American Psychological Association Guidelines on Psychometric Assessment Validity and Reliability Requirements

Eric Gehrig, PhD*
Senior Research Scientist
Research & Development
Target Training International

April 17, 2019

Introduction

Establishing evidence of validity and reliability of psychometric assessments is more important, and more complex, today than every before. The days of claiming an assessment is valid because it correlates well with another "established" assessment are long since past. Concepts of "face" validity and comparison studies are no longer widely accepted as establishing evidence of validity for an assessment.

One of the major challenges of a company such as TTI Success Insights is to address the rigorous demands of the academic professional world as it relates to acceptably valid and reliable assessments. These demands have presented another obstacle in the form of a seemingly ever moving target as well as little in the way of guidance as to what constitutes an acceptable level of "validity" evidence. Following the evolution of the concept of validity evidence uncovers many, and often competing, views of the concept of validity. This short exposé attempts to place in a concise form the current view of the American Psychological Association (APA) on evidence of validity and reliability as presented in the *Handbook of Testing and Assessment in Psychology, Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, [2], from here on referred to as *The Handbook*.

The Handbook is really a collection of three volumes, each playing a central role in a specified area of testing and assessment in psychology. Volume 1, the main subject of interest in this short paper, is comprised of 38 sections or chapters, each individually written by psychological or assessment professionals on range of topics from (Chapter 1) *Psychometric Characteristics of Assessment Procedures: An Overview* to (Chapter 38) *Legal Issues in Industrial Testing and Assessment*. We concentrate on Chapters 2 (Reliability) through 7 (Item Analysis) in the remainder of this work.

Evidence of Validity

The APA Handbook on Testing and Assessment in Psychology has a chapter specifically devoted to measuring validity. The Handbook presents five sources of validity evidence, referencing the

*PhD, Mathematics, Arizona State University, 2007.

Standards (Standards for Educational and Psychological Testing, 3rd Ed., 1999, [1]). These sources are a) test content, b) relations to other variables, c) internal structure, d) response processes, and e) consequences of testing. We give a brief description of each source of evidence along with an example of a possible way of collecting such evidence.

Test Content Based Validity Evidence

In order to discuss providing evidence of validity in any form, in any setting, one must first define what is being measured. In the world of psychometric assessments, this is often referred to as defining the construct or, in the current language of the APA, defining the content domain of definition. Sireci (The Construct of Content Validity, 1998, Social Indicators Research) categorizes test content into four main areas: (1) construct definition, (2) construct relevance, (3) construct representation, and (4) appropriateness of test construction procedures.

Construct Definition

The construct definition, i.e., what is to be measured, is usually addressed during the early stages of development of an assessment and is usually evaluated by a group of content experts. Additionally, the content experts are usually external to and independent of the assessment development process. The content evaluators typically focus on whether important areas have been omitted or whether superfluous areas are included (The Handbook, pp. 66).

Construct Relevance and Construct Representation

Construct relevance and representation are closely related and are presented together even though they represent two distinct areas of content validity evidence. Construct relevance focuses on whether each individual item is actually relevant to the construct of interest. Construct representation is a measure of whether the individual items span the full range of the content domain and avoid items that may be irrelevant to the intended construct.

It should be noted that The Handbook states the following on pp. 66:

> Traditional content validity studies use subject matter experts to evaluate test items for relevance and representativeness. Items that are not judged relevant to the domain are eliminated and new items are added if the experts conclude that aspects of the domain are underrepresented. ... An evaluation of the appropriateness of the construct definition focuses on how well that definition captures the *consensus understanding* of the construct.

The emphasis on consensus understanding is placed by the authors of this piece, not the authors nor editors of The Handbook.

Appropriateness of Test Construction Procedures

Appropriateness of construction procedures is evaluated based on the processes in use by the assessment developer. Items considered include, but are not limited to, adequacy of item development and selection processes, scoring procedures, and quality control.

Some Comments on Test Content Based Validity Evidence

There are two main takeaways from the previous discussion on test content based validity evidence. The first, and perhaps not so obviously presented, is that test content based validity evidence is generated and measured during the assessment development process. The second and more explicitly noted is that test content based validity evidence is traditionally evaluated by content or subject matter experts, usually as an independent third party evaluation, i.e., capturing the consensus understanding of the construct.

Relations to Other Variables Based Validity Evidence

Relations to other variables is an interesting portion of the evidence of validity approach for any assessment. The section on Validity Evidence Based on Relations to Other Variables starts with the following paragraph, see pp. 66 in The Handbook:

> Although criterion-related validity evidence has been established as insufficient to fully support claims of validity for score interpretations and use, depending on the purpose of the test such evidence may be quite useful in building the overall validation argument.

The reference to criterion-related validity evidence is a clear nod to the earlier concepts of concurrent and predictive validity based evidence, and these two concepts are still largely part of the relations to other variables concept currently in use today.

There are two key points The Handbook points to in the process of selecting appropriate criterion variables and measuring the outcomes. First, there are two questions that should be a main focus of any effort in this area: (1) Is the rationale for selecting criterion variables and demonstrating their suitability appropriate? (2) Are the patterns observed between the test scores and external variables consistent with prior expectations?

The second key point is best summarized in the following quote from The Handbook, pp. 67:

> Selecting criteria that are consistent with the theory underlying the construct is easier said than done. Valid external criteria are hard to obtain, expensive, or both. In other cases, valid external criteria may simply not exist. Often, collecting criterion data is impractical, and even when such data are gathered, they may be of questionable reliability or corrupted by biases (e.g., supervisors' ratings used in an employment setting or teachers' ratings in an educational setting).

The main takeaways from attempting to use criterion-related evidence to formulate a validation argument is that the information needs to be highly scrutinized prior to use as part of the validation argument and any use of criterion-related evidence can never be a complete validation argument. It should be noted that some traditional measures of criterion-related evidence of validity are no longer widely accepted as providing evidence of validity. For example, face validity and comparisons to other assumed valid assessments are no longer considered part of criterion-related validity evidence by the APA.

Internal Structure Based Validity Evidence

According to The Handbook, the term internal structure refers to the dimensionality or underlying factor structure of an assessment, see pp. 72. The Handbook goes on to note that there are three

main ways to measure validity evidence based on internal structure: (1) internal consistency, (2) dimensionality (factor analysis), and (3) measurement invariance.

It is the composition of all of the above measures of internal structure that provide a full analysis of the evidence of validity of an assessment based on this part of a validation effort. In some cases, one may simply look to an item response theory model to measure how well assessment items cover the spectrum of location and individual ability. More advanced approaches look to provide more dimensionally based analysis such as exploratory or confirmatory factor analysis.

Measurement invariance is used to determine whether an assessment consistently measures the construct of interest across a multitude of differing characteristics such as gender, ethnicity or country of origin, age, etc. In the setting of psychometrics, understanding measurement invariance considers an application of confirmatory factor analysis across these diverse groups of individuals.

Response Process Based Validity Evidence

Response Process related evidence of validity is an interesting and in some sense fun area of the validation argument to discuss. The Handbook begins its discussion of response process validity with the following, pp. 76:

> Gathering validity evidence based on response processes is perhaps the most difficult validity evidence to gather because it involves demonstrating that the examinees are invoking the hypothesized constructs the test is designed to measure in responding to test items. ... Gathering this type of evidence is difficult because one cannot directly observe the cognitive processes going on within people's heads as they respond to test items.

.
As was presented in an earlier edition of this newsletter, this may no longer be the case. In an article published in October, 2018, the Target Training International Research team presented the protocols for a study cross-walking between the statistical and psychometric based evidence provided by an item response theory study and evidence gathered on EEG while individuals responded to assessment items. The results of this study are forthcoming sometime during 2019.

Consequences of Testing Based Validity Evidence

Conceptually, consequences of testing based validity evidence is fairly straightforward. Simply put, one needs to assess the intended and unintended consequences of an assessment program.

Generally speaking, assessments are designed with positive consequences in mind. Such positive consequences include, but are not limited to, improved instruction and education, higher levels of accuracy in diagnoses, both medical and psychological, and more. Possible negative consequences may include adverse impact that leads to decreased opportunities for specific groups, poor decisions for resource allocation, and possibly adoption of bad social policy. The debate on this topic is not closed. As an example, the section on Validity Evidence Based on Consequences of Testing in The Handbook states the following on pp. 78:

> Whether validity evidence based on consequences of testing is relevant in evaluating the validity of inferences derived from test scores is a subject of some controversy. Considerations of testing consequences are an important social policy issue, but may test

specialists believe they are extraneous to validity. However, others see the evaluation of testing consequences as a critical element in evaluating the appropriateness of using a test for a particular purpose. We believe this debate is one of nomenclature, and given that virtually all testing programs have consequences on some level, it is important to evaluate the degree to which the positive outcomes of the test outweigh any negative consequences.

Closing Remarks on Evidence of Validity

It should be noted that throughout the above remarks, care has been take to use the term evidence of validity, as well as not referring to an assessment as being valid or not. The Handbook as well as The Standards are quite clear on these topics. First and foremost, it is the *use* of the assessment and *interpretation* of its scores that are the subject of validation efforts, not the assessment itself. Second, there are many areas that evidence of validity may be gathered in. The Handbook condenses these to five main areas that have been discussed above.

Each of the five areas has a specific goal in mind. Each addresses areas of the assessment process from the development of the assessment to justification of its uses and interpretations to the cognitive functions of the individual respondent. Validation of the uses and interpretations of an assessment is a long process and the ultimate answer to the question of validity is truly the cumulative effect of many years of effort studying and understanding the whole.

Evidence of Reliability

The main idea behind the concept of reliability is to have a measurement that is as free from random error as possible. The key here is random error. In all measurements there is expected to be error. The concern is when one is unable to account for the error. For example, differences in scores that can be attributed to differences in ability level should not be considered errors in the sense of randomly generated errors.

Perhaps a more apt term is to consider consistency rather than the term reliable. It is the aim of all assessments to consistently measure the construct of interest. Thus, consistency could be considered to be across individuals on the assessment items (internal consistency), across time for individuals (temporal or test-retest consistency), across the construct regardless of the individual items (alternative forms consistency), or across a host of other measurable facets (generalizability).

Internal Consistency Reliability

Internal consistency reliability is a form of measuring reliability or consistency that relies on a content sampling process. There are essentially two forms of internal consistency measurement, the so-called split-half techniques and the homogeneity techniques. As an example, the split-half techniques, in its simplest form, literally splits the assessment in half and compares responses to assessment items based on the splitting. For example, one may look at all odds items compared to all even items, or one could split the test as close to the middle item as possible, etc.

Homogeneity techniques really are a superset of the split-half techniques. As an example Lee Cronbach showed in his original work on the so-called alpha coefficient, that his measure of internal consistency was the average of all possible measures of split half reliability coefficients.

Some of the advantages of using the internal consistency approach as compared to the alternative forms or generalizability theory approach is that internal consistency methods require a single administration of the assessment to a group of respondents. Alternative forms requires at least two administrations, one for each version or form, and generalizability theory requires as many administrations of the same or different versions as is required by the number of facets one is interested in studying. See the sections below on the topics of alternative forms consistency and generalizability theory.

One of the disadvantages of such methods is that these measures are essentially nothing more than scaled versions of inter-item correlation measures. Another disadvantage is that one may improve the internal consistency measure by simply adding more items of the same type already present in the assessment. Finally, there is the existence of the attenuation paradox due to Loevinger (1954, *The attenuation paradox in test theory. Psychological Bulletin, 51, 493-504*). The attenuation paradox states that validity is not a monotonic function of reliability.

To give a relatively gentle introduction into what it means for validity to **not** be a monotonic function of reliability, simply increasing the number of items measuring the exact same portion of the construct will increase the measure of internal consistency, by definition. However, this process will also cause the scale to become highly unidimensional. In other words, one may increase the internal consistency of a scale by adding more and more of the same or similar items, the cost of which is increasing the correlation to the point of showing we are measuring but a single aspect of the underlying construct, thereby lowering the overall validity measurement.

Temporal Consistency Reliability

The Handbook refers to temporal consistency or test-retest reliability as *stability*. The key idea at play in stability of an assessment is not that any given individual taking the same assessment on multiple occasions necessarily has identical scores. Rather, it is more appropriate to consider a larger group of individuals and be more concerned with the *stability of scores* which The Handbook defines as persistent relative standing within a group rather than consistent absolute value of a score.

For the aforementioned reason, it is natural to consider the use of the standard correlation coefficient from statistical analysis as a measure of stability of scores. Traditionally, we consider the correlation between scores on an assessment, or scale of an assessment, based on that the individuals in the study have taken the assessment at least twice within some specified time frame. The closer to one the correlation coefficient is, the more we consider the assessment scores to be stable over time.

It is important to note that The Handbook points out four areas affecting whether memory is the major influence on any measure of stability of assessment scores. These are: (1) length of time between applications of the assessment, (2) length of the test, (3) nature of the test materials, and (4) nature of the characteristic being measured.

We briefly address each. If the amount of time is too short between administrations of the assessment, there is a higher likelihood of rote memorization. Similarly, if the time is too long between administrations, the underlying characteristic of the underlying individuals may have changed. Likewise, the length of the test, if too short, may be consistent with memorized responses, or if too long may inadvertently contribute to lower test-retest correlations via test fatigue.

If test items or stimuli or test problems are very distinctive, they may promote memorization issues as well. Finally, some characteristics of a construct may promote memorization responses

(knowledge based responses) while physiological measurements may not promote such influences.

Simply put, simply reporting high or low levels of correlation for a test-retest study does not imply high or low levels of temporal consistency. More information is required. If a test-retest study is conducted on a group of individuals highly knowledgeable about a subject area with limited time between administrations, one would expect highly correlated responses between the two data sets. However, this would not imply the underlying assessment elicits responses that are highly stable.

Alternative Forms Consistency

Alternate forms consistency is just what it sounds like. In a perfect world, measuring a construct would come in many forms. As an example, one may attempt to develop two completely independent versions of an assessment designed to measure a given construct or characteristic. These two forms may then be used in separate administrations of the assessments utilizing the same group of respondents.

Ideally, the respondents would score the same in the sense of stability as defined in the previous section on temporal consistency. In the opinions of the authors of this piece, there are many issues to discuss that are most likely better left to a future article. However, we mention at least one here. The underlying assumption is that at least one of the assessments is already reliable and validated. If one of them is not already evaluated as such, what then is the point of the exercise?

Having said that, if one is working on a new generation assessment measuring a similar construct to that which has already been measured by a tool whose use and interpretation have been judged to be valid, one may consider using the older version as a measuring stick for a follow on version. All in all, the use of alternative forms of consistency appears to have many logistical and cost challenges in addition to the usual challenges of collecting appropriate data to study such concepts.

Generalizability Theory Based Reliability

Generalizability theory is a topic that entire text books have been written on. We cannot hope to do it all justice here. The topic is named generalizability theory because it is intended to be a generalization of the concept of internal consistency reliability in the sense of Cronbach. In fact, Cronbach was one of the driving forces behind the development of the theory in the 1970s.

In a nut shell, we consider temporal consistency by looking at two administrations of the same assessment to the same group of individuals over a specified time period. We then consider the correlation between the sets of scores to measure an aspect of consistency or reliability. In generalizability theory, such a study could spawn a one and two facet generalizability model in which we search for and measure sources of variations in the response data across time (for the one facet model) and across time and by individual item (for the two facet model). The individual scores may seem at first glance to be a facet, but traditionally they are not considered one.

The goal is to identify as much of the variation as possible and that the variation is assigned to the appropriate areas. As an example, one does not want the items themselves to be a large source of variation while one would expect that a large source of the variation is the individual or person differences. If the items are the source of large portions of the variation, then the items are not invariant across the population bringing into question the consistency of the assessment as well as possible the internal structure of the assessment.

It is also necessary to discuss the variation that is attributed to randomness. In this sense, we do

not discuss randomness in the sense of quantum theory. Rather, randomness is a phrase used to describe a source of error that is as yet unknown because no one has found the source of the error or no one has chosen to attempt to find the source.

Closing Remarks on Evidence of Reliability

Evidence of consistency or reliability is an important part of evaluating the overall effectiveness of the use and interpretation of assessment scores. Similar to discussing validity in terms of the use and interpretation of scores based on an assessment, we also speak of reliability in these terms. In other words, one should discuss the concept of evidence of reliability of the scores of an assessment or the interpretation or use of those scores and not the assessment itself.

Closing Remarks

Evidence of reliability and validity go hand in hand in establishing the argument for validity of the use and interpretation of an assessment. Such an argument cannot be created solely based on one aspect of either area. Establishment of these arguments takes time and solid performance of the results of the assessment over many years.

One misconception that appears to have been propagated through the commercial assessment community is that a measure of internal consistency such as the so-called Cronbach's alpha is a proxy for a validation argument. This is a false argument. Said in one way, reliability is a necessary but not sufficient condition for validity. Said in another, reliability is an important aspect of establishing the validation argument, but is by no means the entire argument.

We finish with a simple example. Consider the home thermometer an individual may use if they feel ill. If that individual measures their own temperature several times over the next few hours and the temperature consistently reads as the same over that time, the thermometer is to be considered reliable. However, for the thermometer to be considered valid, it must not only be consistent, it must also be correct.

References

[1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and psychological testing, 2014.

[2] Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors. *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, 750 First Street NE, Washington, D.C. 20002-4242, 2013. American Psychological Association.