RTD Initial Guide to Fairness in Item Development

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

Fairness is one of the three "foundations" chapters in the 2014 edition of The Standards, which says, "Fairness is a fundamental validity issue" (p. 49). This guide is intended to provide the kinds of guidance to practitioners -- particularly those involved in item development -- that the Standards cannot. It provides a framework for thinking about ten fairness issues that can both serve as an introduction for those new to this work and a reference document for veterans.

Stated simply, fairness in test development is about *maximizing accessibility of tests for the widest range of people from the widest range of background possible*. Historically, this goal has been addressed with lists of words and topics to avoid, but *true* accessibility requires a more nuanced approach -- one that take advantage of the intelligence and professional expertise of item developers.

Standardized tests are supposed provide a standardized test experience, "so that testing conditions are the same for all test takers" (p. 224). This logic, taken from *The Standards* suggests that test takers experience with individual items should also be the same -- or as similar as possible -- for all test takers. The fairness considerations explored in this guide are all aimed at furthering that goal. Regardless of a test taker's ELL status, disability status, race/ethnicity, gender or other demographic or background, they have a right to equal opportunity to demonstrate their mastery of the KSAs (knowledge, skills and/or abilities) defined in the test construct(s). Fairness aims to minimize the construct irrelevant variance that stems from test takers' backgrounds and demographics. *When construct irrelevant elements bias an item or a test for or against a particular relevant subgroup, there is a fairness issue.*

There is also another crucial aspect to fairness: facial validity. This will be discussed in its own section on page [x], but this idea is important enough to bear mentioning here. There are strong reasons to consider facial validity (i.e., political and public perceptions of tests and their contents) when developing items and tests.

Obviously, fairness "requires attention throughout all stages of test development and use" (p. 49). This begins as early as construct definition, and runs all the way through test administration, scoring and score reporting. However, this guide -- intended as it is for item developers -- focuses exclusively on fairness consideration in item production, particularly item drafting, editing and refining.

Psychometricians have developed tools to examine and catch a number of potential issues, but fairness considerations should include far more differentiators than available data could ever allow such techniques to address. Furthermore, their dependency upon field and operating test data means that they can only catch problematic items *after* all of the expense of developing items has already been incurred. Thus, anticipating and addressing fairness issues throughout the item development process *saves* money for test publishers and developers by reducing late-stage item loss.

Fairness is a goal of test development, but not one that can always be met perfectly. Furthermore, it is such a complex and nuanced problem -- even for an individual item -- that no single person could reasonably be expected to anticipate and address every potential fairness issue. Therefore, fairness considerations require a team approach to the work, even one that includes bringing in outside perspectives to help item developers to recognize and alleviate issues.

The Purpose of this Introductory Guide

Fairness work may be the most difficult and uncertain aspect of content development professionals' work. This guide offers a small set of concepts and terms that serve as the basis for how the RTD approach thinks about *fairness* to help these professionals – both those newer to this work and those with more experience – approach this work with more nuance as they further develop their professional judgment. Of course, this guide falls well short of providing a full discussion of any of these topics. It falls short of providing sufficient guidance to anyone to engage in any aspect of fairness work with real confidence. However, it does provide a starting point for content development professionals to refine their own understanding, examine their own experiences and bring in their own professional values so that they may move forward with their colleagues to improve their organization's approach to fairness considerations.

- Relevant Subgroups
- Construct Irrelevance
- Perspectives of Test Takers
- Clarity of Language
- Representation
- Familiarity
- Sensitivity
- Facial Validity and Politics
- Committees
- Measurable Bias
- Professional Judgment

Comparing the RTD and Traditional/Psychometric Approach to Fairness

The traditional approach to test development put a lot of focus on tests and on recognizing multiple demands. It assumes that a test, taken together, must meet many demands, needs and challenges, simultaneously. It often does so my including on a single test, different items that meets the needs of different test takers. For example, in order to provide information about test takers at a variety of levels, it includes items at a variety of difficulty levels. This traditional approach has a tolerance for suboptimal – or even perhaps mildly problematic – items for some subgroups because it works to make sure that these groups may be better served by other items. Thus, taken together, a single test can meet the needs of a diverse population of test takers.

RTD, on the other hand, always looks more closely at the contents of individual items than the traditional/psychometric approach.. It seeks to increase the value of each item items on a test, in part by lessening the weaknesses of individual items. While the traditional approach focuses on the fairness of entire tests, RTD seeks to address the fairness of individual items, as much as possible. Most importantly, the RTD approach to fairness is *very* careful about the potential of an individual item to distract some test takers so much that it impinges on their ability to focus on *other* items, as well.

Relevant Subgroups

The first -- and perhaps the most obvious -- fairness consideration is the question of which subgroups should be considered when evaluating an item or test for fairness. This will vary from test to test, and largely stems from the nature of the test population. Of course, almost all test production efforts must consider standard demographic characteristics (e.g., race/ethnicity, gender, SES). Other relevant subgroups differentiators could include such factors as urbanity,

region, age, English language learner status, or even such factors as experience in family roles, experience with disease, or any number of other differentiators.

A list of relevant subgroups should not contain every imaginable subpopulation of the testing population. Rather, it should include subgroups whose background and/or experience may give them a (dis)advantage. For example, lower-SES and minority groups may be disadvantaged when words/jargon from topics particularly associated with activities or hobbies rarely taken in part in by members of such groups. The classic case of this was when the word regatta (i.e., a term from the sports of rowing and sailing) appeared in a vocabulary item on a major test.

Unfortunately, it not possible to consider *every* conceivable (dis)advantaged subgroup with *any* degree of (dis)advantage. For example, two otherwise quite similar families may differ in how they spend their family vacations, and thus have quite different experiences with camping instead of hunting the beach for interesting shells. Recognizing appropriate subgroup size and even appropriate bases for relevant subgroups requires nuance and expert judgment, and sometimes deliberation between those involved in item and test development.

Construct Irrelevance

Inevitably, many items will tap knowledge, skills and abilities that are not a part of the targeted construct (i.e., that which the test is supposed to measure). For example, instructions are written in English on math, science and social studies tests. Stimuli may include references to events and activities that are not part of the targeted construct, such as a passage on a science or ELA test about the importance of putting babies to sleep on their backs. All of these are construct irrelevant skills, knowledge and/or background.

The use of construct irrelevant elements in items is a tricky topic. Obviously, sometimes it is necessary. Both advanced mathematics tests and science test usually must assume that test takers possess basic math skills. Reading passages will often assume that readers have some knowledge of living in families and in society. "Real world" or applied problems will assume that test takers have some knowledge of how the world itself works. On the other hand, items that depend on construct irrelevant elements may prevent some test takers from demonstrating KSAs that they actually possess (i.e., create *construct irrelevant variance*) . For example, a math item about baseball statistics may be much easier for people who already know a lot about baseball than it is for people who know nothing about baseball, even if the item is intended simply to measure basic arithmetic skills.

Construct irrelevance is a problem when it significantly advantages or disadvantages a particular group of test takers by unreasonably enhancing or limiting their ability/opportunity to demonstrate their proficiency with the targeted KSA(s) of an item.

Perspectives of Test Takers

Test and item development teams almost invariably differ from their tests' testing population in a number of ways. For example, test and item developers have far more experience with the content of the test than test takers, are usually much older than test takers, have quite different work histories, in addition to other potential demographic differences. Some of these

issues are addressed with the use of Bias/Sensitivity/Fairness Committees, who bring additional perspective and backgrounds to the item development process. However, the diversity of potential perspectives and backgrounds of test takers requires each and every item developer be mindful of perspectives *other than their own* (or those in which they have particular interest) throughout the item development process -- and not merely include representations of them in the test.

Item developers must consider the experience of test takers as they attempt each item, and how test takers' backgrounds impact or shape that experience. Item developer need to anticipate how different backgrounds or experiences can lead to the kind of strong positive or negative reaction that can affect test takers' performance on an item. They need also to anticipate how (lack of) familiarity with construct irrelevant material in an item can affect test takers' performance (i.e., create construct irrelevant variance).

Item and test developers should

- Each consider multiple perspectives of individuals when examining in-process items, in addition to thinking about representations of different groups in the test.
- Constantly ask such questions as, "Who might find this item alienating?" and "Who might be significantly (dis)advantaged due to construct irrelevant elements in this item"?
- Periodically review lists of Relevant Subgroups to remind themselves of the perspectives they need to be considering as they do their work.
- Consider both positive and negative reactions to items and their contents.

Clarity of Language

Differences in test takers' backgrounds (e.g., regionality, native language, familiarity with a topic) can lead to different interpretations of the language and phrases used in items and stimuli. This, in turn, can impact performance and thereby create construct irrelevant variance. That is, language that might otherwise appear clear and unambiguous to a thoughtful reader can be unclear or have a different meaning even to another thoughtful reader with a different background. These differences may be found between relevant subgroups or even within relevant subgroups.

Item and test developers should

- When examining stems, prompts and stimuli for clarity and ambiguous language, consider the impact of test-taker background on their interpretations.
- Be aware of regional idioms and regional differences in the way words are used.
- Pay particular attention to how English language learners and speakers of English as a second language will understand stimuli and items on non-ELA tests.
- Be careful in use of construct irrelevant jargon or technical terms that may advantage those with prior knowledge of a topic.
- Consider using a picture, illustration, diagram or chart to enhance clarity.

Representation

Depictions of people, groups and activities on a test can create serious fairness issues. These issues can arise through offensive or insulting depictions within items and/or stimuli, and/or through the absence of appropriate depictions. Obviously, the former should be avoided. The, latter however, is much more difficult to address, as it requires both examining individual items *and* the range of depictions across the item pool and on individual forms. Stereotypical representations can also create fairness issues, particularly when not appropriately balanced with non-stereotypical depictions of that same group and/or activities and artifacts associated with that group.

Item and test developers should

- Be careful of negative depictions of identifiable groups.
- Include depictions of individuals in both traditional and nontraditional gender roles (e.g., both male and female doctors).
- Vary the cultural origin of names used for people mentioned in items.
- Vary the urbanicity and region of the locations and situations depicted in items and stimuli.

Familiarity

Test development has addressed problematic representation on tests so well that familiarity is now a bigger issue for fairness -- perhaps because representation is simply easier to address. Some groups of test takers have sufficiently greater familiarity with construct irrelevant topics and events in items and stimuli that they have a significant advantage in understanding an item and/or stimulus -- creating construct irrelevant variance. As research has long established, background knowledge can lower the effective difficulty of a reading passage, and therefore, familiarity is a key validity issue. A test taker's background and experience (i.e., membership in one or more relevant subgroups) should not impact his/her performance; tests and individual items should strive to be as test-taker-background-neutral as possible.

Item and test developers should

- Make sure to understand how background knowledge and experience can impact item and passage difficulty.
- Think carefully about the background knowledge/familiarity that an item may assume.
- Look both for vocabulary and activities with which some relevant subgroups of test takers make be significantly more familiar with than others.

Sensitivity

There are two quite different aspects of sensitivity, though in practice there is quite a bit of overlap between them.

First, items and stimuli that evoke a sufficiently strong emotional reactions in relevant subgroup(s) of test takers create construct irrelevant variance when these emotional reactions distract test takers from being able to demonstrate their proficiency. This can arise from problematic depictions, use of offensive or derogatory words, and/or inappropriate inclusion of potentially distracting topics (e.g., domestic abuse, human sexuality, violence).

Second, test sponsors may wish developers to avoid particular topics in items and stimuli that some portion of the public find highly objectionable -- regardless of a sponsor's own view of the language or topic. This dynamic relates to the issues discussed in *Facial Validity and Politics* (below). In addition to the kinds of sensitivities mentioned above, this may include political hot button issues of the day.

When handled poorly, the first aspect of sensitivity can devolve into *over*sensitivity, and taint the credibility of all fairness issues. The fact that one might imagine someone being offended in some way by an item or stimulus is not sufficient to object to it on the grounds of sensitivity. Rather, the distraction or public objection must both pertain to a relevant or large enough identifiable group *and* be sufficiently strong as to create a problem (i.e., sufficient construct irrelevant variance or sufficient public backlash). Thus, when handled properly, the credibility of remaining claims will be bolstered, as will the entire topic of fairness.

Item and test developers should

- Be aware of topics and language that can trigger strong emotional reactions in test takers sufficient to hamper their performance.
- Consult test sponsors regarding sensitive topics and how (or whether) they want them included.
- When faced with a potential sensitivity objection, consider carefully the thresholds for the size of the impacted group and the size of the impact of a sensitivity concern.

Facial Validity and Politics

It is easy to decry many issues of fairness and efforts to address them as "mere politics" or "just facial validity;" this would be a mistake. Facial validity and politics *are* important in assessment for multiple reasons.

First, a test that is not seen as valid (and fair) by the public has a limited commercial and political future. If the public is sufficiently opposed to a particular test, decision-makers (e.g., politicians, high level government officials, school district leaders) will select another test instead. They might even abandon a test they have already selected. This can critically hurt a publishers' chance to recoup their investment in development and can damage a developer's credibility come the next RFP. Simply as a matter of business, facial validity and politics matter a great deal.

But facial validity is also an important validity issue. Collective public perception of a test can shape individual test takers' perceptions of that test. Test takers who view a test as biased or unfair are unlikely to commit the same effort as those who believe the test *is* fair. Thus, politics and a lack of facial validity can create a new source of construct irrelevant variance in test scores.

These matters are important *whether or not the underlying objections are accurate*. Managing public perception of a test *is* often a political effort, and one that is well worth the effort and expense.

Test and item developers should

- Throughout the item development process, try to minimize the potential for items to which significant elements of the public may object on fairness grounds.
- Include individuals and representatives of appropriate groups in the item development process who may specifically alleviate the particular concerns about the test (e.g., on Bias/Sensitivity/Fairness and Content Validity Committees).
- Respond to complaints (both from the public and from participants in the review process) about potential fairness as though each is made is good faith.
- When creating ancillary documentation to support or educate the public about the test, make sure to include appropriate documentation of the overall fairness process, as well as annotated representative items that reflect the principles of this document.

Committees

Bias/Sensitivity/Fairness Committees are invaluable to the creation of fair tests, particularly because they bring additional and outsiders' views to the test development process. Unfortunately, when not properly trained, they are far less able to provide the advice and insight that is so valuable. Individual members may limit their contributions to concerns pertaining to the subgroup of which they themselves are either members or in which they have specialized professional interest. Committees may get bogged down in tangents and/or oversensitivity. Despite the fact that the Standards only mention fairness committees' responsibility to identify aspects of items that "could be seen as offensive" (p. 64), when utilized to best advantage fairness committees should consider any issue in items or stimuli that (dis)advantage relevant subgroups of test takers.

Test and item developers should

- Carefully staff fairness committees with professionals who both can thoughtfully speak to their own perspectives and to those of others.
- Take care to properly train committee members on their responsibilities and the issues of fairness they should consider (i.e., those explained in this guide).
- Include in fairness committee training a review of the relevant subgroups of the testing population for the particular test.
- Encourage each committee member to consider all relevant subgroups as they would have others consider the subgroups most important to them.
- Regularly revisit with the fairness committee the appropriate thresholds for fairness issues (i.e., likelihood of creating construct irrelevant variance for a relevant subgroup).
- Facilitate fairness committee meetings and ask questions of the committee, without participating in discussion themselves.

- Use nuanced professional judgement when reviewing committee discussions and feedback (i.e., view it as valuable advice, rather than as commands).

Measurable Bias

Unfortunately, despite care and diligence through the item development process regarding potential fairness issues, items slip through that function differently for different subgroups. Luckily, psychometricians have developed and refined techniques for catching such items in the field testing stage (e.g., DIF studies). Reliance on catching problem items in post-fielding analysis is expensive, as these items have already gone through the entire item development process, only to be rejected or incur the additional expense of going through virtually the whole process anew (with edits). Furthermore, these methods are limited by the demographic and background that are collected on test-takers.

Item develops should

- Learn the strengths and limitations of statistical techniques for flagging items with potential fairness issues.
- Work closely with psychometricians to find and investigate items that may have performed differently for different subgroups in field testing.
- Work closely with psychometricians to monitor operational test items for potential fairness issues.
- Not rely on psychometric statistical techniques in place of thoughtful consideration of fairness issues throughout the item development process.

Professional Judgment

Professionals in all fields develop professional judgment, based on their training, their experience, their values and what they learn from their colleagues. Unfortunately, content development professionals hampered in this process because of the paucity of training, the lack of good resource materials and the time pressures that often prevent sustained collaboration and/or discussion with colleagues around the thorniest matters they encounter. As a result, content development professionals are often left to reinvent the wheel, either on the organization or team level, or even on the individual level. They try to think though matters, as best they can, but lack opportunities to check their thinking or learn from the best thinking of others in their field. Thus, the development of their professional judgment is too often limited to their personal judgment, through no fault of their own.

Item develops should

- Seek out peers – in other organizations if necessary – to discuss their own questions about fairness in their work.
- Try to adopt a language that allows them to document and discuss the issues and conflicts they encounter when addressing fairness in their work.

- Take on an intentional and mindful learning stance as they develop their own professional judgment around fairness considerations.
- Remain humble so that they may learn from those with other perspectives and experiences about how fairness issues may manifest for others.

Lists of Fairness and Sensitivity Topics

Many organizations have developed lists of topics that layout particular areas, concerns, topics and/or subgroups that have or could introduce fairness concerns. Perhaps none is more thorough than the *ETS Guidelines for Fair Tests and Communications*. However, no such list can be truly comprehensive. For example, list of controversial topics to avoid cannot be set in stone. More importantly, thresholds for inclusion on such lists should be products of carefully deliberated professional judgment and may vary by content area, test population and/or client. At their worst, they can actually interfere with the ability to address some constructs. This may be an easier response for those who do not work closely with item development, but content development professionals can do better.

Item develops should

- Make sure to know and understand every entry on whatever sensitivity concerns list are attached to their tests and/or come from their clients.
- Be familiar with the more thorough publicly available lists and the reasoning behind the topics included in them.
- Use existing lists to help development their professional judgment, including raising their awareness of concerns with which they may have had less experience.
- Be careful *not* to let such lists serve as substitutes for their professional judgments.