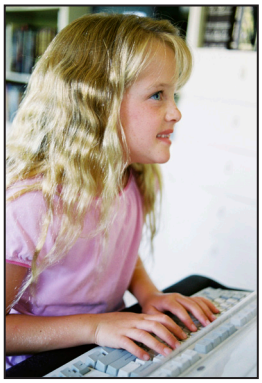
The University of Iowa
ITP
Iowa Testing Programs

Content Validity for Large-Scale Assessment



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.

What is content validity?

Validity refers to the evidence we have to support the way test scores are used and the impact these uses can have on individuals. We use the scores from tests to make inferences about what students know and can do. Validity affects the inferences we are able to make from test scores.

Content validity is one source of evidence that allows us to make claims about what a test measures. It is the degree to which the content of a test is representative of the domain it is intended to cover. In order to use a test to describe achievement, we must have evidence to support that the test measures what it is intended to measure. For instance, if after administering a test we want to make statements about how a student reads, it is imperative that the test comprehensively measures the most important, relevant topics essential to the subject and skill of reading. All educational assessments aim to reason from specific things students do, make, or say, to broader inferences about their knowledge and abilities. Without evidence of content validity, we cannot have confidence in these inferences.

How do we establish content validity evidence?

Articulation of test purpose

The purposes of a test define how the test should be used, who should use it, who should take it, and what types of interpretations should be based on the results. This is why the purposes of a test must be clearly stated at the outset of the assessment development process. Once the test purpose is defined, the test can be developed such that the outlined purposes are always at the forefront of the development process. Then it becomes possible to evaluate how the items are selected, how a test is used, and what is done with the results relative to the articulated test purpose.

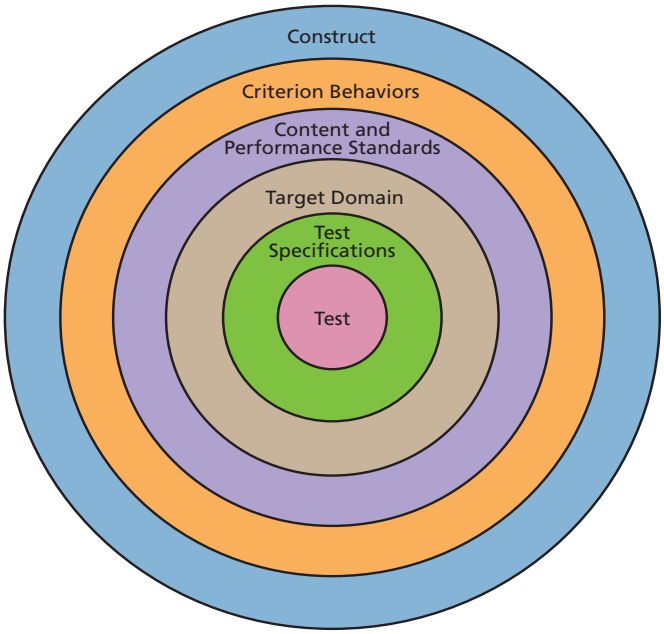
"The decisions which are made preliminary to actual test construction are, from the broadest point of view, far more important or crucial than those which follow."
E.F. Lindquist, founder of Iowa Testing Programs

Content Validity for Large-Scale Assessment

A Content Validity Perspective

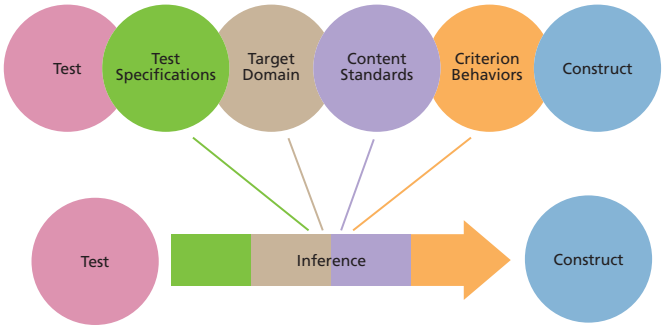
Once the test purpose is clear, it is possible to develop an understanding of what the test is intended to cover. It is the test developers' responsibility to provide specific evidence related to the content the test measures. In evaluating large-scale assessments, this requires a very specific statement of the test content, or test domain. Often this comes in the form of content and performance standards as well as test specifications, which together outline what can be covered on an assessment.

It is possible to think of the process of defining test content in terms of concentric circles. The largest and most encompassing circle is the construct. The construct is the concept or characteristic that a test is designed to measure. It may be a broad range of knowledge and skills represented by subject area domains. Next, it is necessary to identify the student behaviors that are examples of those constructs, and then determine what types of tasks or situations can be used to elicit those behaviors.



Once criterion behaviors are established, it is possible to develop content and performance standards that appropriately communicate them. From there we can define the target domain and the types of items that appropriately sample that domain by creating test specifications to guide the development of the test.

In large-scale assessment it is not possible to directly measure all student performance. The full range of performance must be inferred from observations collected from students. In quality assessments, this evidence is representative of the set of standards, or domain of knowledge and skills, to which we want to make inferences. The evidence we have about each of the concentric circles contributes to the inference we make about what students know and can do related to the construct.



Inferences are made from the test, which represents a sample of the target domain. The test must present situations to the test taker that are specifically designed and selected to elicit the desired behaviors. Given the content and performance standards, the target domain for large-scale assessment is established from these standards. This is how we determine which standards are appropriate for large-scale assessment and which standards are better evaluated with classroom projects or other formative assessments.

Sampling is the process whereby test developers articulate the target domain. This is done by establishing evidence for what defines the domain, as well as evidence for what is and what is not assessable. Sampling also determines what proportion of the assessable content and skills will appear on the test. This is an important distinction that must be made during sampling. Establishing content validity is not only about providing evidence supporting what makes up the target domain, but it is also about providing evidence for what can and cannot be tested reasonably and efficiently within that domain. This is not to say that all of the content within the target domain is not important. Quite the contrary, this process provides evidence such that important content can be evaluated in other equally important ways, outside of large-scale assessment.

Returning to the concentric circles, let us operationalize our understanding by using the subject of reading as an example.

Content Validity for Large-Scale Assessment

Construct	Criterion Behaviors	Content and Performance Standards	Target Domain for Large-Scale Assessment	Test Specifications	Test
Reading	Students are able to read widely and deeply from a range of high-quality literary and informational texts	Key ideas and details Craft and structure Integration of knowledge and ideas Range of reading and level of text complexity Comprehension and collaboration Presentation of knowledge and ideas	Key ideas and details 1. 2. 3. Craft and structure 4. 5. 6. Integration of knowledge and ideas 7. 8. 9.	Items • Multiple Choice • Constructed response Type of text • Fiction • Nonfiction Percentages • Key ideas and details • Craft and structure • Integration of knowledge and ideas	Reading test

The figure above expands the concentric circles into a more detailed framework for defining test content. The **construct** of reading is essential for students of all ages and is often measured in large-scale assessment. This is the outermost circle and the first section in the chart. The **criterion behavior** is for students to be able to read widely and deeply from a range of high-quality literary and informational texts. This criterion behavior helps give context to the construct, and assists us in further defining the standards. In order to effectively determine whether a student is able to demonstrate this criterion behavior, **content and performance standards** are developed. These standards include components of reading such as key ideas and details, craft and structure, integration of knowledge and ideas, range of reading and level of text complexity, comprehension and collaboration, and presentation of knowledge and ideas. The purpose of these standards is to help us elicit the criterion behavior when designing the test.

For the purposes of sampling, it is essential to identify the standards that can be assessed in a way that allows for efficient and reasonable measurement of content and skills. This process is how we better define the **target domain**. It is possible that test developers may decide that requiring students to read "widely and deeply from a range of reading and levels of text complexity" is more ambitious than a large-scale assessment can accommodate—meaning it may be challenging to develop test items that will measure this component of reading in a reasonable amount of time. Again, it is important to reiterate that this is not a statement about the importance of a given standard. It is instead a process that helps to more clearly define the domain of the test. This reading component clearly has an important place in the curriculum for reading, but it may be more appropriate to obtain evidence of it in ways other than on large-scale assessments.

Once the target domain it defined, it becomes necessary to rely on samples of items that match the test specifications to estimate an individual's domain score. The **test specifications** detail the type and quantity of items to be included on the assessment. Taking care to adhere to the specifications helps ensure that the **test** will adequately sample the target domain. The quality of the inferences made from the test scores is directly related to the quality of the sampling from this domain. The items must be developed to clearly assess the domain. Thus, the domain must be well-defined and the sample of items must be relevant to and representative of it. The question about inference becomes: to what extent is a score on this test reflective of a test taker's understanding of the target domain? This is the essential question

Content Validity for Large-Scale Assessment

of validity. Once we have accumulated validity evidence surrounding the content of a test, we can confidently use scores from the test to make inferences about what a student knows and can do as it relates to the construct. Each of the circles is essential to making a valid inference.

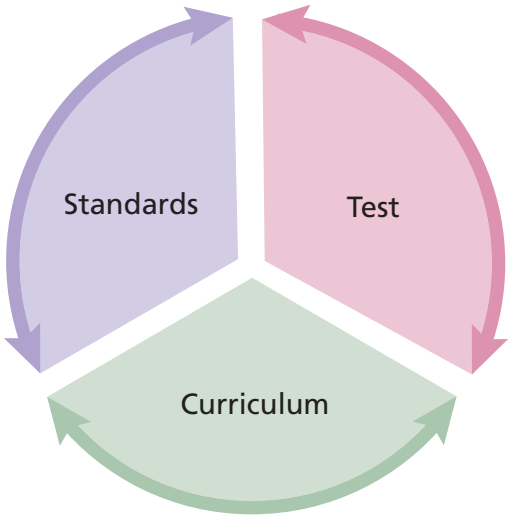
In sum, by defining the **reading construct**, we are able to identify the core reading **criterion behaviors** that are relevant to the construct. Once those are identified, **content and performance standards** can be written to help further articulate those behaviors. The **target domain** derives from those content standards and defines what is and what is not appropriate for large-scale assessment. We are then able to develop **test specifications** and develop a **test** that is representative of those specifications. When we have amassed content validity evidence, we can confidently make interpretations about students' knowledge and skill as it relates to the **construct of reading**. In order to make inferences back to the construct, we must carefully gather validity evidence throughout the process.

Alignment

Alignment studies can help establish the content validity of an assessment by describing the degree to which the questions on an assessment correspond, or align, to the content and performance standards they are purported to be measuring.

During an alignment study, evaluators examine whether the categories of content and levels of cognitive thinking that appear in the standards also appear on the assessment. In addition to alignment to specific standards, an alignment study should also take into consideration how well the test is representative and aligned to the curriculum as a whole. Ideally, the curriculum should be a reflection of the content and performance standards, and vice versa. It is by examining the intersection of these three components – test, standards, and curriculum – that an alignment study documents evidence of content validity.

For a given test, an initial alignment study may be performed by the test developers themselves; additionally, independent evaluators may perform a parallel alignment study to establish yet another piece of evidence towards showing how well the test is aligned.



Conclusion

Content validity evidence allows us to make claims about what a test measures. It is the degree to which the content of a test is representative of the domain it is intended to cover. Articulating the purposes of the test, understanding and clearly defining the target domain, and working to ensure alignment of test items can provide validity evidence that allows us to confidently make inferences about a test taker's knowledge and skills with respect to the construct. Accumulating content validity evidence requires developing an understanding of the essential aspects of the path from a construct definition to the design and development of the test that measures it. What the test measures, what it does not measure, and how the scores can be used to effectively and accurately communicate what students know and can do are fundamental aspects of content validity.

Iowa Testing Programs
College of Education
University of Iowa
Iowa City, IA 52242-1529

steve-dunbar@uiowa.edu
catherine-welch@uiowa.edu

www.education.uiowa.edu/itp

The University of Iowa

Iowa Testing Programs