Rigorous Item Feedback

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.

Table of Contents

Because item (and test) development is a highly collaborative process, feedback is a critical part of CDPs' (content development professionals) work. There are many kinds of feedback, as we explain in our *RTD Item Feedback Typology*, here we address the elements of that must be included when a particular kind of feedback – one we call *Rigorous Item Feedback*. That is, the four elements of item feedback on deep and substantive issues that make it clear, understandable, evaluable and perhaps even actionable.

Because *valid items elicit evidence of the targeted cognition for the range of typical test takers*, Rigorous Item Feedback must be aimed at improving the quality of the evidence that items elicit for some group(s) of test takers. It should focus on reducing:

- False positives/Type I errors (i.e., when test takers respond correctly to items despite lacking proficiency with the targeted cognition).
- False negatives/Type II errors (i.e., when test takers respond incorrectly to items despite possessing sufficient proficiency with the targeted cognition.

Decisions on which feedback to act on, how to reconcile different feedback and exactly what to do remain in the hands of the CDPs, but RIF (Rigorous Item Feedback) can come from many different kinds of reviewers and many different kinds of reviews.

Outside the Requirements of Rigorous Item Feedback

There *are* types of feedback that are not bound to the four elements of Rigorous Item Feedback.

- Pointing out actual content errors in items does not require RIF's four elements.
- Problems with *item hygiene* can simply be highlighted without RIF's four elements. These are generally issues in an item's form and presentation arise from sloppiness in building the item and should be caught before they are presented to review panels. (See *RTD Item Hygiene*.)
- CDPs simply have no control over the standards and are bound to assess them. That is simply the job. However, sometimes some standards may seem objectionable (e.g., as not sufficiently important, as inappropriate for this level of cognitive development). While CDPs cannot alter or ignore standards, there may be some value in hearing objections to the standards (i.e., to refine their understandings of the domain) and in unburdening oneself. While such objections should be thoughtful, because they are not about individual items, RIF's four elements simply do not apply.
- There is often a bit of interpretive work to be done with standards, and this shapes how standards appear in items. Concerns in this area are not item-specific concerns and therefore do not require RIF's four elements. However, interpretation of standards is very delicate and context-specific and therefore such feedback should be given quite carefully (see below).

We strongly agree that assessment should *not* limit instruction – neither in terms of the content itself or the pedagogy used to teach it. We believe that the standards themselves should be a floor or foundation upon which content and pedagogy is built, but that they should not limit or constrain *instruction*. A full education and a rewarding educational experience is more than just the standards – and certainly more than a series of isolated standards. On the other hand, the *standards act as a ceiling for assessment*. If assessment goes beyond the standards, teachers and students in different classroom, schools and districts may be evaluated on things without notice. Assessment's limited opportunities to examine test takers requires a kind of construct isolation that would be detrimental if enforced in instruction.

In fact, in order to be useful for instructional (and other) purposes, assessment must have this different relationship to the standards. These two different relationships to standards (i.e., as a floor for instruction and as a *ceiling for assessment*) necessarily lead to different interpretations of individual standards. Instruction builds on the standards and often interprets standards in the context of other standards. Assessment's limitations require such a different kind of focus that it cannot rely upon the kinds of interpretations that drive high quality instruction.

Four Elements

Rigorous Item Feedback contains the following four (overlapping) elements:

- Who — Which group of test takers are at risk?
- Where — Where is the item (or stimulus) does the problem appear?
- How — How will the test taker's cognitive path be inappropriately disrupted?
- Which — Which KSAs (knowledge, skills and/or abilities) are implicated?

Who?

Different test takers adopt different strategies and approaches to items and they bring different experiences and understandings with them to items. Other than actual content mistakes in items, no one should expect that *all* test takers will be effected by any particular problem in an item. Rigorous Item Feedback should show the care and thought that went into it by making clear what test takers the reviewer believes might be effected. Otherwise, it is difficult for a CDP to see that the reviewer is not merely unmindfully projecting their own confusion onto test takers.

Where?

Rigorous Item Feedback must identify exactly where in the item the problem appears. That is, it should name the specific aspect or feature(s) of the item will prompt the anticipated problem for test takers. Reviewer should identify the specific word(s) or phrase(s), the part(s) of the figure(s) or any other spot in the item or stimulus that they

think will lead some test takers astray or otherwise trigger a false positive or false negative outcome.

The problem in the item or stimulus may be one of omissions, rather the of commission. That is, the reviewer may point to an spot or element that is missing something – a word, an explanation or something else.

How?

As tests of cognition, problems in items or stimuli necessarily produce bumps or issues in test takers' cognitive paths through the item. That is, somewhere in the series of thoughts, realizations, decisions, use of knowledge and/or applications of skills, something problematic is evoked. Rigorous Item Feedback identifies how effected test takers' cognitive paths are disrupted, precisely where along those paths and what problems inappropriately result from that disruption – and may even prevent test takers from reaching *any* answer.

This is often the most difficult RIF element for reviewers, as it requires them to consciously think through the cognitive path of test takers who may be very different from themselves. It requires recognizing that most people's cognitive paths through items are not quite orderly and do not resemble a computer responding to a linear series of instruction. It also calls for some degree of what we call *Radical Empathy* (see our *What is Radical Empathy?* white paper) to imagine another's cognitive path in sufficient detail to understand how it can be affected – as opposed to handwaving that this issue will interfere with test takers *somehow*.

Which?

Problems in item necessarily inappropriately tap KSAs that test takers cannot be expected to possess. They may inappropriately assume certain background knowledge, inappropriately require test takers to use techniques that are beyond their grade level, expect level of mastery that is unrealistic, require test takers to the test developers' minds or any number of other problematic expectations. Rigorous Item Feedback must be explicit and clear about precisely which KSAs items require of test takers that they should not.

Overlap

The four elements of Rigorous Item Feedback are *not* entirely distinct, and instead often overlap. For example, many item issues impact test takers who would following a particular strategy in responding to an item, thus identifying strategy both identifies the group (i.e., *who*) and almost identifies cognitive bump (i.e. *how*). Identifying and explaining the cognitive bump (i.e., *how*) quite often lays out which KSAs the item inappropriately requires (i.e., *which*).

When Rigorous Item Feedback is given for items that support an alternative task (i.e., a path to the correct answer that does *not* rely appropriately on the Targeted

Cognition) elements will certainly overlap. For example, test takers who possess a particular KSA (i.e., both *who* and *which*) might be able to respond to the item by following [this] path (i.e., a longer *how*). The elements in the item that support the alterative path (i.e., *where*) may not be as small as with risks of false negative responses.

The fact that these elements often overlap does *not* mean that they do not each need to be recognize and articulated clearly. The overlap should not mask a lack of clarity with any of them. That is, they should each be completely explained to ensure that the reviewer is fully expressing their concern and to give the audience (e.g., a CDP) the best chance to understand it.

CDP Use of Rigorous Item Feedback

When content development professionals give feedback to each other on items, they should be explicit and clear exactly what problem they see, including all four of the elements of Rigorous Item Feedback. This kind of effort is a product of CDPs' necessary mutual investments in each other's professional growth, in addition to their commitment to improving item quality. It is a complement to take the time to give a colleague careful feedback. CDPs should be prepared to do this, themselves.

First, reviewing their feedback themselves for these four elements forces them to make sure that they understand their own concerns and have thought them all the way through. It is simply a matter of respect for one's colleagues to try to avoid giving erroneous negative feedback.

Second, providing Rigorous Item Feedback gives the recipient the best chance to understand exactly the nature of the objection or concern. It is also a matter of respect for one's colleagues to try to be as clear as possible when giving critical feedback.

Third, this format for feedback give the listener the best opportunity to *evaluate* the feedback. The CDP can examine the logic and reasoning for themselves and deliberate on whether the issues cited are significant enough to merit altering the item – or even if the objection is actually accurate. For example, it also allows the CDP to consider whether the implicated KSAs are, in fact, part of the standards and therefore entirely appropriate to require. This last use is especially important when a test developer includes reviews from CDPs who are *not* expert in the content area (e.g., disability specialists), more junior CDPs or cross-content CDPs.

Review Panel Use of Rigorous Item Feedback

Although panelists who serve on review panels obviously cannot be expected to be proficient with sharing Rigorous Item Feedback, collection of such feedback should be the aspirational goal of those who facilitate review panels. Moving review in the direction of Rigorous Item Feedback relies on a two-prong strategy.

First, review panelist should receive some minimal training in the elements of Rigorous Item Feedback. It can be offered as a framework for explaining feedback and for group discussions among the panelists about the concerns raised (e.g., see *Rigorous Item*

Feedback handout at the end of this packet). Panelists can be asked to consider these four elements/questions when thinking about their concerns and can be invited to support each other by helping to refine or expand objections of their fellow panelists by discussing each element of the concern.

Second, review panel facilitators can model this kind of support themselves by asking questions that draw out each element, that ask panelists to clarify an element or even to share their own thoughts regarding an element of one of their fellow panelist's feedback. This framework can provide facilitators with anchor points that they can use to connect feedback across item, to invite people into the discussion and otherwise deepen discussions across the days of the review panel. This framework can also help to focus the panel on sharing and explaining their concerns, and move them away from offering the kinds of advice and fixes that CDPs easily recognize as being unwise for one reason or another. Facilitators can explain that they understand the feedback and all four of its elements, and then move the group on the next objection or the next item.

Using this framework to record review panel feedback is immensely valuable when CDPs must later decide how to respond to it. Review panel feedback can be *particularly* useful because panelists can bring personal experiences, personal perspectives and experiences with a range of potential test takers that the internal content development team lacks. However, because they are not full-time professional CDPs, they often misunderstand how items function and certainly are not in position to be mindful of all of the concerns that CDPs have to balance and/or address. If the feedback is recorded with this framework, CDPs can more quickly recognize the nature of the concern and evaluate its scope and its alignment with the relevant standard. Furthermore, use of this framework makes it easier to postpone evaluation of the feedback until later, making it more practical to use less experienced CDPs and/or CDPs out of their content areas to facilitate review panel committees and/or record their feedback.

Rigor

We refer to the idea of *rigor* quite a bit when we think about, discuss, write about and explain content development practices. While we acknowledge doing is an intentional act of provocation, it is done to send a very deliberate message. Standardized tests should *not* be judged by their difficulty, which is how the term "rigor" is so often intended in this context. Tests should be as difficult as the standards call for, no more/no less. Instead, *the practices and procedures of test development* should be rigorous. That is, they should be demanding on content development professionals and other who work with them *in the right ways.* In our view, that means that test development practices and procedures should challenge and support all who are involved to do a better job of producing *items that elicit evidence of the targeted cognition for the range of typical test takers.*

This kind of rigor is fully on display in the Rigorous Item Feedback framework. It requires those giving feedback to express their concerns by specifying which test takers, the nature of the suspect cognition (i.e., both in the cognitive path and in terms of the

implicated KSAs) and where in the item the issue appears. This framework focuses reviewers on issues that impact items' ability to avoid false positive and false negative responses.

Yes, sticking to the this framework *is* demanding. Some will chafe at the limits it puts on what they can object to and/or the shift from wordsmithing and advise giving. There is a real learning curve to thinking in terms of these four elements. It requires *rigor* from those involved in the content development process.

That is why we call it *Rigorous Item Feedback*.

Rigorous Item Feedback

Rigorous Item Feedback has four elements. There are other types of feedback (e.g., item hygiene, content errors) that do not have these same elements. Ideally, feedback about how test takers will understand and make sense of items will include each of these elements.

Who
Which group of test takers are at risk? • Are they identifiable by their demographics? • Are they identifiable by the strategies they adopt? • Are they identifiable by their command of content? • Are they identifiable by their abilities/ disabilities?

Where
Where is the item (or stimulus) does the problem appear? • Is there a problematic or inadequate word or phrase? • Is there a spot where information or direction is missing? • Is there a particular problem with a graphic? • Where in the item or stimulus is a problematic idea prompted?

How
How will the test taker's cognitive path be inappropriately disrupted? • Where in their cognitive path might they get inappropriately distracted? • Where in their cognitive path might they get inappropriately led astray? • What new cognitive path will result from this disruption?

Which
Which KSAs (knowledge, skills and/or abilities) are implicated? • Which specific knowledge, skills and/or abilities does the item require that it should not? • Which specific KSAs must be misapplied to get to the 'correct' answer? • Which specific KSAs give some test takers an inappropriate advantage?

Feedback on the Standards
Feedback on the standards is usually not about test takers' understanding of an individual item. Though it often addressed important issues, it is not about a particular item. It is often focused on issues that are beyond CDPs' authority. • This standard/the targeted cognition should not be taught at this grade. • This standard/the targeted cognition is not taught at this grade. Interpretation of the core of the standard (e.g., whether an item addresses the most important part(s) of the standard) is actually an issue to be addressed on the level of the standards, rather than of the individual item.
The objection that *an item does not address a standard* can be made by laying out the four RIF elements to describe the cognitive path that leads to a correct response, who will take that path and what in the item prompts them to do so. This likely will include KSAs that let them identify this alternative task.

ALEDEV CONSULTING