Spotlight on Assessment

Content Validity Evidence in the Item Development Process

Catherine Welch, Ph.D., Stephen Dunbar, Ph.D., and Ashleigh Crabtree, Ph.D.

Test items (questions) are the backbone of every test. They are the manifestation of the content standards, meaning they are the bridge from what students know and can do in the classroom to the standardized evaluation of learning and growth. Much of the content validity of test scores rests squarely on the test items and their assembly into test forms.

Content validity is the most fundamental consideration in developing and evaluating tests. Without content validity evidence, we are unable to make statements about what a test taker knows and can do.

The development of test items is a careful, meticulous process with safeguards along the way to ensure content accuracy, fairness, statistical integrity, and accessibility for students.

Item development is a careful, multi-step process that begins once the test purpose has been defined, once a common understanding of the target domain exists, and once content and performance standards have been developed. Then it is possible to write items that measure the content standards. During the design of the test, specifications are developed that outline the content that is to be assessed by the test. Item specifications are an important component of the process for the item writer. These specifications are the framework used by writers to create and develop the test items. Once the test specifications are developed, the item development process can begin.

Figure 1 below illustrates what happens during the life cycle of an item. This brief will describe the process and discuss each stage and how it relates to the content validity of test scores. Initially, **item writers are selected** who have the appropriate experience, expertise, and background to generate content for the test. The writers generally pass a rigorous selection process to ensure that the writers are familiar with both the content of the test and the group for whom the test is being written. Item writers must have an understanding of both the content and the examine population by possessing exceptional content knowledge, experience with the age group taking the test, and the ability to communicate ideas and content clearly.

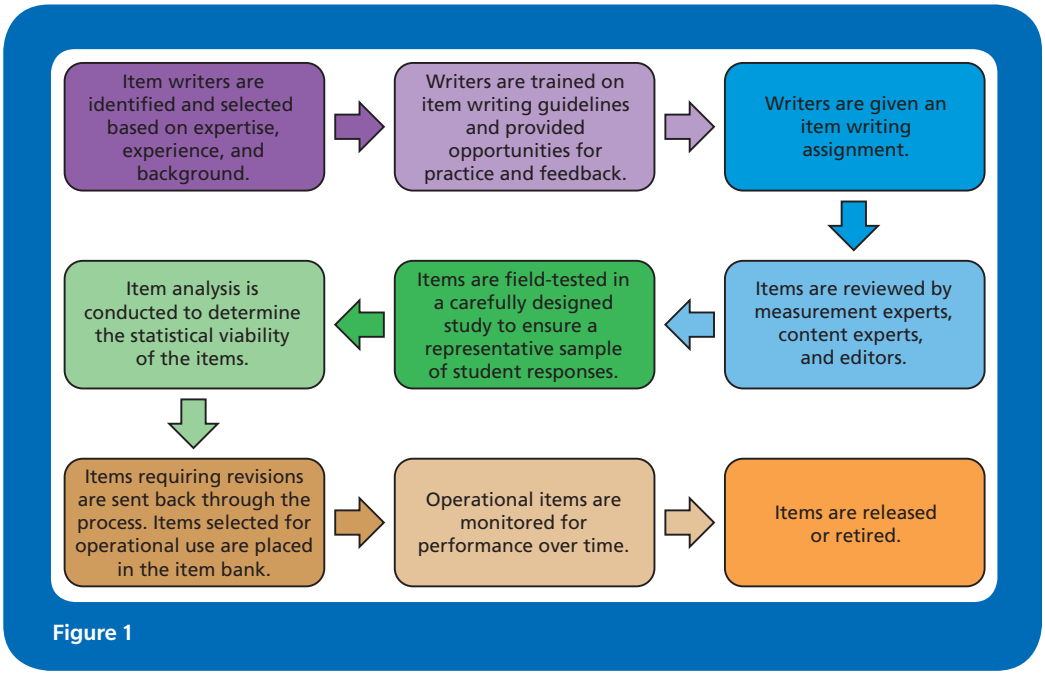


Figure 1

"The greater effort expended to improve the quality of test items in an item bank, the greater degree of validity we are likely to attain. As item development is a major step in test development, validity can be greatly affected by a sound, comprehensive effort to develop and validate test items."

- Haladyna and Rodriquez, 2013

Next, **item writers receive intensive training**. This is another key way that content validity can be assured. The training must be robust and intensive in order to assure that the writers are able to communicate and write effectively. Item writers learn how to transition from content standards to questions that address the appropriate content and cognitive demand of the item. The spirit of the content standards is represented in the group of items selected for the test, so it is *always* important that there is a clear link from the items to the content standards. Once the item writers have been properly trained on best practices in item writing, **they receive their item writing assignments**. The allocation of item writing assignments is where the content coverage of a test begins to take shape. In order for a test to measure the content standards according to the cognitive and content demands outlined by the test specifications, careful assignments must be distributed.

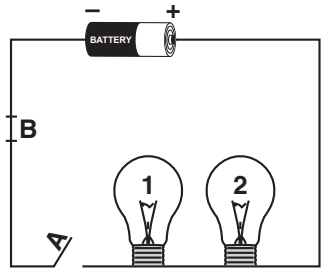
The next step in the development process is for the **items to be reviewed** by several people who are looking to review different aspects of the item. Measurement experts review the items according to best practices. Content experts are convened and the panel reviews and discusses the merits and shortcomings of the items. The panel makes certain the items add value to the test or the item bank; that the content is accurate, relevant, and significant to the content area and the grade-level being assessed. Also, at this phase of development, items often undergo additional reviews for fairness and accessibility. We are then able to more confidently maintain that the items can contribute to score interpretations.

Field testing is arguably one of the most critical aspects of the development process. **Items are field tested** to amass statistical evidence regarding whether they may be useful in the interpretations we make about what students know and can do in a particular content area. The integrity of the item-level statistics that are used in forms assembly rests with the field test. At the item level, field test opportunities provide valuable information tied directly to the integrity of the items in that the information helps the developer to evaluate, revise, and select items eligible for forms assembly or pool inclusion.

Generally, at this phase we learn how difficult students find the items and how well the items discriminate among higher and lower achieving students. A strong field test enables the next step in the item development process—item analysis. **Items are then analyzed** to determine the statistical appropriateness for inclusion on an operational form.

Let us consider an example, Figure 2.

20 **In the diagram, which of these steps will make bulbs 1 and 2 light up?**



A Close switch A
B Open switch B
C Change the direction of the battery
D Change the size of the light bulbs

2014 -2015	ITEM 20			
GR:	A	B	C	D
04	58%	10%	18%	13%

Figure 2

This is an item statistic card for a science item that was tried out on a population of fourth graders. This information tells us that the item was tested in grade 4, and 58% of the students who responded to the question answered correctly by selecting option A: *Close Switch A*. We see that 10% of students selected option B; 18% selected C; 13% selected D. In a test measuring learning outcomes, this type of information helps us to better identify whether students have mastered the content measured in this item. Items that require revisions are sent back through the editing and field testing process. This is an important part of accumulating content validity because it deals directly with the accuracy and significance of the item content.

Once items have proceeded through the development process, the **items are then eligible to appear on an operational test form**. It is essential to monitor the items performance over time and with each new operational

sample of students. As new items make their way through the development process and become eligible for operational testing, **older items are then retired**. They are removed from the test form and sometimes released to the public in the form of practice items or sample items.

The item development process is a careful progression of steps that together work to ensure the appropriateness of the items that will end up on an operational test form and make up an operational item pool. Ultimately, these items will be used in the assessment of what test takers know and can do. Thus it is of utmost importance that we work to validate them and accumulate content validity evidence supporting their use.

Validity Evidence According to Test Purpose: Item Development

When developing items for a test, purpose must still serve as important motivation for the writing, review, field testing, and eventual approval or revision of test items. While the item development process as described remains consistent for many testing purposes (assessment of learning outcomes, growth, readiness, etc.) some steps may require more attention, a nuanced perspective, or a slightly different focus.

Student learning outcomes generally describe what students know and are able to do at particular points in time during their educational experience. Developing a test to measure whether students have obtained these outcomes may coincide with determining student proficiency, as well as evaluating students against a particular criterion, or set of standards. When determining if a student is proficient in a particular content area, it is especially important that items appropriately sample the test specifications and target domain.

During the item writing phase, the writers must be specifically trained to write items that assess the range of content and cognitive demands outlined in the standards and test specifications. Of course, this is always a critical part of the item development process, regardless of purpose. However, in the case of evaluating student outcomes, it is imperative that the items on the test appropriately sample the content and are written to cover the variety in both content and cognitive demand required of the standards and the curriculum.

Training for an item writer who is writing for a test to evaluate student outcomes or student proficiency may include guidance as to how to appropriately assess the standard while paying important notice to the cognitive demand. Figure 3 illustrates how that training may be presented.

Level	Description	Suggestions for item writing prompts
Knowing	Recall or recognize knowledge	Define, label, identify, match
Comprehending	Understand knowledge and make comparisons	Compare, explain, paraphrase
Applying	Use knowledge in a new situation, adapt information to a new setting	Compute, extend, generalize, modify
Analyzing	Break down knowledge, differentiate among fact, opinion, and hypotheses	Categorize, diagram differentiate

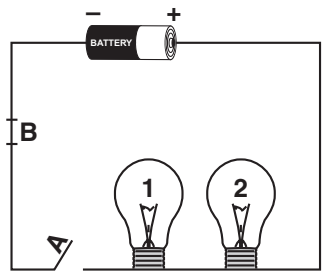
Figure 3

Field testing is also a part of the process that may be altered slightly depending on the purpose of the test. For instance, a test to measure student proficiency in a particular subject at a particular grade level—or time of year—will sample heavily from a population of students in the appropriate grade level who are receiving instruction in the subject at the appropriate time of year. However, for a test that is also designed to measure growth, it is important to represent the adjacent grade levels in the field test sample. This type of nuance is especially important and will have implications for item analysis.

Consider the example (Figure 4, on page 4) from the item card that was introduced on the previous page. This item shows the same item presented previously. This is an example of an item that can meet several purposes when carefully written and evaluated. The statistics shown earlier are used to help measure whether students have obtained the outlined learning outcome. By collecting different data, the same item can be used to evaluate growth.

The statistics in this item show that the item was tested not only at grade 4, but also at grades 5 and 6. As we would expect, the item gets progressively easier for each grade level, indicating some degree of growth. At fourth grade, 58% of students get the item correct, but in sixth grade 81% of students correctly answer this item. This item also shows a quality index (QI). This value is another way we learn how the item is performing among testers. Any value greater than .30 is considered acceptable. We can use this value to determine in what sample the item best distinguishes content that students know from content students do not know as well.

20 **In the diagram, which of these steps will make bulbs 1 and 2 light up?**



A Close switch A
B Open switch B
C Change the direction of the battery
D Change the size of the light bulbs

2014 -2015	ITEM 20				
GR:	A	B	C	D	QI
04	58%	10%	18%	13%	.49
05	75%	2%	11%	6%	.53
06	81%	1%	6%	8%	.67

Figure 4

Assuming the content standards have been written to account for a continuum of learning in the target domain, the results of this item analysis reveal that students are able to show growth with respect to the knowledge required of this item. Again, this reiterates the idea that an item can be written to serve several purposes. If this item was appropriately aligned to the content standards, it would be possible to use the item at the fourth grade on a test to indicate proficiency, but also use the item across grades on a test used to measure growth.

Conclusion

Validity is greatly influenced by a rigorous effort to develop and validate test items. As developers struggle to measure new and changing constructs, the item development process, field test sample, and the resulting item-level statistics must be evaluated with respect to appropriateness and fidelity against the interpretation and use of this information.

By carefully attending to test purpose during the accumulation of validity evidence, it is possible for items to serve several purposes and allow for varied interpretations according to test use. It is the responsibility of the test developer to align the item development process and the validity evidence associated with it to the purposes of the test.

Authors

Catherine Welch, Ph.D. is a professor of Educational Measurement and Statistics at the University of Iowa. She teaches graduate-level courses in educational measurement and conducts research in the areas of test design, interpretation, and growth. Dr. Welch has responsibilities with Iowa Testing Programs, where she directs statewide testing for the ***Iowa Assessments***™ and the Iowa End-of-Course Assessments. She is a principal author of the ***Iowa Assessments***.

Stephen Dunbar, Ph.D. is the Hieronymus-Feldt Professor of Educational Measurement in the College of Education at the University of Iowa, where he has taught since 1982, and also serves as Director of Iowa Testing Programs. His primary research interests are in the areas of test development and technical applications in large-scale assessment. He is a principal author of the ***Iowa Assessments***.

Ashleigh Crabtree, Ph.D., is an Assistant Research Scientist for the Iowa Testing Programs.

To learn more about the *Iowa Assessments*, please go to riversideinsights.com to view author video clips and download informational brochures, scope and sequence resources, and additional white papers. Contact your Assessment Account Executive or call Customer Service for a presentation.

Connect with us:

Riverside Insights