Rigorous Test Development (RTD): Theory of the Item

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.

RTD Theory of the Item

We say (and write) that *valid test items elicit evidence of the targeted cognition* – and we stand by that. But that is a much simplified distillation of the *RTD Theory of the Item.* In this chapter, we explain more deeply what we think makes for valid items and explain what is really going on. Our theory of the item drives almost everything we think about, in terms of test development practices – certainly everything we preach. It is the connective tissue of Rigorous Test Development. Because we cannot read test takers' minds to be sure of what they know and can do, *we must use items to prompt tasks to generate work product that contains evidence of the targeted cognition* – as we explain in this chapter..

As is so often the case, when diving deeper and thinking theoretically, jargon can be very helpful. The goal of that jargon is to differentiate between ideas or issues that conventionally are usually not carefully considered as distinct idea. In this chapter, we rely on the terms *test taker* (TT), *item*, *targeted cognition* (TC) and *task*. We add two new concepts, *intended task* (IT) and *alternative task* (AT).

Test Taker (TT): The person taking the cognitive (e.g., educational, professional certification, psychological) test.

Item: What is presented on the page or screen for test takers (TT) respond to. What a lay person might refer to as the test question, though it can include instructions, potential answers and other features.

Targeted Cognition (TC): The knowledge, skill or ability that the item is aimed at assessing mastery of, in the test taker (TT)

Task: What test takers (TT) actually *do*.. The cognition, cognitive processes and cognitive steps that TTs engage in to respond to their understanding of an item.

Intended Task (IT): The cognition, cognitive processes and cognitive steps that content development professionals *intend* test takers (TTs) to engage in, upon reading an item. The IT appropriately relies upon the Targeted Cognition (TC), while also including other cognition.

Alternative Task (AT): Cognition, cognitive processes and cognitive steps that a test taker TT engages in *instead of the intended task*, upon reading an item. An AT might or might not rely on the Targeted Cognition (TC), and might not rely on it appropriately.

Other Cognition (OC): Cognition that a test taker engages in when responding to an item, in addition to the cognition they use to complete their understanding of the task. OC is almost invariably distracting and impairing of their ability to engage in whatever task they have gone to.

Work Product (WP). Whatever the test taker (TT) produces in response to an item, be it the selection of an answer option, or some constructed response.

Idealized Story of the Item

We begin explaining the RTD Theory of the Item with an idealized story of everything going smoothly.

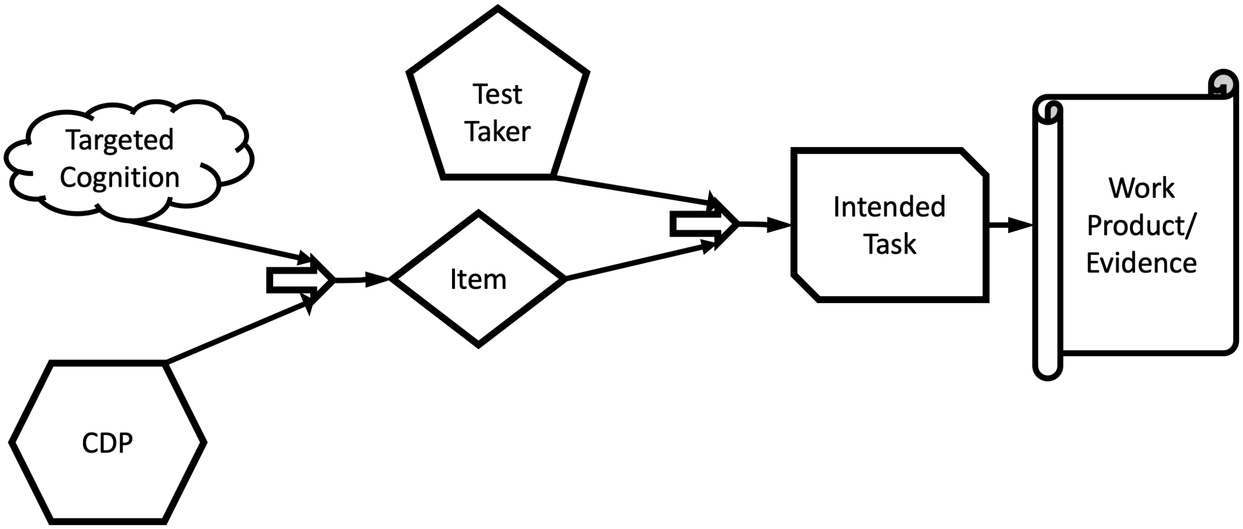

Figure 1. Idealized View of the Item

Content development professionals (CDPs) start working with some targeted cognition (TC) in mind. This TC comes from the standards, assessment targets or some other organization of the content domain. They declare to themselves, "This item will be aligned with this TC, and it will elicit evidence of that TC."

The CDPs produce such an item, and it is presented to test takers. TTs read the item and correctly understand the task that it is prompting them to engage in – an intended task that is built around and depends on the targeted cognition (TC). Because of the skill and expertise of the CDPs, completion of item the produces some work product that contains evidence of the TC.

In those two paragraphs, a lot is going on that we usually take for granted. There are no bumps. But reality is messier than that.

What Went Right

First, for our Theory of the Item, we do not question the appropriateness of the targeted cognition (TC). We accept that it was identified well and is appropriate for the test. That is a question for test design, and we focus more on test production. In this story, the TC is well selected.

Second, CDPs' understanding of the TC was deep and substantial enough that they could design a task that actually depended on that TC in the right ways, featuring the TC without letting other KSAs or cognition muddle the item's alignment.

Third, CDPs understood test takers (TT) well enough to write the item (i.e., the presentation on the page or screen) in a way that actually prompts TTs to engage in the intended task (IT), rather than some other related or similar task. That is, the TTs would engage in a task that is built on and depends on the TC, rather than some alternative task that might otherwise suit, and they will not be distracted by anything else.

Fourth, TTs' work product (WP) is not just a result of engaging in such a task, but actually includes, contains or constitutes evidence of that *intended* task (IT), rather than an alternative task, because only such a WP would contain or constitute evidence of the targeted cognition (TC).

That is a lot to go right. CDPs must have great skill and expertise to do that. Unfortunately, reality makes that much harder than it appears in this little story.

A Story of It Going Wrong

Some time ago, one of us was complaining to the other about how hard it is write well-targeted mathematical calculation MC items, these days. We do not to need to point fingers, but *she* was rightly making very careful observations about test takers and contemporary mathematics curriculum and pedagogy. The other – and we still do not need to point fingers – heard her complaints and while *he* granted the difficulty, insisted that it was possible. He granted that it might not be productive enough, but it was *possible*.

The next day, while he trying to figure out something else, he somehow had a flash of insight about how to do it. He could create two-digit multiplications items with a sufficient number of high quality distractors (see the RTD Rule for Distractors in chapter [x]) that the item would be perfectly aligned and valid. He started with the most mundane two-digit multiplication problem that he could think of: 45 x 67. He wrote it up, quickly writing out the rationales for three distractors – all he needed! But he realized that he had more, too. Six distractors. Eventually, he had seven – SEVEN, ah ah ah! – high quality distractors.

He looked at them to be sure. Did each of them obey all of the requirements of the RTD Rule for Distractors? Was each the product of a specific misunderstanding or misapplication of the targeted cognitions? Oh, yeah. Certainly. No doubt. Here are all the ways that a second or third graders could misunderstand the algorithm for two-digit multiplication. He hubristically thought his write up was so good that he said it was practically an item template. Here is the actual quote from the chat transcript, “OK. Not a template. I have too many plausible distractors to be a template!” Yeah, he said that.

Sixteen minutes later, she replied, “I don't think your distractors meet the criteria.”

That prompted a phone call.

She explained to him that his distractors would not catch the kinds of mistakes that students *today* might make because they do not simply use the old algorithm – the way that we were all taught in school to multiply multiple digit numbers in school in the 1970’s and 1980’s – to do that kind of problem. They are taught tricks of grouping and borrowing and reshaping the problem, to say nothing of integrating estimation into problem-solving (i.e., her original concern).

He did not realize that test takers would *not* (necessarily) engage in *his* targeted cognition or intended task to reach the answer. They might not just use the old algorithm, and in fact were not *that* likely to do so. His distractors were not appropriate to the tasks that so many test takers *would* engage in in response to his simple item, and therefore would not catch their mistakes. And because they would not catch their mistakes, when they made them they would not have an appropriate distractor available to them. This lack of an answer option that matched their mistake would cue them that they made an error and thereby prompt them to find it an correct it.

He did not understand the construct – in this century – well enough to even understand the problem she described. He did not understand test takers well enough to understand what task they would *actually* engage in. And so, he did not supply distractors to catch the kinds of mistakes might emerge from that.

(Most importantly, he did not listen closely enough to his partner to really understand what she was complaining about. It turns out that old ideas about the importance of assessing strict mathematical calculation skills are running up against improved mathematical instruction that integrates estimation skills as a basic part of solving arithmetic problems. Because of this improved instruction and greater abilities on the part of students, it might no longer be possible to assess strict calculation – separate and apart from more recently emphasized estimation skills.)

Understanding that the Task is Central

When we say or write, "Valid test items elicit evidence of the targeted cognition," we are leaving out the most important part. *How do they do that?* If we want CDPs to do that, what should they be focusing on? As we allude to at the beginning of this chapter, they should think hard about their items' *intended tasks* and potential *alternative tasks* that might supplant them.

Content development professionals (CDPs) write, create, build, edit, refine and develop items. Tests are made up of items. We say that RTD is item-centric because items are what CDPs produce and are what test takers (TTs) are faced with and respond to. However, items are *not* central to our Theory of the Item.

Because the kinds of tests that we are concerned with are all about various kinds of cognition (i.e. thinking), cognition must be at the center of our work. And that cognition is found when the test taker engages in the work of responding to an item, in the cognition, cognitive steps and cognitive processes that take them from a question to an answer, result or work product.

We call that work, all that thinking, all those steps, a *task*.

CDPs want to create items that prompt test takers to engage in tasks, but not just *any* tasks. Rather, they want test takers to engage in tasks that appropriately depend on the targeted cognition (be it a content standard, assessment target or some other KSA). Therefore, that targeted cognition must be a critical element of the activity that makes up the task – the most critical element, in fact.

Unfortunately, items do not automatically and/or magically prompt the intended task and prompt test takers to engag in the cognition, cognitive steps and cognitive processes that CDPs intended. Especially poorly written items might tend to prompt something else entirely. Poorly considered items might prompt the intended tasks for some test takers but consistently prompt alternative tasks in others, or even other things *in addition* to the intended task.

Understanding the item requires understanding all the ways that this can go wrong.

Now, this this view of the *task* is a bit different than our normal view. Often, we elide the distinction between what an item charges the test taker with (i.e., produce a response to this item) and the actual cognition of the test taker. However, even then – when we are more focused on the work of CDPs – we think that it is critical to pay attention to the cognitive task. For example, asking test takers to select the correct answer to a multiplication question is simply a different task than asking them to supply it in a constructed response item. In all contexts, we think that understanding the task is critical, but in this context of the RTD Theory of the Item, our thinking about tasks is more layered and nuanced.

In this context, as we try to understand how items *really* work, we adopt a more test-taker-focused view of the task. We must think very carefully about intended tasks and consider what else the item might prompt in the test taker.

Understanding the Test Taker

It is the test taker who reads the item and translates the item into some cognitive task to engage in. It is the test taker who produces the work product, and therefore the evidence of the targeted cognition. This is why one of RTD's core principles is, *Test development requires mindfulness of the test-taker's perspective.*

This understanding requires mindfulness of many things.

The first thing that CDPs – and everyone else – must understand is that test takers are different from them. For example, CDPs are almost invariably more expert in the content domain than test takers, and have more subtle, nuanced – and often far more accurate – understanding of the content domain than test takers. Test takers' more novice emerging understanding of the content domain can lead them to understand items differently and engage in different sorts of tasks than experts might, were they to be encounter the same item.

Second, CDPs must be mindful of the fact that test takers are *not* all the same. They come to items with different backgrounds, different training, different experiences and with different senses of selves. Thus, an item might prompt one task in some test takers and prompt another in a different group of test takers – or even prompt additional cognition. Some test takers might know a special trick to solve that kind of problem. Some test takers might find that a story or example in the item (or stimulus) leads them to thoughts that distract them from the intended task, even as they actually attempt the intended task. Some test takers may have more experience with the item's presentation or charge, and others may have less. There is seemingly no ends to these differences that a CDP might consider.

Third, CDPs must be mindful that test takers are…well, test takers. They are likely in a highly attentive state, but also a stressed state. As they likely have encountered other items before this particular item, their state maybe impacted by those earlier items. That is to say, the test experience itself can influence the test taker as they approach the next item.

In short, there is no singular test taker, or even a typical test taker. Rather, one might consider some theoretical group of different typical test taker*s*. Items must be developed with an array of different typical sorts of test takers in mind.

Understanding the Role of the Targeted Cognition in the Task

RTD is a stickler for making sure that the targeted cognition is not merely a *part* of the task, but is actually be most important part. We have seen countless items in which the targeted cognition was required to produce the correct answer, but it was merely one of many KSAs, and was clearly not the most difficult step. We look at those items and declare that they are *not* aligned with their supposed targets and therefore *cannot* provide evidence that supports valid inferences about test takers mastery of the targeted cognition.

When other KSAs are even *equally* demanding elements of a task, lack of sufficient mastery of those other KSAs can prevent a test taker from demonstrating the targeted cognition. If test takers *might* fail to produce evidence of mastery of the targeted cognition for any reason *other* than lack of mastery of the targeted cognition, we simply cannot ever know whether the failure was because of the targeted cognition or because of *the* other KSAs. Thus, we could never have solid evidence of shortfalls in the targeted cognition. If items allow test takers do engage in an alternative task that does *not* even depend on the targeted cognition to arrive at the correct answer or work product, then even the value of seemingly affirmative evidence will always be questionable. Items and their tasks, therefore, must *depend* upon the targeted cognition.

This is *not* to say that the targeted cognition – which could very well be comprised of a collection of KSAs – must be the only cognition that a task includes. That would be ridiculous! For example, nearly every test requires test takers to have either to be able to use a writing implement or a computer. Also, most mathematics tests have non-trivial amounts of reading on them. Furthermore, most interesting tasks are built of many KSAs.

Rather, these other KSAs should be much less demanding than the targeted cognition. High school mathematics tests should *not* require a high school reading level, or else we could not tell whether test takers got items wrong because of poor math ability or because of poor reading ability. Other KSAs should only be included if they are at levels whose mastery can be *safely* assumed in a test taker who can perform the targeted cognition.

The targeted cognition must be the key step of the task, the distinguishing step, the most important step. Exactly what this means will look different in different content domains and even for different potential targeted cognition. Regardless, CDPs should keep in mind that mere relevance or inclusion of the targeted cognition in the intended task is simply not enough.

Understanding the Importance of Clear Evidence

Above, we began to address the quality of the evidence than an item might elicit, particularly when a task might not appropriately depend on the targeted cognition. This issue, however, is more complicated than that.

As a classroom teacher, we loved assignments that required students to integrate multiple skills, apply a broad range of knowledge and tap a range of abilities in order to produce work that felt real and engaging. One of the great advantages of teaching lessons based on literature is the ability to use the art as a launching off point for a wide array of KSAs and other lessons (e.g., critical thinking, communication, what it means to be human, the nature of society). Though we used rubrics, we created room for students to show what they *could* do, and not just what they couldn't. Unfortunately, this experience and instinct is entirely inappropriate to assessment. You see, despite the rubrics, there was sufficient subjectivity and room for demonstrations of additional ability that it was not always clear our students had the mastered the KSAs that unit was intended to target. It certainly was not always in the grade we recorded for the assignments.

When test takers' work products are more than selection of a correct answer, the quality of the evidence in the work product becomes more complicated to examine. Items must be *very* clear regarding what is expected of test takers and scoring guides must be carefully aligned with those targeted constructs. Test takers must be prompted (by items) to generate work products that contain clear evidence of the targeted cognition and that evidence must be recognized when it is present and recognized when it is absent.

When evidence is unclear, inferences based on the item are weak. When inferences based on the item are weak, inferences based on the test are weak. And when inferences based on the test are weak, the test is not valid.

Consider an item that prompts a test taker to produce some sort of work product, but it is not clear about whether a particular element is *required*. Can the absence of that element in the work product be taken as evidence that the test taker lacks mastery of the cognition associated with that element? The absence of evidence does not tell us whether the test taker has mastery of that cognition, or not. It suggests that they might not have it, but perhaps they were lazy, or rushing, tired, or just did not see the need to include it. That suggestion, thus, is rather weak evidence to use against the test taker when calculating their score.

Understanding the Artificiality of Tests

Tests – be they fancy formalized standardized tests or smaller less formal classroom tests – are intrinsically artificial and/or heightened situations. Test takers know they are being assessed and that alters their state. Some may be nervous and some may focus more intently. Sometimes, a test might be administered in a unusual physical location or under unusual conditions.

The importance of the targeted construct to items and their intended tasks discussed above is one key aspect of the artificiality of tests. Authentic application of knowledge, skills and abilities are not so constrained. Authentic challenges are often amenable to multiple solutions, relying on different potential collections of tools. Authentic application of KSAs does not required just one of them to be dominant. The basic need of tests to know what they are assessing makes them inauthentic.

Furthermore, test takers are primed when they are taking tests to show off the KSAs that they know the test is targeting. They might have studied and/or might have been told what to look out for, explicitly. Almost invariably, they are *trying* to focus and give their best performance.

Is the demonstration of the KSAs that make up the content domain and the use of the targeted cognition in response to these artificial items *really* the same thing as authentic proficiency with those KSAs in the real world, in more authentic contexts? Obviously, it is not *exactly* the same thing. We simply cannot know how much a difference it makes. But we know that it makes *some* difference, and that causes us to be aware that whatever we are measuring is not quite what we would *like* to be measuring.

This puts a ceiling on the quality of any evidence that tests, their items and their associated tasks elicit. All of that evidence is just a *little* bit off, at best. And because the evidence is always at least a little bit off, the inferences that are made based upon it are a little bit weaker. CDPs and everyone else involved in testing – including the audience for test scores and reports – should know that.

However, we do *not* interpret that to mean that we should not bother or try. Rather, for us, it just underscores the importance of CDPs doing everything they can to ensure the highest quality items they can. Sure, the ceiling might be a bit lower than we want, but we can still work to get as close to that ceiling as possible.

The RTD Theory of the Item

Putting all together the RTD Theory of the Item looks like this:

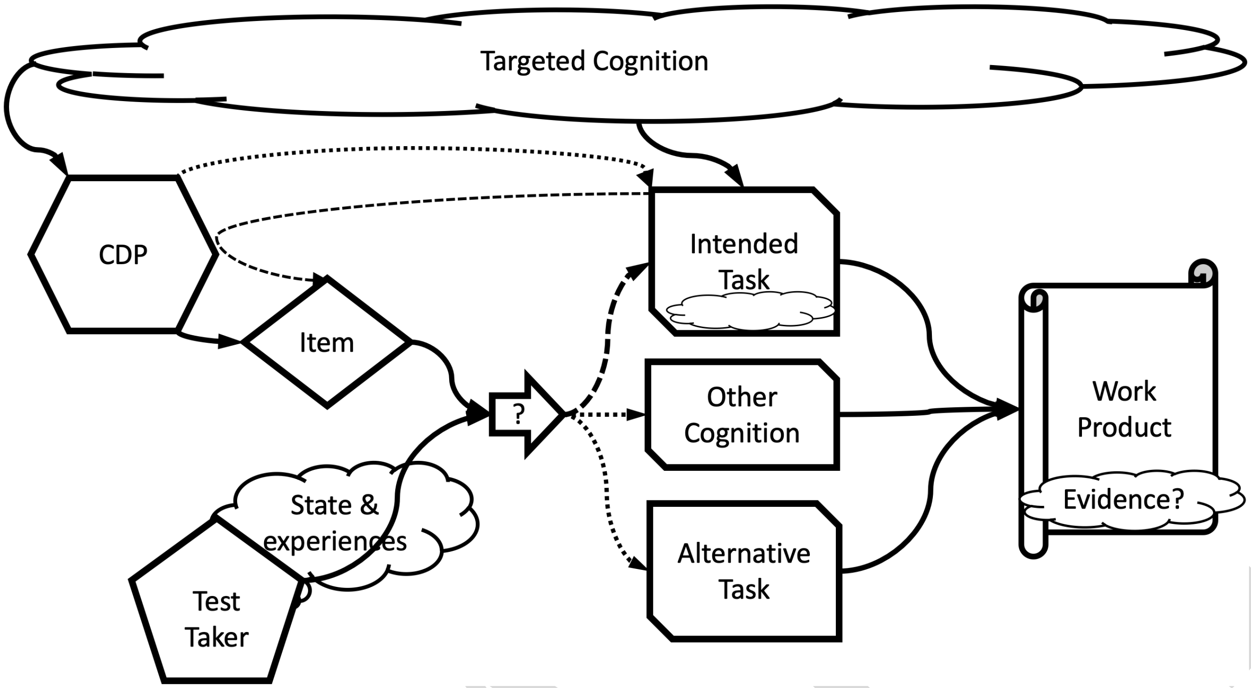

Figure 2. RTD Theory of the Item

- The content development professional takes some targeted cognition (TC) and tries to envision an intended task that appropriately depends on that TC and will produce evidence of that that TC.
- They create a task calculated to prompt test takers to engage in the intended task.
- Test takers respond to the item by either engaging in the intended task or in some alternative task, with the possibility of engaging in some other cognition in addition to either task, and produce some work product.
- That work product may or may not contain evidence of the TC, and its absence will be taken as evidence of the lack of sufficient mastery of the TC.

Understanding the Demands on CDPs

Not surprisingly, the RTD Theory of the Item highlights the demands on content development professionals required to develop valid items (i.e., items that elicit evidence of the targeted cognition).

First, CDPs must understand the targeted cognition well enough to understand how it might be used in a variety of potential tasks. This is one element of the kind of content knowledge that they must have.

Second, CDPs must be able to envision an intended task that appropriately depends on the targeted cognition. Yes, this is a kind of content knowledge, but it assessment-specific content knowledge, because envisioning tasks with that kind of dependency is simply not needed in other contexts. It also requires a kind of creativity that is surprisingly difficulty – and therefore specialized.

Third, CDPs must be able to envision an intended task that will produce a work product that contains high quality evidence of test takers' level of mastery of the targeted cognition. We term this content-specific assessment knowledge, because understanding this idea of evidence is a general CDP skill, but actually is a little different for each content domain.

Fourth, CDPs must be able to see items (and tasks) through the view of test takers. They must understand test takers well enough – the range of typical test takers and perhaps a range of less typical test takers – to be able to design an item that will actually prompt test takers to engage in the intended task (rather than some alternative task), without any distracting or inappropriate other cognition. This is neither content knowledge nor assessment knowledge. Rather, it requires understanding how the minds of a range of takers of *this* test actually work. The RTD view is that this is the *most* critical skill to do CDP work well. We have some ideas about what sorts of backgrounds might make it more likely, but none that assure us of finding it consistently.

Fortunately, CDPs do not work in isolation, responsible for all of this on their own. Content development work is done in teams, with many different people contributing in different ways. No one should be expected to produce valid items from scratch all by themselves. That would simply be too much to ask of any one person.

Alternative Views of the Item

We know of two widely subscribed to views of the item that we know are wrong, and fail to recognize what content development work is really about.

The dominant view among [the public?] is that test takers come to a test with their KSAs (i.e., a sort of toolbox), and demonstrate those KSAs as instructed in items. This view allows that some marginal issues might distract the test taker, but generally the idea of taking a test is pretty damn straightforward. *Just answer the questions*. Unfortunately, this view utterly fails to acknowledge the complexity of test takers and the challenge of eliciting high quality evidence of particular cognition without being able to read test takers' mind. It does not have room for ambiguity or miscommunication and it fails to grapple with the subtleties of content domains.

The second alternative view of the item is the dominant view within the assessment industry because it is the view of most psychometricians. In this view, items are black boxes that produce points for some test takers and not for others. In this view, items are to be understood by patterns of those points, without any reflection on content, tasks or cognition. Items may also be examined by looking at patterns of those points relative to patterns in externally determinable patterns in test takers (i.e., demographics). They are not content people, and focus more-or-less entirely on the data produced from field and operational testing, and their various analyses based on that data.

Obviously, we think that both of these views are, *when thinking about the validity of items, tests and the inferences made upon them*, utter garbage. They are insufficient because they include no effort to consider how items work or how they might lead to the kind of evidence that would justify the inferences we make based on test taker performance. Items should be understood as prompting cognition and understanding how – or whether – that has worked is the first part of how we should understand them. Items should thereby elicit evidence of that cognition, and how – or whether – that works is the second part. Any view of items that is not entirely grounded in test taker cognition simply cannot have anything to offer regarding the *validity* of a cognitive test.

Implications of this Theory of the Item for RTD

Virtually everything in RTD is connected to this view of the item. Everything.

This Theory of the Item explains the role of every person involved in content development. It explains the need for Fairness, Sensitivity and/or Bias committees.

This Theory or Item explains why content development work is hard and why some people have such trouble with it. It explains many of the skills and understandings that should be looked for when hiring CDPs.

This Theory of the Item explains what makes for valid items and what detracts from item validity. It provides a foundation for larger ideas about test validity.

As it is considered and applied across content areas, this Theory of the Item can be a guide for improved content development work and, thereby, more valid items and more valid tests (i.e., the inferences made upon them).