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 (Figure 1). 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.

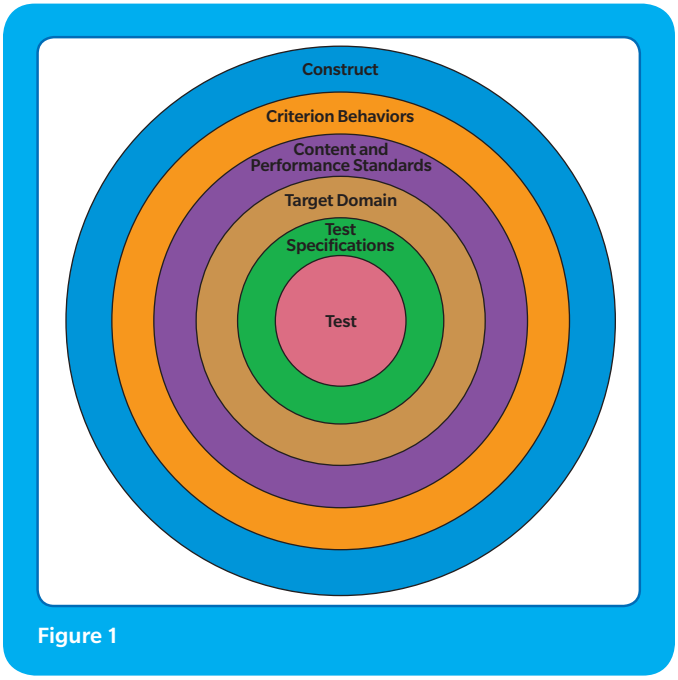
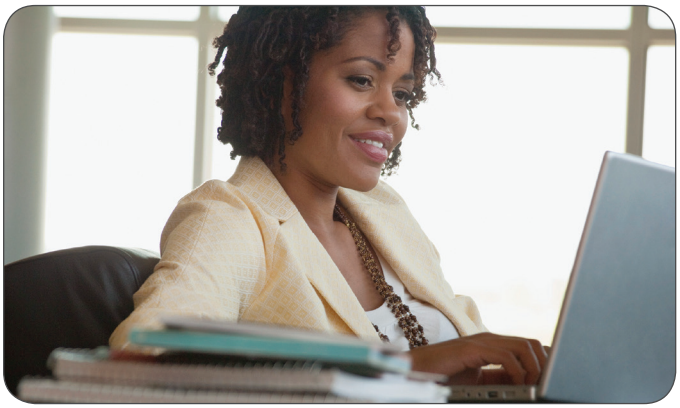


Figure 1

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 development of the test.



In large-scale assessment, it is not possible to directly measure all student performance. The full range of performance instead 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 (Figure 2).

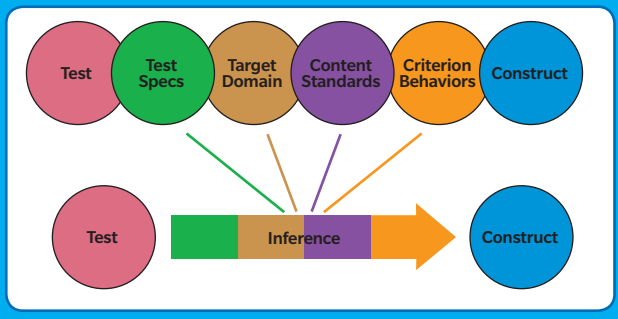


Figure 2

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 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.

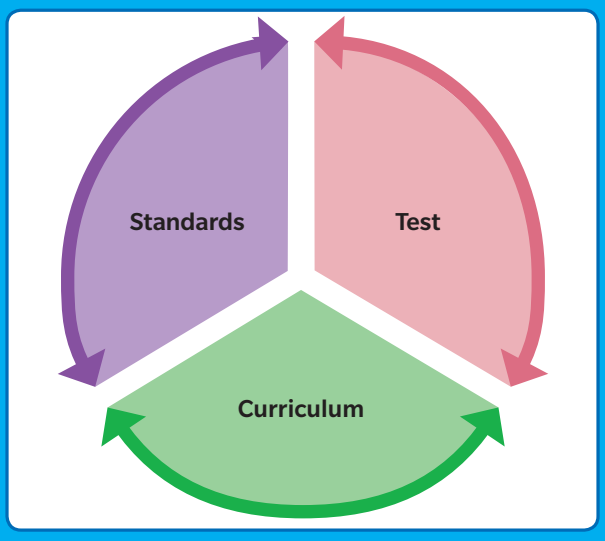


Figure 4

The relationship between the content of a test and the standards provides important validity evidence. The extent to which the same categories of content and levels of cognitive demand appear in both the standards and the assessments is examined during alignment. Results of these categorical comparisons can be presented in a way that communicates the degree to which content that is covered on the test represents the content found in the standards (Figure 4). Yet, the test must not only be aligned to the standards, it must also be representative and aligned to the curriculum. The curriculum should be a reflection of the standards and the standards must also be reflective of the curriculum. It is at the intersection of these three components that alignment begins to be a manifestation of content validity.

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.

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.