RTD Extended Response Rangefinding Procedure (Multi-Trait Rubric)

By: Marjorie Wine
Alexander Hoffman

The RTD (Rigorous Test Development) project is an attempt to build a professionalized content development practice that focuses on individual item quality, particularly by leaning into the importance of validity throughout the content development process. It assumes that content development professionals develop professional judgment that can be raised, honed and calibrated by providing frameworks and clarifying expectations in ways that account for the constraints and demands of typical practice within test development, today. RTD is a conscious and deliberate attempt to respond to the disparity in status, training and shared knowledgebases between psychometrically oriented professionals and content development professionals.

Extended Response Rangefinding (Multi-Trait Rubric)

The purpose of rangefinding is to determine the range of responses that satisfy the rubric-based criteria for each score point. Further, though rangefinding, responses' true scores may be determined, so that Training Sets may be built and used to prepare scorers to score consistently and accurately.

Materials

- Preliminarily ordered stack of test-taker responses, prepared by scoring vendor.
- Copies of scoring guides, Rubrics and Items for each committee member.
- Scoring record sheets for facilitator (and committee members).

Procedure:

I. Provide overview of process and purpose for committee members.

II. Score responses using the first Rubric Trait (may take a whole day, or more)

A. Review the Trait criteria established in the Rubric

1. Review and discuss the first Dimension
   a. Read aloud the description of the Dimension (i.e. the first bullet) for each scorepoint, starting with the maximum score.
   b. Lead a discussion of the differences between demonstrated performance of that Dimension between each adjacent at each scorepoint and across the entire range.

   Note: You are the expert and the committee should not impose their own differing interpretations upon you or the group.
2. Repeat the Dimension review (i.e. II.A.1) for each remaining Dimension, one at a time. ↩

B. Score the responses for the first Item for the first Trait:

1. Have the committee read the entire stimulus silently.
2. Read the prompt aloud to the committee.
3. Select the first guidepost response
   a. Select a potential guidepost response for the lowest scorepoint from the stack of responses.
   b. Lead committee discussion of this response's score on just this Trait. If there is no consensus on the score, repeat with another potential lowest scorepoint guidepost response.

   Note: If you and/or the committee reach a consensus, but feel(s) that the response is too close to scorepoint boundary, repeat with another potential guidepost response.
4. Repeat guidepost selection (II.B.3) for each remaining scorepoint, one at a time. ↩

Extended Response Rangefinding (Multi-Trait Rubric)

Procedure (continued):

5. Score the first 20 responses:
   a. Have the committee members score the next 20 responses on Trait 1, individually.
   b. Add each member's scores to master scoring sheet.
   c. Lead discussion of any responses that lack consensus scores (i.e. more than 1 person disagrees with the rest of the group).
   d. Record the "true score" (i.e. consensus score) for each response on your score sheet.

   Note: Responses should be marked "DNU" (i.e. do not use for scorer training) if
   - the committee cannot reach a consensus
   - the responses containing something indicating a threat to someone and/or the test-taker his/herself (i.e. Alert Responses).
   - the response satisfies one of the condition codes listed in the rubric.
   - the content might especially appeal to or unsettle scorers and therefore introduce opportunity for bias (e.g. passionate expression of political partisanship, deep exploration of personal religious views, etc.).
   - the response is otherwise anomalous.

6. Repeat the scoring process (i.e. II.B.5) for each remaining set of 20 responses. ↩

III. Repeat the Trait scoring process (i.e. II) for the each remaining Trait, one at a time. (Do not move on to the next Trait until all responses for all Items have been reviewed for the previous Traits.)

Training Sets

Anchor Set: a set of 10 responses reflecting the full range of score points, containing the clearest exemplars of low, typical and high responses at each score point. Each response is annotated to explain how it satisfies the criteria enumerated in the rubric for its assigned score point. Anchor sets are used to calibrate scorers at the onset of scorer training and are used by scorers as a reference throughout live scoring.

Practice Sets: two sets of 10 responses reflecting the full range of score points, containing the widest variety of responses available in the rangefinding stack. Responses selected for the practice sets should represent particular challenges to the scoring criteria (e.g., they are near score point boundaries, they satisfy scoring criteria across multiple score points, etc.). Practice sets expose scorers a wide range of approaches that test-takers may take in answering the prompt and are used by scorers as a reference throughout live scoring.

Qualification Sets: two sets of 10 responses reflecting the full range of score points, each containing typical responses that represent fairly clear cut scoring decisions. At the end of scorer training, scorers must score a pre-determined percentage of the responses in the qualifying sets the same way as the rangefinding committee did in order to qualify for live scoring. They are also used by scorers as a reference throughout live scoring.

Validity Set: a set of 10 responses reflecting the full range of score points, containing typical responses that represent fairly clear cut scoring decisions. The validity set is administered to scorers midway through live scoring to help scoring directors monitor scorer accuracy and prevent scoring drift.

Extended Response Rangefinding (Multi-Trait Rubric)

Calibration Questions:

When the committee appears to be embarking on an inconsistency, ask the committee compare response being discussed with a similar response previously discussed.

How does Response 28, to which we gave a score of 3, compare to Response 14?

Is scoring Response 36 as a 2 consistent with the decision we made on Response 21?

When the committee is unsure of a response's score, ask the committee to explicitly address (i.e. score) each bullet, and then come to an overall score for that Trait.

How does this response score on each bullet? How do we want to reconcile that mixture of bullet scores?

When then committee is having trouble reaching consensus, ask an individual member from each group to present his/her entire line of reasoning for his/her score.

Lynn, can you please explain to all of us, from the beginning, why you think this response is a 4 for this Trait?

Issues to Monitor:

Equal weighting of each bullet in a Trait: monitor whether the committee is stressing one bullet disproportionately to the other bullets (e.g. excellent use of prior content knowledge on a social studies response is not alone sufficient for a response to earn a 3 or a 4).

I am concerned that we are not paying enough attention to how well-organized these longer papers are. We need to be careful not pay so much attention to idea development that we are not properly weighing all the bullets. Some of these longer papers are really poorly organized.

Calibration of individual committee members: monitor whether the any one committee member is consistently in disagreement with the rest of the group.

I have heard your concern, however we are going to proceed as the rest of the committee has determined. If you want to talk about thus further, let's talk during the next break.

Rubric-based scoring: monitor whether committee members are using their own criteria for scoring, rather than the criteria in the Rubric.

I understand that those are certainly qualities of good writing that one would generally look for, however, we are trying to measure very specific skills on this test. The Rubric is designed to focus scorers on those specific skills.

Construct isolation: monitor whether committee members are relying upon the Trait being discussed or on other Traits.

You're right. The grammar is realty weak in this response. However, we have another whole Trait with which we will scores the grammar and conventions. Right now, we need to focus on idea development and organizational structure.

Extended Response Rangefinding (Multi-Trait Rubric)

Consistency: monitor whether the committee is scoring the current prompt as it did previous prompts.

How did we deal with this issue on the first prompt we looked at? Can we make sure that we are using the same standard across all the prompts?