# Modeling Rater Effects in a Formative Mathematics Alignment Study

An integrated assessment system considers the alignment of both summative and formative assessments with academic content standards. Although generally viewed within the realm of summative statewide accountability assessments, the alignment of formative assessments is just as critical. A considerable body of research has been devoted to studying alignment within the realm of large-scale summative assessments, both the methods for doing so and the alignment of the tests themselves (see Webb, 1999). Far less research, however, has been devoted to the alignment of formative assessments. Like summative assessments, the validity of the inferences made from formative assessment results (e.g., intervention and instructional decisions) depend in part upon the degree to which items align with the given standards. In both cases, misaligned items may have serious consequences for the validity of test-based inferences.

In alignment studies raters are typically asked to gauge the degree of alignment between items and the targeted content standards. Analyzing data from individuals making qualitative judgments poses certain challenges. Both the *between* and *within* rater variance should be considered, as well as the overall leniency/severity of a given rater. Erratic ratings, or systematically lenient/severe ratings, could threaten the validity of inferences based on the alignment results. For example, in the same way that inferences based on a student's writing ability are threatened by the tendencies of the individual judging the writing sample, so too are the alignment results for any one item. If the student (or item) is "unlucky" and is judged by severe or erratic raters, the resulting score may not adequately reflect the underlying trait.

In this study, we apply a many-facets Rasch model (MFRM) to explicitly model and control for rater effects. MFRM was developed by Linacre (1989) as an extension of the basic Rasch model to parameterize other facets of the measurement process (Bejar, Williamson, &

Mislevy, 2006). We apply the model to data from an alignment study between a formative middle school math assessment and the Common Core State Standards (CCSS), while parameterizing a “rater” facet. We then use the results to examine both the degree of item alignment and how the results may have differed had other approaches been used.

## Methods

Fifteen participants (raters) were recruited for this study. All participants were either teachers or district-level math coaches, and were selected based on (a) mathematical content knowledge, and (b) knowledge of the CCSS. Approximately 450 items in each of grades 6-8 were used, which were randomly selected for review from a larger pool of items. Each item was subsequently randomly assigned to one of five 90-item sets per grade. Each rater was assigned 3 sets to review for a total of 270 items. Raters were stratified across items sets so there were always 3 raters per set. Additionally, all raters were linked across sets, as shown in Table 1, so that all ratings could be placed on the same scale during the MFRM analysis. During the review, items were paired and displayed with intended standards using a secure online distribution tool, and raters were asked to gauge the item’s standard alignment based on a 4-point (0-3) ordinal scale.

When data follow an ordinal scale, the MFRM can extend either Masters’ (1982) partial credit model – when each item follows its own unique scale – or Andrich’s (1978) rating scale model, when item thresholds can be assumed equivalent across items. Because all items were rated on the same 4-point ordinal scale there was little theoretical reason to presume that thresholds would vary dramatically across items. Further, for many items there were unobserved categories along the 4-point scale, and because the rating scale model is robust against empty categories (Linacre, 2000), we decided to apply the MFRM extension of Andrich's rating scale model, defined as

$$ln\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_k \qquad (1)$$

where $P_{nijk}$ is the probability that item $n$ is rated into category $k$ on latent trait $i$ by rater $j$. Generally, the $B_n$ term represents the estimated ability of examinee *n* while $D_i$ represents the difficulty of item *i*. When applying the model to alignment data where the object of measurement is test items and not individuals, the terms need to be slightly redefined, although estimation is equivalent. In the current study the $B_n$ term instead represents the *item's* level on the latent trait – "alignment" – while the $D_i$ term represents the raters' overall willingness to endorse an item as aligned (i.e., the "difficulty" of endorsing an item as aligned). The $Cj$ and $F_k$ terms are identical regardless of the object of measurement, representing the severity of rater *j* and the Rasch-Andrich threshold for the $k - 1$ category, respectively, where $F_k$ is estimated once and fixed across all items.

The MFRM analysis allowed us to compute an adjusted alignment rating for each item, controlling for rater severity. The adjusted rating is, in a sense, an estimate of what the rating on the item would have been had it been rated by a judge with the average leniency/severity. Rater severity was included as the sole facet in the MFRM analysis in a fully crossed design, and was conducted with FACETS, version 3.70 (Linacre, 2012). Following the MFRM analysis, all items were tabulated into *aligned* (rating > 2) and *not aligned* (rating < 1) categories based on the adjusted MFRM rating. The results of the MFRM analyses were then compared to the results of other alignment methodologies.

## Results

Overall, 1,180/1,345, or 87.73% of all items had an adjusted MFRM rating above 2.0, suggesting the item was aligned to its corresponding standard. A summary of rater effects is displayed in Table 2. Of particular note is the severity statistic, which is reported on the logit scale, and differs substantially amongst raters. An examination of the average rating column indicates that the most severe Rater 3 scored items, on average, nearly a full category below the most lenient Rater 11. These differences in rater severity can be seen visually in Figure 1. The left most column of the figure displays logit values from the analysis, mapped vertically. To the right of these values is the distribution of rater severities, mapped against the logit values. To the right of the rater labels are the distribution of items, mapped against the same logit scale. Finally, the furthest right column of the figure displays the raw rating scale ranging from 0-3. The figure thus displays how the distribution of rater severities compared to the distribution of item endorsabilities on a common scale.

The fit statistics in Table 2 are also important given that they provide an indication of the consistency with which raters made decisions. That is, given the severity of the rater, as determined by his or her scoring of all other items, and the endorsability of the item, does he or she consistently rate items as would be expected. The outfit statistic is perhaps most useful for this purpose, with values above 1.0 indicating underfit to the model (i.e., unexpected ratings) and values below 1.0 indicating overfit to the model (i.e., overly consistent ratings, as in a rater who judges all items to be perfectly aligned). These fit statistics provide an indication of the intra-rater reliability.

Overall, the MFRM results produced more conservative estimates of alignment than alternatives. For example, while the MFRM estimated 165 items to be not aligned with their corresponding standards, use of an unadjusted average would have resulted in only 133 items rated as not aligned. Similarly, if a consensus approach had been used only 100 items would have not met the alignment criteria. However, inspection on an item-by-item basis reveals that, while the MFRM adjusted the ratings of many items from *aligned* to *not aligned*, there were also items that were adjusted from *not aligned* to *aligned*. This shifting of alignment was due solely to rater severity, which was unaccounted for in both the consensus and unadjusted average techniques.

## Discussion

This proposal represents, to our knowledge, the first application of MFRM to alignment data. While the results suggest more conservative estimates of alignment relative to alternative methods, it is likely that the results more accurately represent the "true" alignment of the items, given that threats of systematic rater variance were minimized. For the full paper, a more complete account of the procedures for alignment will be detailed, as well as a more in depth discussion of the inter- and intra-rater reliability. We will also provide further discussion around the latent "alignment" trait, and why we viewed the definition as invariant across grades 6-8 for formative math items. Further, the differences between formative and summative alignment will be reviewed and discussed. For example, while the general processes used to gauge alignment with summative assessments have been largely successful (see Webb, 1999), the intended use of formative assessments differ dramatically. Alignment, then, may need to be re-conceptualized to match this differing use. We plan to discuss our approach to this quandary in further detail. Finally, the full paper will provide a more in depth account of the results and implications as well as limitations and directions for future research.

# References

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573. doi: 10.1007/BF02293814

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49-82). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2000). Comparing "partial credit" and "rating scale" models. *Rasch Measurement Transactions.* Retrieved July 5, 2012, from http://www.rasch.org/rmt/rmt143k.htm

Linacre, J. M. (2012). Facets computer program for man-facet Rasch measurement (Version 3.70.0). Beaverton, Oregon: Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Pychometrika, 47*(2), 149-174. doi: 10.1007/BF02296272

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.

Table 1

Teacher Sampling Plan

| Grade 6 | | | | | Grade 7 | | | | | Grade 8 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 |
| a | a | | | | | | | | | | | | | a |
| | b | b | b | | | | | | | | | | | |
| | | | c | c | c | | | | | | | | | |
| | | | | | d | d | d | | | | | | | |
| | | | | | | | e | e | e | | | | | |
| | | | | | | | | | f | f | f | | | |
| | | | | | | | | | | | g | g | g | |
| h | | | | | | | | | | | | | h | h |
| i | i | i | | | | | | | | | | | | |
| | | j | j | j | | | | | | | | | | |
| | | | | k | k | k | | | | | | | | |
| | | | | | | l | l | l | | | | | | |
| | | | | | | | | m | m | m | | | | |
| | | | | | | | | | | n | n | n | | |
| | | | | | | | | | | | | o | o | o |

*Note.* Each letter, a – o, represents an individual rater. Each rater is further represented by one and only one row. The table displays the overlap of raters across items so that each item is rated by at least 3 teachers, while the raters themselves link across item sets, allowing all raters to be calibrated on the same scale during the MFRM analysis. Each set contained 90 items.

Table 2

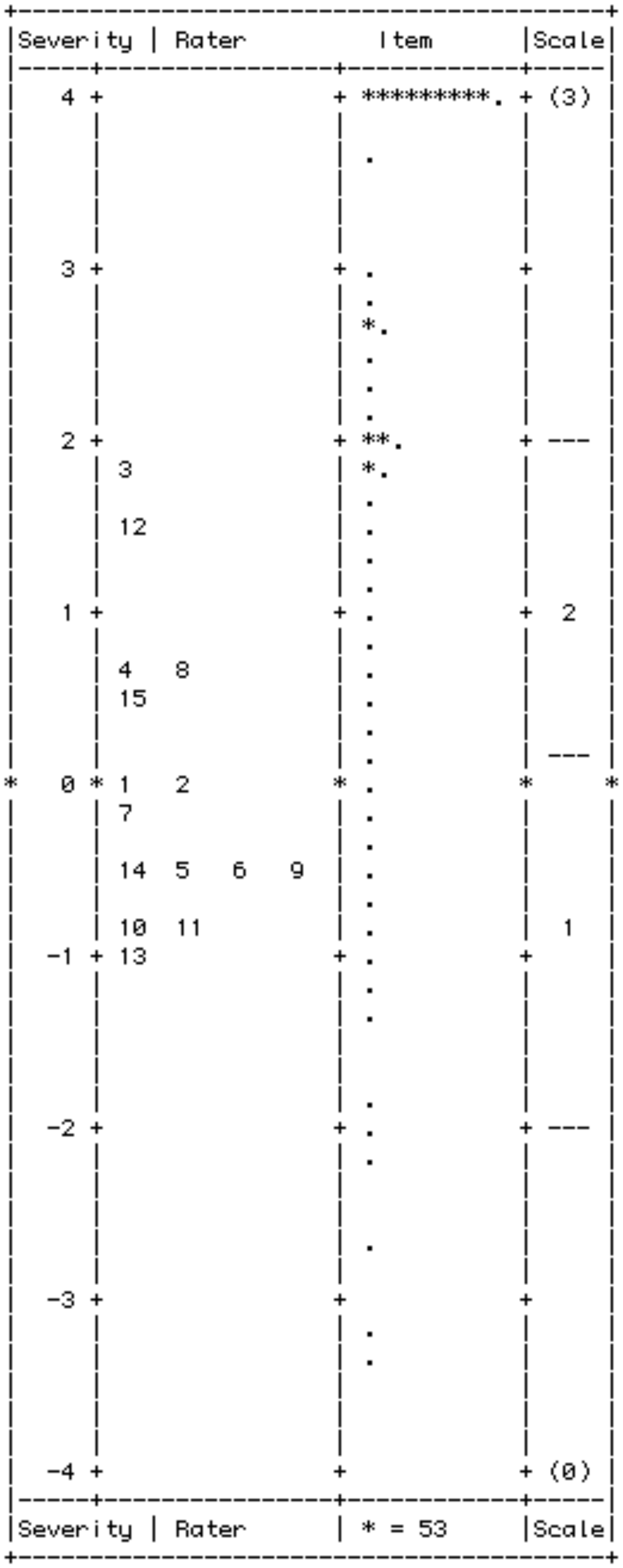
Summary of Rater Effects

| | | | | | | Fit Statistics | | | |
|---|---|---|---|---|---|---|---|---|---|
| Rater | Score Tot | Count | Avg. Rtg | Severity | S.E. | Infit | Infit Z | Outfit | Outfit Z |
| 1 | 728 | 268 | 2.52 | -0.01 | 0.14 | 1.25 | 1.83 | 0.97 | -0.15 |
| 2 | 683 | 270 | 2.19 | -0.07 | 0.12 | 0.96 | -0.34 | 1.04 | 0.34 |
| 3 | 537 | 269 | 1.71 | 1.78 | 0.09 | 0.75 | -3.2 | 0.76 | -2.98 |
| 4 | 690 | 268 | 2.3 | 0.64 | 0.11 | 0.91 | -0.88 | 0.87 | -1.21 |
| 5 | 713 | 270 | 2.36 | -0.44 | 0.13 | 1.2 | 1.57 | 1.07 | 0.56 |
| 6 | 688 | 269 | 2.23 | -0.49 | 0.12 | 1.28 | 2.38 | 1.16 | 1.31 |
| 7 | 660 | 269 | 2.06 | -0.23 | 0.11 | 1.19 | 1.84 | 1.16 | 1.48 |
| 8 | 638 | 270 | 1.79 | 0.66 | 0.12 | 0.99 | -0.02 | 0.96 | -0.35 |
| 9 | 717 | 269 | 2.44 | -0.54 | 0.13 | 0.98 | -0.16 | 0.92 | -0.48 |
| 10 | 736 | 269 | 2.53 | -0.8 | 0.15 | 0.78 | -1.71 | 0.92 | -0.41 |
| 11 | 750 | 269 | 2.68 | -0.92 | 0.15 | 0.87 | -0.84 | 0.99 | 0.04 |
| 12 | 586 | 269 | 1.88 | 1.49 | 0.1 | 0.76 | -2.97 | 0.8 | -2.36 |
| 13 | 728 | 269 | 2.47 | -1.00 | 0.14 | 1.34 | 2.43 | 1.12 | 0.82 |
| 14 | 746 | 268 | 2.67 | -0.55 | 0.15 | 1.22 | 1.49 | 0.89 | -0.55 |
| 15 | 692 | 269 | 2.28 | 0.45 | 0.12 | 0.97 | -0.21 | 0.91 | -0.79 |

*Note.* Severity reported on logit scale, with higher values indicating a more severe rater.

Figure 1

Item-Rater Map



```
+-----------------------------------------+
|Severity | Rater         Item      |Scale|
|-----+----------------+-----------+-----|
|   4 +                + *********. + (3) |
|     |                |            |     |
|     |                | .          |     |
|     |                |            |     |
|     |                |            |     |
|     |                |            |     |
|   3 +                + .          +     |
|     |                | .          |     |
|     |                | *.         |     |
|     |                | .          |     |
|     |                | .          |     |
|     |                | .          |     |
|   2 +                + **.        + --- |
|     | 3              | *.         |     |
|     |                | .          |     |
|     | 12             | .          |     |
|     |                | .          |     |
|     |                | .          |     |
|   1 +                + .          +  2  |
|     |                | .          |     |
|     | 4   8          | .          |     |
|     | 15             | .          |     |
|     |                | .          |     |
|     |                | .          | --- |
*   0 * 1   2          * .          *     *
|     | 7              | .          |     |
|     |                | .          |     |
|     | 14  5   6   9  | .          |     |
|     |                | .          |     |
|     | 10  11         | .          |  1  |
|  -1 + 13             + .          +     |
|     |                | .          |     |
|     |                | .          |     |
|     |                |            |     |
|     |                |            |     |
|     |                | .          |     |
|  -2 +                + .          + --- |
|     |                | .          |     |
|     |                |            |     |
|     |                |            |     |
|     |                | .          |     |
|     |                |            |     |
|  -3 +                +            +     |
|     |                | .          |     |
|     |                | .          |     |
|     |                |            |     |
|     |                |            |     |
|     |                |            |     |
|  -4 +                +            + (0) |
|-----+----------------+-----------+-----|
|Severity | Rater      | * = 53     |Scale|
+-----------------------------------------+
```