive combination of the two.

Judging Plans

Because the contest is finished and your goal is to improve judging in future speech contests, let us consider some possibilities. First of all, you could simply treat the three rooms as separate contests and award prizes to the top performers in each room. However, if the best speeches of the day happen to take place in the same room, speakers in the "strong room" would be at a disadvantage. A better approach would be to create a judging plan that rotates judges through the rooms as the contest progresses. This would link all judges, eliminate the disjoint subsets, and allow you to create a fair and dependable scale that applies to all participants independent of their room assignment.

Table 1 shows how such a judging plan would work. After 6 speeches, a short break is called and three judges rotate to other rooms. After another 6 speeches, a second set of judges rotate. With this simple plan, 6 pairs of judges would rate 11 speeches in common, 3 pairs would rate 6 speeches in common, and 3 pairs would have no common ratings but would be indirectly linked via two other judges. This would be sufficient to eliminate the disjoint subsets and create a common rating scale applicable to all rooms.

Table 1. Simple judging plan for speech contest

Session	Room 1 (Judges)	Room 2 (Judges)	Room 3 (Judges)
1 (speeches 1-6)	1 2 3	4 5 6	7 8 9
2 (speeches 7-12)	1 2 9	4 5 3	7 8 6
3 (speeches 13-17)	1 8 9	4 2 3	7 5 6

Group-anchoring

Another approach to dealing with disjoint subsets is to employ group-anchoring. Group-anchoring allows us to specify which measurement facet, in this case speakers or judges, will be considered the source of any variance between the subsets. For example, before starting the estimation process we specify that the mean speech performance measure for each room will be set to zero logits. Estimates of judge severity and individual performance within a room are then calibrated in relation to that benchmark. In effect, this forces the mean performance measures for each room to be equal, adjusting severity measures to compensate. Conversely, we could specify that the mean severity for each judging panel be set to zero. This would put faith in the judges and attribute group differences in performance measures to lower quality speeches.

Although group-anchoring may appear arbitrary, we can usually build a case that it is preferable to anchor one facet rather than the other if there are disjoint subsets. The following are some issues to consider when specifying group-anchoring:

1. *Sample Size.* Group-anchoring can take sample size into account. In the speech contest, the speakers outnumber the judges, so speech performance means are less likely to be affected by sampling error. With only three judges per room, a single strict judge, assigned by chance, can substantially skew the group mean. With 17 speakers per room, it would require about 6 weak speakers to similarly skew the mean. While 17 is hardly a robust sample, it is certainly better than 3.
2. *Incorporating additional information.* There is often anecdotal or other extraneous information to support anchoring one facet over the other. In the speech contest, one judge appeared to be quite tough based on the raw scores awarded. Apart from the speech contest, is he or she known to be a tough grader? No information was provided, but were speakers assigned to rooms randomly, or were room assignments related to classes, departments, levels, or other factors that might affect speech performances? If there are reasons to believe a priori that students in one room were of lower proficiency, one could argue for anchoring the judges rather than the speakers.
3. *Transparency.* Group-anchoring, even when wrong, creates transparency. For example, if we elect to group-anchor the speakers, we can add a caveat such as "assuming that the speeches delivered in each room were of equal quality on average, speakers 5 and 9 in room 1 and speaker 3 in room 2 delivered the best speeches of the day." Because the Rasch estimates which Facets provides are ambiguous when there are disjoint subsets, it can be advantageous to designate an hypothesized source of variance and then state the limitation. In addition, consider that the raw scores used to award prizes in the speech contest were essentially anchoring the judges. With no adjustment for judge severity, raw score comparisons assume that differences between rooms are the result of lower quality speeches. By not specifying group-anchoring, we might be accepting a default assumption that we would reject if it were made transparent.

Group-anchoring by design

In your speech contest, group-anchoring could be used, with some reservation, as a post hoc repair to compensate for a flawed design. There are instances, however, where group-anchoring can be advantageously and validly employed as part of an a priori design. To extend the discussion, let us consider the following example from an English speaking test that I helped administer several years ago.

Approximately 120 students were divided into groups of 4 and randomly assigned one of three topics for a 10-minute discussion test. Discussions were observed by two teacher-raters who did not participate in the discussion. Group assignments were quasi-random, mixing students from two or more classes, and raters were rotated frequently. The three topics were based on themes from the textbook and were known to students in advance. Topics were assigned at the start of the discussion by drawing them one by one from a canister until all had been used. This insured that the distribution of topics was equal across students and raters.

A problem with this design was that each student discussed only one topic, creating 3 disjoint subsets of unlinked topics. Consequently, it was not possible to unambiguously determine whether the topics were of equal difficulty. To exacerbate matters, there was a widespread belief among both students and teachers that Topic 3 was more challenging due to vocabulary and cognitive demands and would disadvantage students to whom it was assigned.

From a measurement perspective, the ideal solution would have been to have each student discuss two topics, preferably in different groups with members who had discussed different topics previously. This would have eliminated the disjoint subsets and allowed us to estimate the degree to which topic assignment affected performance scores. Two discussions, however, would have required an extra class period

to administer. Pedagogically, it was questionable whether the lost teaching time could be justified by minor improvements in testing accuracy.

In this context, a strong argument can be made for the validity of anchoring the students and attributing differences between subsets to topic difficulty:

1. The sample of student participants is robust. With approximately 40 randomly assigned students attempting each topic, it is reasonable to expect the mean speaking ability of each subset to closely approximate the overall mean.
2. There is an a priori prediction that Topic 3 is more difficult. If this is verified when anchoring students, it bolsters the argument that variation due to topic difficulty is the true source of any mean differences in the subset measures.

In addition, if the topic assignments are shown to affect scores, maximizing adjustments to offset this enhances face validity. Students will feel the test is more fair if they know that differences in topic difficulty are recognized and compensated for. Even though group anchoring is a compromise from a measurement perspective, the pedagogical benefits of reducing test administration time justify the cost.

Figure 1 is a Facets vertical ruler from an early administration of this test. The figure shows the relative measures of the four facets—students, judges, topics, and categories—relative to a single, logit-delineated scale along the left. As was predicted, Topic 3 is slightly more difficult than Topics 2 and 3. In comparison to the variation in rater severity and student ability, however, variations in topic and category difficulty are quite small. In practice, we found only two cases where an adjustment for topic difficulty might have been large enough to alter a student's final grade. Nevertheless, knowing that their discussion test scores took into account both the severity of the raters and the topic they were assigned seemed to increase student confidence in the overall fairness of the test. Logistically, this test would have been difficult to administer without relying on group anchoring.

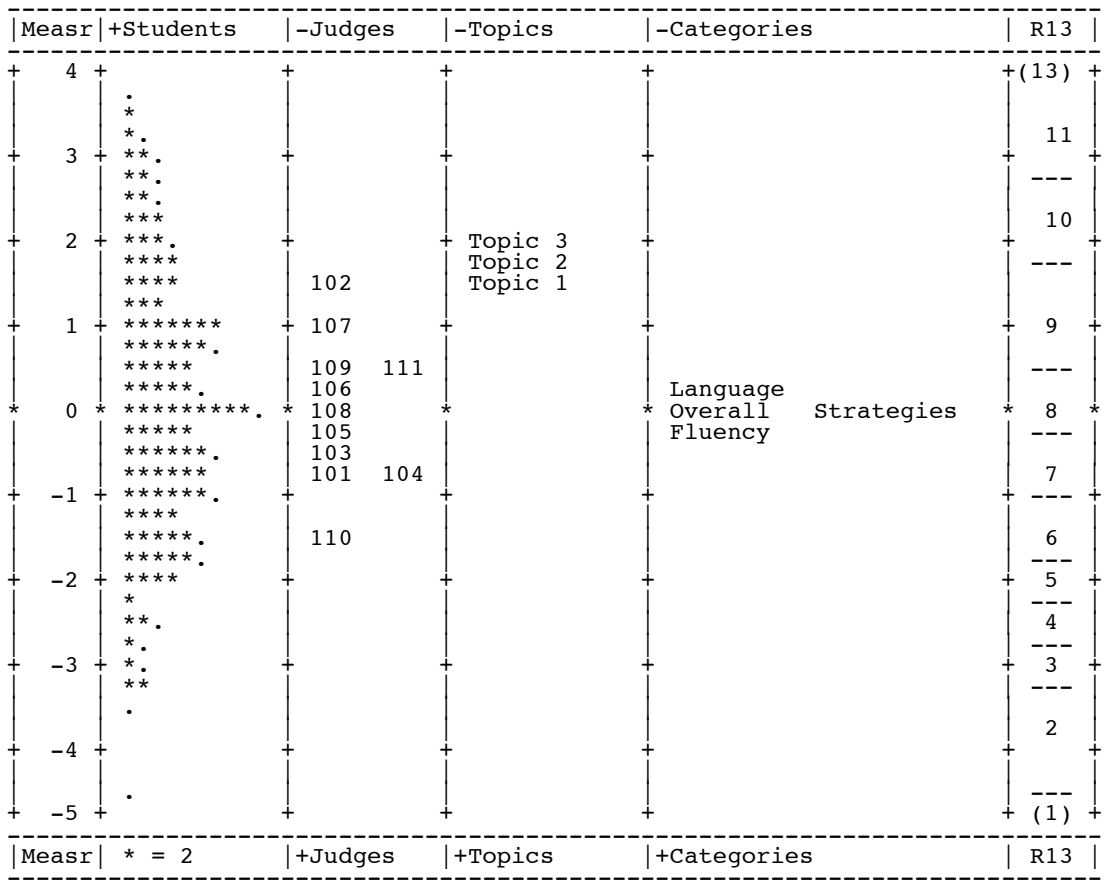

Figure 1. **Vertical ruler for a group discussion test**

More information about judging plans can be found in Linacre (1997). Details of how to specify group anchoring can be found in the Facets manual (Linacre, 2012c). An excellent tutorial on judging plans, disjoint subsets, and group-anchoring is also available at the Winsteps and Facets website (Linacre, 2012b).

References

Linacre, J. M. (1997). Judging plans and Facets. *MESA Research Note #3.*

Linacre, J. M. (2012a). Facets (Version 3.70) [Computer Software]. Chicago: Winsteps.com.

Linacre, J. M. (2012b). *Many-facet Rasch measurement: Facets tutorial #4.* Retrieved from http://www.winsteps.com/a/ftutorial4.pdf

Linacre, J. M. (2012c). *A user's guide to Facets: Program Manual 3.70.* Retrieved from http://www.winsteps.com/a/facets-manual.pdf