RTD Short Answer Rangefinding Procedure

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.

ALEDEV CONSULTING

©2018
Copies and derivative works must retain AleDev attribution

Short Answer Rangefinding Procedure

The purpose of rangefinding is to determine the range of responses that satisfy the scoring-guide-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 Provisional Scoring Guides and Items for each committee member.
- Scoring record sheets for facilitator (and committee members).
- Copies of indicators

Procedure:

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

II. Score the first item:

A. Review Scoring Guide

1. Have the committee read the item content and Provisional Scoring Guide silently.
2. Lead a discussion of the differences between score points as listed in the Provisional Scoring Guide.

   Note: You are the expert and the committee should not impose their own differing interpretations upon you or the group.
3. Lead a discussion of potential Scoring Guide amendments.

B. Select Guidepost responses

1. Select a potential guidepost response for the lowest scorepoint from the stack of responses.
2. Lead committee discussion of this response's score. 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.

C. Repeat guidepost selection (II.B) for each remaining scorepoint, one at a time. ↩

D. Score the first 20 responses of the first item:

1. Have the committee members read and score the next 20 responses, individually.
2. Add each member's scores to master scoring sheet.
3. Lead discussion of any responses that lack consensus scores (i.e. more than 1 person disagrees with the rest of the group).
4. Table any responses that the committee believes the scoring guide is inadequate for scoring (i.e. those which demonstrate in a manner not anticipated in the scoring guide partial or full mastery of the KSA in the indicator and/or correctly answers the item).

   Note: Push back against tabling if you suspect it is just a matter of scoring difficulty.
5. Record the "true score" (i.e. consensus score) for each response on your score sheet.

AleDev Consulting ©2018
Copies and derivative works must retain AleDev attribution

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.
- 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).
- the response is otherwise anomalous.

E. Repeat the scoring process (i.e. II.D) for each remaining set of 20 responses. ↩

F. Group the tabled responses by issue and review each group to ensure you understand how the committee believes the scoring guide is inadequate.

G. Record recommendation for changes to the scoring guide, as appropriate.

III. Repeat the item scoring process (i.e. II) for the each remaining item. ↩

IV. Scoring Guide amendments for the first item (after rangefinding committee disbands):

A. Review the suggested edits and amendments to the scoring guide.

B. Evaluate which recommendations are valid (i.e. both provide additional guidance to scorers *and* address common responses to the item).

C. Incorporate changes to the Scoring Guide

VII. Repeat Scoring Guide revision process (i.e. IV) for each remaining items. ↩

Training Sets

Anchor Set: a set of 8-12 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 8-12 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 8-12 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 8-12 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.

©2018
Copies and derivative works must retain AleDev attribution

Short Answer Rangefinding Procedure

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 score point in the scoring guide.

How does this response score in relation to each scorepoint described in the scoring guide?

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 2?

©2018
Copies and derivative works must retain AleDev attribution

Short Answer Rangefinding Procedure

Issues to Monitor:

Correctness of the scoring guide: monitor whether the committee regularly wanting to award one or more points for a common type of response that is not explicitly described in the scoring guide.

> *I am observing that we really seem to believe that this alternative configuration really should get a point. Do we need to add this type of response to the scoring guide as an alternative method for getting a 1?*

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.*

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

> *I understand that those are certainly qualities of a response that one would generally look for, however, we are trying to measure very specific skills on this test. The scoring guide is designed to focus scorers on those specific skills and particular aspects of the responses.*

Construct isolation: monitor whether committee members are relying upon skill being demonstrated in the responses or on other skills irrelevant to the designated indicator.

> *You're right. The grammar is realty weak in this response. However, we really need to focus on how well they are able to design a controlled experiment.*

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?*

©2018
Copies and derivative works must retain AleDev attribution