ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);158-168

Interview with Stephen G. Sireci on Validity

Interviewer: Nuri DOĞAN*

Interviewee: Stephen G. SIRECI**

INTRODUCTION

The following questions were created within the scope of Classical Test Theory course (OLC720) offered in the doctoral programme of Measurement and Evaluation in Education branch in the Educational Sciences Department of Hacettepe University. The question of "what three questions about validity would you ask if an internationally famous expert was in front of you?" was asked 23 students taking the course, and the participants were asked to write down their questions. 69 questions in total were revised by the lecturer of the course in terms of scope, importance and clarity, and the number of questions was reduced to 37 by removing 32 of them. The students participating in the classes were then asked to order the questions selected by the lecture from the most important to the least important. The order of importance for each question was determined by adding up the scores given by the students. 15 out of the 37 questions which had been ordered according to the students' rating were selected to be answered.

QUESTIONS and ANSWERS of INTERVIEW

1. ***What does the fact that different types of evidence have increased and that there are no criteria as to what type of evidence we should prioritize make you think about classifications in relation to types of validity? What is the most valid definition and classification of validity in your opinion?***

2. ***There are different approaches in defining validity types. For example, some methods of gathering evidences for validity are mentioned in the last version of Standards Book, but not validity types as content, criterion-related or construct validity. What are the reasons for disagreements concerning these different approaches?***

I will answer these two questions together. I think the best definition of validity is provided by the current version of the *Standards for Educational and Psychological Testing*. The current version was published in 2014, and it is the 6th version. They define validity as, "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). This definition is essentially the same as that provided in the 5th edition (AERA, APA, & NCME, 1999). It is important because from the definition we can see that,

- Validity must be evaluated with respect to a particular *purpose* or *use* of a test.
- Tests are not "inherently" valid or invalid. What must be validated (that is, supported by research and theory) is the use of a test for a particular purpose.

*Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey e-mail:nurid@hacettepe.edu.tr
**Prof. Dr. ,University of Massachusetts Amherst , College of Education, Amherst–United States, e-mail: sireci@acad.umass.edu

- Validation requires both evidence and theory to support the use of a test for a particular purpose.

There is not one study that can be done to validate the use of a test for a specific purpose. There are different types of validity evidence, and the types of evidence used to defend the use of a test for a particular purpose will vary based on the purpose of the test. The last two versions of the *Standards* specify five sources of evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (AERA et al., 2014, p. 13). These sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing.

These sources of validity evidence described by the *Standards* (AERA et al., 1999, 2014) are not the same as described in earlier versions of the *Standards*. In Table 1, I list the different versions of the *Standards* and how they described different categories of validity or types/sources of validity evidence. Note that the current version describes "sources of validity evidence" because there are not different types of validity. Validity is a unitary concept, whether you put the word "construct" in front of it, or not.

I think the current (AERA et al. 2014) definition of validity, and the five sources of validity evidence are the best ways to describe validity for several reasons. First, they are not the product of one person. For over 60 years, three organizations have worked together to come to some consensus about what validity means and how test scores (uses) should be validated. Second, the definition emphasizes that validity refers to test use, and that validation requires both theoretical justification and empirical evidence. These are truisms that are hard to reject.

Tablo 1. Categorization of Validity Evidence Over Time in the *Standards*

Publication	Validity Classifications
Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal (APA, 1952)	Categories: predictive, status, content, congruent
Technical recommendations for psychological tests and diagnostic techniques (APA, 1954)	Types: construct, concurrent, predictive, content
Standards for educational and psychological tests and manuals (APA, 1966)	Types: criterion-related, construct-related, content-related
Standards for educational and psychological tests (APA, AERA, & NCME, 1974)	Aspects: criterion-related, construct-related, content-related
Standards for educational and psychological testing (AERA, APA, & NCME, 1985)	Categories: criterion-related, construct-related, content-related
Standards for educational and psychological testing (AERA, APA, & NCME, 1999)	Sources of evidence: content, response processes, internal structure, relations to other variables, consequences of testing

3. *How is the response process a test used as evidence for validity? Can response process be used as evidence for validity in examinations with extensive participation?*

Validity evidence based on response processes refers to "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers" (AERA et al., 2014, p. 15). Examples of this type of evidence include interviewing examinees about their responses to test questions, systematic observations of examinees responding to test items, evaluation of the criteria used by judges when scoring performance tasks, analysis of item response time (chronometric analysis), tracking students' eye movements, and evaluation of the reasoning processes examinees use when solving test items. This evidence is particularly useful for evaluating the degree to which tests tap higher-order skills and for evaluating how well students in different subpopulations understand the test items. Given that many new educational tests emphasize higher-level cognitive skills, evidence will be needed that these tests adequately measure these skills.

4. *Should different ways be followed in collecting validity evidence in cases when absolute evaluation and relative evaluation are made?*

I don't understand what absolute and relative evaluation mean. However, let's talk about how different sources of validity evidence are put together to make a "validity argument." Kane (1992, 2006, 2013), suggested that validating the use of a test for a particular purpose is tantamount to developing a sound and logical argument that use of the test for a particular purpose is justified. The *Standards* essentially adopted this perspective by claiming that the five sources of evidence should be coherently synthesized to support use of a test for a particular purpose. For example, they state "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (AERA et al., 2014, p. 21).

The argument-based approach to validation is similar to defending the use of a test in a courtroom. The idea is to present a preponderance of evidence that would support the use of a test for a particular purpose. This body of evidence should include evidence that the test is fulfilling its intended objectives and is not producing undesired consequences.

5. *What can be done to collect evidence for validity if adequate sample is not available to perform validity study?*

Validity evidence based on test content is typically gathered using subject matter experts, and usually represent very small numbers of experts (e.g., 10 or fewer). So, please remember validity studies are not all based on analysis of item responses or test scores. Content validity evidence is fundamental and necessary for educational tests. Much of the research on validity evidence based on response processes also uses very small sample sizes.

6. *How can the response process of a test be used as evidence of validity? Is it possible to use the response process as evidence for validity in exams with large rates of participation?*

See answer to question #3.

7. *How do we explain the relations between multiple methods and multiple traits analysis?*

I recommend reading Campbell and Fiske (1955).

8. *Boud (1995) and Messick (1995), have raised the concept of consequential validity for alternative assessment methods. Which techniques are used in determining the consequential validity based on the influence of the assessment on learning?*

Messick did not use the term "consequential validity." The *Standards* describe "validity evidence based on consequences of testing." Validity evidence based on consequences of testing refers to evaluating the intended and unintended consequences associated with a testing program. Tests are used to promote positive consequences such as appropriate diagnosis of psychological disorders, protection of the public, improved instruction, and better understanding of the constructs measured. Unintended positive consequences that were not explicitly intended or envisioned may also emerge. However, unintended negative consequences may also occur in a testing program. Examples of unintended consequences may be adverse impact that leads to decreased education and employment opportunities for certain groups, increased dropout rates in schools, and poor decisions regarding resource allocations or employees' salaries based on test performance.

Validity evidence based on the consequences is particularly important in considering the validation of tests for some purposes, such as accountability (e.g., using tests to evaluate schools or teachers). In the USA, accountability testing is required by federal educational policy, and typically comes with a

theory of action outlining the intended consequences for stakeholders. For example, using students' test results to evaluate teachers encourages teachers to teach the intended curriculum, and it is assumed this more focused instruction will improve student learning with respect to that curriculum. The degree to which these intended consequences are realized, and other, unintended consequences (e.g., decreased teacher morale, narrowing the curriculum in a way that decreases student learning) are minimized is essential to investigating the validity of educational tests for accountability purposes. Other testing purposes, such as using a test for high school graduation or college admission, also have consequences that should be evaluated.

9. ***There are also different approaches and classifications on the validity in the literature. The question is based on a hypothetical situation in order to privatize the situation. -You want to measure top-level skills (e.g. problem solving, critical thinking). The result of test scores will be used in decisions that have a high stakes qualification. How do you provide validity evidence in the process of developing the test to the meaning and use of the scores to ensure validity?***

See answer to #4.

10. ***Scores for individuals are calculated by adding up the numerical values of responses to items in Likert type scales. Yet, it is also clear that the degree to which each item serves to the relevant structure differs. How correct is it to collect the scores in a straightforward way and how high is the validity of the findings/inferences made on the basis of those scores?***

Remember that validity refers to the use of a test for a particular purpose. There are typically no one-item tests and so inferences and decisions are made on the basis of scores calculated across many items. If there is a problem with an item or two, it may or may not be important, depending on how it affects the total test score and the interpretation. It is good for validity analysis to include analyses at different levels, such as the item level, total score level, and subscore level. If you consider the 5 sources of validity evidence, all levels are accounted for. Item analyses and differential item functioning are part of validity evidence based on internal structure. Dimensionality analyses also typically focus on the item level, as do studies based on test content. However, relations with other variables (e.g., predictive validity, differential predictive validity, MTMM) will focus at the total test score level. The specific validity question to be evaluated will dictate the level of analysis.

11. ***We rather focus on test reliability in Turkey and we see that our institutions have not put most of the applications concerning validity into practice yet. Let us assume that a commission authorised in tests in Turkey would like to work with you. What applications and how would you like to change by considering validity on the basis of scientific and social values? What would you recommend?***

Reliability is important, but it is more important to remember that tests that are not valid for a purpose can still be reliable. For example, a college admissions test may produce reliable scores, but it would not be valid for assigning grades to students in a math course. I think my answers to the previous questions describe how I would go about the process of test validation.

12. ***It is known that visually impaired students are given test booklets printed in differing font sizes according to the degree of their sight and students who cannot see are not held responsible for visual questions and are asked to answer fewer questions with extra time with the support of a reader in examinations held by Centre for Measurement- Selection and placement in Turkey. In addition to that, there are differing practices for those who certify their handicap. By taking the above mentioned situations into consideration, should validity evidence for a test be considered on normal conditions, or should different validity evidence be searched for the sub-groups of different characteristics?***

It depends what the validity questions are. Remembering that validity refers to a specific testing purpose, one question might be if the test is similarly valid for standard test administrations and accommodated test administrations. Another question might be if the scores from standard and accommodated administrations are comparable. A more specific question might be whether the accommodation has changed the construct measured. There is a great deal of literature and applied research in these areas.

13. ***Can a solution be found to the problem of range narrowing with a statistical approach?***

Restriction of range is really important when evaluating test data. I have seen many studies that have concluded a lack of invariance across groups (at either the item or total test score level), that is probably just picking up on differential restriction of range. Statistics such as item biserials, reliability coefficients, factor loadings, and etc., need variability. If a test is too easy or too hard for one group, it will look like a source of bias, but really it is just an artifact of restriction of range.

One way we have handled this problem is to sample from the unrestricted group in a way that matches the distribution in the restricted group. That strategy controls for differential restriction of range. Of course, disattenuation for restriction of range is also handy when appropriate, and when you have the data to do it.

14. ***Messick (1995) states that validity can be defined broadly as the result and use of both evidence collection and score interpretation and sees content validity and criterion-based validity as sub-parts of construct validity. Messick points out that the concept of unitary validity demonstrates the construct validity of a test. The concept was criticised in that it could not answer simple questions about what a test measured and that it was rather related with complex score interpretations which were explained with nomological network. Considering the fact that different types of validity serve to different purposes, is it possible to form an umbrella term for validity (as different from construct validity)?***

15. ***Messick (1995) states that validity can broadly be defined as the result and use of both evidence collection and score interpretation and sees content validity and criterion-based validity as sub-parts of construct validity. Accordingly, how correct is it to investigate construct validity independently of content and criterion-based validity? Should the validity of a measurement always be investigated as a whole? Which type of validity should firstly be looked at?***

Please see the Sireci (2012) for an answer to these two questions.

Stephen G. Sireci ile Geçerlik Üzerine Söyleşi

Söyleşiyi yapan: Nuri DOĞAN[1]
Söyleşi yapılan: Stephen G. SIRECI[2]

Çeviri: Nuri DOĞAN
Burcu ATAR[3]
Nermin KIBRISLIOĞLU UYSAL[4]
Osman TAT[5]
Cem MALAKÇIOĞLU[6]

GİRİŞ

Aşağıdaki sorular Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı doktora programındaki OLC720 Klasik Test Kuramı dersi kapsamında üretilmiştir. Derse katılan 23 öğrenciye "Eğer karşınızda 'geçerlik' konusunda uluslararası tanınırlığa (üne) sahip bir uzman olsaydı, ona geçerlikle ilgili sormak istediğiniz en önemli üç soru ne olurdu?" diye sorulmuş ve bu soruları yazmaları istenmiştir. Böylece 69 soru elde edilmiştir. Elde edilen 69 soru kapsam, önemlilik ve anlaşılabilirlik bakımından ders sorumlusu tarafından gözden geçirilmiş ve 32 soru çıkarılarak 37 soruya indirgenmiştir. Derse katılan öğrencilerden ders sorumlusu tarafından seçilen soruları en önemliden en önemsize doğru sıralamaları istenmiştir. Öğrencilerin verdiği puanlar toplanarak soruların önem sırası belirlenmiştir. Öğrencilerin puanlarına göre sıralanmış 37 sorudan 15'i cevaplandırılması için seçilmiştir.

Cevapların çevirisi beş farklı araştırmacı tarafından bağımsız olarak yapılmış sonrasında aralarındaki uyuma bakılmıştır. Çeviri sırasında bire bir çeviri yerine anlam bütünlüğü en doğru ifade eden cümleler kullanılmıştır.

SÖYLEŞİ SORULARI ve YANITLARI

1. ***Farklı kanıt türlerinin çoğalması ve farklı kanıt türlerinden hangisine öncelik vereceğimize ilişkin bir ölçütün olmaması geçerlik türlerine ilişkin yapılan sınıflandırmaların gerekliliği noktasında size ne düşündürtmektedir? Bu kapsamda, size göre en geçerli geçerlik tanımı ve sınıflandırması nedir?***

2. ***Geçerlik türlerinin tanımlamada farklı anlayışlar bulunmaktadır. Örneğin Standartların son versiyonunda geçerlik kanıtları elde etme yolları vardır. Kapsam, ölçüt dayanaklı, yapı geçerliği gibi türler yoktur. Bu ilgili tanımlamalara yönelik görüş ayrılığı ne gibi nedenlerden dolayı ortaya çıkmıştır?***

1-2. Bu iki soruyu birlikte cevaplayacağım. Geçerliğin en iyi tanımının Eğitimsel ve Psikolojik Testler için Standartların güncel baskısında verildiğini düşünüyorum. Güncel baskı 2014 yılında yayımlandı ve bu 6.baskıdır. Onlar geçerliği "kanıt ve kuramın testlerin amaçlanan kullanımları için test

[1] Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta:nurid@hacettepe.edu.tr
[2] Prof. Dr., University of Massachusetts Amherst , Amherst –Birleşik Devletler, e-posta: sireci@acad.umass.edu
3 Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta: burcua@hacettepe.edu.tr
[4] Arş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta: nkibrislioglu@hacettepe.edu.tr
[5] Arş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta: osman.tat@hacettepe.edu.tr
[6] Arş. Gör., İstanbul Medeniyet Üniversitesi, Edebiyat Fakültesi, İstanbul-Türkiye e-posta: cemm@medeniyet.edu.tr

puanlarının yorumlanmasını ne ölçüde desteklediği" şeklinde tanımlamaktadır (Amerikan Eğitimsel Araştırma Derneği, Amerikan Psikoloji Derneği ve Eğitimde Ölçme Ulusal Konseyi, 2014, p. 11). Bu tanım esasen 5. baskıda verilen tanımla aynıdır (AERA, APA, & NCME, 1999). Bu tanım önemlidir çünkü tanımda şunları görebiliriz;

- Geçerlik bir testin belli bir amacına veya kullanımına göre değerlendirilmelidir.
- Testler "doğası gereği" geçerli veya geçersiz değildir. Geçerlenmesi gereken (tabii ki, kuram ve uygulama ile desteklenerek) testin belli bir amaç için kullanımıdır.
- Geçerleme hem kanıt hem de kuramın bir testin belli bir amaç için kullanımını desteklemesini gerektirir.

Bir testin belirli bir amaçla kullanımını geçerlemek (validate) için yapılabilecek tek bir çalışma yoktur.

Geçerlik kanıtının farklı türleri vardır ve bir testin belli bir amaç için kullanımını desteklemek üzere kullanılan kanıt türleri testin amacına göre değişecektir. Standartların son iki baskısı "belli bir kullanım için test puanlarının amaçlanan yorumlanmasının geçerliğinin değerlendirilmesinde kullanılabilecek" (AERA ve diğerleri, 2014, p. 13) kanıtların beş kaynağını belirtir. Bu kaynaklar (a) test içeriğine, (b) yanıt süreçlerine, (c) içyapıya, (d) diğer değişkenlerle ilişkilere ve (e) testin sonuçlarına dayanan geçerlik kanıtlarıdır.

Standartların son iki baskısında (AERA ve diğerleri, 1999, 2014) tanımlanan geçerlik kanıtlarının bu kaynakları Standartlar'ın daha önceki baskılarında tanımlananlar ile aynı değildir.

Tablo 1'de Standartlar'ın farklı baskılarını ve geçerliğin farklı kategorilerini veya geçerlik kanıtı türlerini/kaynaklarını nasıl tanımladıklarını listeliyorum. Güncel baskının "geçerlik kanıt kaynaklarını" tanımladığına dikkat edin, çünkü geçerliğin farklı türleri yoktur. Önüne "yapı" kelimesini koysanız da koymasanız da geçerlik bütünsel bir kavramdır. Geçerliğin güncel tanımının (AERA ve diğerleri, 2014) ve geçerlik kanıtlarının beş kaynağının birçok nedenden dolayı geçerliği tanımlamanın en iyi yolu olduğunu düşünüyorum. İlk olarak, bunlar tek bir kişinin ürünü değildir. 60 yılı aşkın bir süredir, üç kuruluş geçerliğin ne anlama geldiği ve test puanlarının (kullanımlarının) nasıl geçerli kılınması gerektiği hakkında fikir birliğine varmak için birlikte çalışmaktadır. İkinci olarak, tanım geçerliğin, test kullanımına işaret ettiğini vurgulamaktadır ve bu geçerleme hem kuramsal doğrulama hem de deneysel kanıt gerektirmektedir. Bunlar reddetmesi zor, apaçık gerçeklerdir.

Tablo 1. Standartlar 'da Geçerlik Kanıtlarının Zaman İçindeki Sınıflandırılması

Yayım	Geçerlik Sınıflandırmaları
Psikolojik testler ve tanılayıcı yöntemler için teknik tavsiyeler: Bir ön öneri (APA, 1952)	Sınıflamalar: Yordama, durum, kapsam, Uyum (uygunluk)
Psikolojik testler ve tanılayıcı yöntemler için teknik tavsiyeler (APA, 1954)	Türler: Yapı, eşzamanlı, yordama, kapsam
Eğitimsel ve psikolojik testler ve kılavuzlar için standartlar (APA, 1966)	Türler: Ölçüt dayanaklı, yapı dayanaklı, kapsam dayanaklı
Eğitimsel ve psikolojik testler için standartlar (APA, AERA, ve NCME, 1974)	Yönler: Ölçüt dayanaklı, yapı dayanaklı, kapsam dayanaklı
Eğitimsel ve psikolojik testler için standartlar (AERA, APA, ve NCME, 1985)	Sınıflamalar: Ölçüt dayanaklı, yapı dayanaklı, kapsam dayanaklı
Eğitimsel ve psikolojik testler için standartlar (AERA, APA, ve NCME, 1999)	Kanıt Kaynakları: Kapsam, yanıtlama süreci, iç yapı, diğer değişkenlerle ilişkiler, testin sonuçları

3. Bir testin cevaplanma süreci geçerlik kanıtı olarak nasıl kullanılmaktadır? Geniş katılımlı sınavlarda cevaplama süreci geçerlik için kanıt olarak kullanılabilir mi?

Yanıtlama sürecine dayalı geçerlik kanıtları "yapı ve testi alan alanlar tarafından gerçekten ortaya konan performans ya da tepkinin ayrıntılı doğası arasındaki uyumla ilişkili kanıtlara" karşılık gelmektedir. (AERA ve diğerleri, 2014, p. 15). Bu kanıt türünün örnekleri katılımcılarla test sorularına verdikleri yanıtlar hakkında görüşme yapmayı, test maddelerine cevap veren katılımcıların sistematik bir biçimde gözlemlenmesini, puanlayıcıların performans görevlerini puanlarken kullandıkları ölçütlerin değerlendirilmesini, madde yanıtlama süresinin analizini (kronometrik analiz), öğrencilerin göz hareketlerinin takibini ve katılımcıların test maddelerini çözerken kullandıkları akıl yürütme sürecinin değerlendirilmesini kapsar. Bu kanıt(lar) özellikle testin üst düzey becerileri ortaya çıkarma derecesini ve farklı alt evrenlerden gelen öğrencilerin test maddelerini ne kadar iyi anladıklarını değerlendirmede oldukça kullanışlıdır. Eğitim alanındaki pek çok yeni testin üst düzey bilişsel becerileri vurguladığı düşünülürse, bu testlerin söz konusu yetenekleri yeterince ölçtüğüne dair kanıtlar gerekecektir.

4. Mutlak değerlendirme ile bağıl değerlendirme yapılan durumlarda geçerlik kanıtlarının toplanmasında farklı yollar izlenmeli midir?

Mutlak ve bağıl değerlendirmenin ne anlama geldiğini anlamadım. Ancak, bir "geçerlik argümanı" ortaya koymak için geçerlik kanıtlarının farklı kaynaklarının nasıl bir araya getirileceği hakkında konuşalım. Kane (1992, 2006, 2013), bir testin belirli bir amaç için kullanımının geçerlenmesinin, testin belirli bir amaç için kullanılmasının doğrulandığı sağlam ve mantıklı bir argüman geliştirmekten farksız olduğunu öne sürmektedir. Standartlar, geçerlik kanıtlarının beş kaynağının testin belli bir amaç için kullanımını desteklemek için tutarlı bir şekilde sentezlenmesi gerektiğini öne sürerek özellikle bu bakış açısını benimsemiştir. Örneğin, "Sağlam bir geçerlik argümanı, mevcut kanıtların ve kuramın belirli kullanımlar için test puanlarının amaçlanan yorumunu destekleme derecesini tutarlı bir açıklamayla birleştirir" diye belirmektedirler (AERA ve diğerleri, 2014, p.21). Geçerlemeye yönelik argüman-temelli yaklaşım, bir mahkeme salonunda bir testin kullanılmasını savunmaya benzerdir. Buradaki fikir, testin belli bir amaç için kullanılmasını destekleyecek çeşitli ve güçlü kanıtlar sunmaktır. Bu kanıtlar bütünü, testin planlanan hedeflerini karşılayan ve istenmeyen sonuçlar doğurmayan kanıtları içermelidir.

5. Bir çalışmada, geçerlik çalışması yapılabilecek yeterli örneklem yoksa geçerlik kanıtı toplamak için ne yapılabilir?

Testin kapsamına dayalı geçerlik kanıtları genellikle alan uzmanları kullanılarak elde edilir ve genellikle çok az sayıda uzmanın ürünüdür (örneğin, 10 veya daha az). Dolayısıyla, geçerlik çalışmalarının tamamen madde yanıtları ve test puanlarının analizine dayalı olmadığını lütfen hatırlayın. Kapsam geçerliği kanıtı eğitimle ilgili testlerde temeldir ve gereklidir. Yanıtlama süreçlerine dayalı geçerlik kanıtına dayanan çalışmaların çoğunluğu da çok küçük örneklem büyüklüğü kullanır.

6. Bir testin cevaplanma süreci geçerlik kanıtı olarak nasıl kullanılmaktadır? Geniş katılımlı sınavlarda cevaplama süreci geçerlik için kanıt olarak kullanılabilir mi?

Üçüncü sorunun cevabına bakınız.

7. Geçerlilik ile çoklu yöntem-çoklu özellik analizi arasındaki ilişki nasıl açıklanır?

Campbell and Fiske'yi (1959) okumanızı öneririm.

8. Boud (1995) ve Messick (1995), alternatif değerlendirme yöntemleri için sonucsal geçerlik kavramını gündeme getirmişlerdir. Değerlendirmenin öğrenme üzerindeki etkisine dayanan sonucsal geçerliğin tayininde hangi yollara başvurulmaktadır?

Messick "sonuçsal geçerlik (consequential validity)" terimini kullanmamıştır. Standartlar, "testin sonuçlarına dayalı geçerlik kanıtlarını" tanımlar. Testin sonuçlarına dayalı geçerlik kanıtı, bir test programı ile ilişkilendirilen kasıtlı (intended) veya kasıtsız (unintended) sonuçların değerlendirilmesini ifade eder. Testler, psikolojik rahatsızlıkların doğru tanılanması, kamunun korunması, öğretimin geliştirilmesi ve yapıların daha iyi anlaşılması gibi pozitif sonuçları arttırmak için kullanılır. Açıkça düşünülmemiş veya planlanmamış kasıtsız pozitif sonuçlar da ortaya çıkabilir. Ancak test programında kasıtsız negatif sonuçlar da ortaya çıkabilir. Belirli gruplar için eğitim ve iş olanaklarında azalmaya neden olan ters etki, okul bırakma oranında artış, test performansına dayalı olarak çalışan maaşları veya kaynak aktarımına ilişkin kötü kararlar kasıtsız sonuçların örnekleri olabilir.

Sonuca dayalı geçerlik kanıtı hesap verilebilirlik (örneğin okul veya öğretmenleri değerlendirmek için testlerin kullanılması) gibi bazı amaçlar için testlerin geçerlenmesi düşünüldüğünde özellikle önemlidir. ABD'de hesap verilebilirliği test etmek federal eğitim politikasınca gereklidir ve genellikle paydaşlar için beklenen sonuçların ana hatlarını belirten bir eylem kuramıyla birlikte gelir. Örneğin, öğrencilerin test sonuçlarına göre öğretmenleri değerlendirmek, amaçlanan müfredatı öğretmek için öğretmenleri teşvik eder ve bu daha çok odaklanmış öğretimin, öğrencilerin söz konusu müfredatı öğrenmelerini geliştireceği varsayılır. İstenen sonuçların gerçekleştirilme derecesi ve istenmeyen sonuçların (örneğin, düşük öğretmen morali, öğrencilerin öğrenmesini azaltacak şekilde müfredatın daraltılması) ne ölçüde en aza indirileceği, hesap verilebilirlik için eğitim alanındaki testlerinin geçerliğinin incelenmesinde önemlidir. Lise mezuniyeti veya üniversiteye giriş için bir test kullanmak gibi, diğer test amaçlarının da değerlendirilmesi gereken sonuçları vardır.

9. Alan yazına bakıldığında da geçerliğe dair farklı yaklaşımlar ve sınıflamalar bulunmaktadır. Durumu özelleştirmek adına hipotetik bir durum üzerinden soru yer almaktadır. -Üst düzey becerileri (ör; problem çözme, eleştirel düşünme) ölçmek istiyorsunuz. Test puanlarının sonucu kritik (high stakes) öneme sahip kararlarda kullanılacak. Geçerliği sağlamak adına testin geliştirilmesinden puanların anlamı ve kullanımına kadar olan süreçte nasıl geçerlik kanıtları sunarsınız?

Dördüncü sorunun cevabına bakınız.

10. Likert tipi ölçeklerde bireylere ait puanların hesaplanması, maddelere verilen tepkilerin sayısal değerlerinin toplanması ile gerçekleştirilmektedir. Fakat her maddenin ilgili yapıya hizmet etme derecesinin farklılık gösterdiği de açıktır. Bu bağlamda, bu gerçeğe rağmen puanların düz bir şekilde toplanması ne derece doğru ve bu puanlardan elde edilen bulguların/ yapılan çıkarımların geçerliliği ne derece yüksektir?

Geçerliğin testin belirli bir amaç için kullanımına karşılık geldiğini hatırlayınız. Genel olarak tek maddelik testler yoktur ve dolayısıyla çıkarımlar ve kararlar pek çok maddeden hesaplanan puanlara dayanarak yapılır. Bir veya iki maddeyle ilgili bir problem varsa bu durum, toplam test puanlarını ve yorumlamayı nasıl etkilediğine bağlı olarak, önemli olabilir de olmayabilir de. Geçerlik analizine madde düzeyi, toplam puan düzeyi ve alt puan düzeyi gibi farklı düzeylerdeki analizleri katmak iyidir. Eğer geçerlik kanıtlarının beş kaynağını dikkate alacaksanız, tüm düzeyler hesaba katılır. Madde analizleri ve değişen madde fonksiyonu içyapıya dayalı geçerlik kanıtının bir parçasıdır. Test kapsamına dayalı çalışmalarda olduğu gibi, boyutluluk analizleri de genellikle madde düzeyine odaklanır. Ancak diğer değişkenlerle ilişkiler (örneğin, yordama geçerliği, değişen yordama geçerliği, çoklu özellik-çoklu metot) toplam test puanı düzeyine odaklanacaktır. Değerlendirilecek olan belli geçerlik sorusu analizin düzeyini belirleyecektir.

11. ***Türkiye olarak sınav uygulamalarında daha çok test güvenliği noktasına odaklanmaktayız ve geçerlik adına ortaya konulan çoğu uygulamaları kurumlarımızın henüz uygulamaya koymadığını görmekteyiz. Diyelim ki Türkiye'de yapılan sınavlar ile ilgili yetkili bir komisyon sizinle çalışmak istedi. Geçerliği bilimsel ve sosyal değerler temelinde düşünerek hangi uygulamaları ne şekilde değiştirmek isterdiniz ve önerileriniz ne olurdu?***

Güvenirlik önemlidir ancak belirli bir amaç için geçerli olmayan testlerin yine de güvenilir olabileceğini hatırlamak daha önemlidir. Örneğin, bir üniversite giriş testi güvenilir puanlar üretebilir; ancak, bir matematik dersinde öğrencilere not vermek için kullanıldığında geçerli olmayacaktır. Sanırım önceki sorular cevaplarım test geçerleme sürecini nasıl ele alacağımı tarif ediyor.

12. ***Türkiye'de ÖSYM'nin (Ölçme-Seçme ve Yerleştirme Merkezi) uyguladığı çeşitli sınavlarda görme engelli adayların görme derecesine göre farklı yazı puntolarında kitapçıklar verildiği, görme yetisi olmayan adaylara ise görsel sorulardan sorumlu tutulmayarak daha az soru sorulduğu, bir okuyucunun desteği ve belirli bir ek süre ile sınava girdikleri bilinmektedir. Bunun yanında çeşitli özür durumları için rapor alan adaylar için de farklı uygulamalar bulunmaktadır (lavabo ihtiyacı gibi). Yukarıda belirtilen benzer durumlar göz önüne alındığında, bir testin geçerlik kanıtları normal şartlar altında mı düşünülmeli veya farklı özelliklere ait alt gruplar için farklı geçerlik kanıtları da aranmalı mıdır?***

Bu geçerlik sorusunun ne olduğuna bağlıdır. Geçerliğin testin belli bir amacına karşılık geldiğini hatırlarsak, testin standart test uygulamaları ve uyarlanmış test uygulamaları için benzer şekilde geçerli olup olmadığı bir soru olabilir. Bir diğer soru, Standart ve uyarlanmış test uygulamalarından elde edilen puanların karşılaştırılabilir olup olmadığı olabilir. Daha özel bir soru ise uyarlamanın ölçülen yapıyı değiştirip değiştirmediği olabilir. Bu alanlarda oldukça geniş bir alanyazın ve uygulamalı araştırma bulunmaktadır.

13. ***Yordama geçerliğindeki ranj daralması sorununa istatistiksel yaklaşımlarla çözüm bulunabilir mi?***

Test verisi değerlendirilirken ranjın daralması gerçekten önemlidir. Muhtemelen değişen ranj daralmasından kaynaklanan gruplar arası değişmezliğin (hem madde hem de toplam test puanı düzeyinde) olmadığı sonucuna varan birçok çalışma gördüm. Madde çift serileri, güvenirlik katsayıları, faktör yükleri vb. gibi istatistikler değişkenliğe ihtiyaç duyar. Eğer bir test bir grup için çok kolay ya da çok zorsa, bu bir yanlılık kaynağı gibi görünecektir ancak bu gerçekte sadece ranj daralmasının yapay etkisidir.

Bu problemle başa çıkmanın bir yolu, sınırlandırılmış grubun dağılımı ile örtüşecek şekilde, sınırlandırılmamış gruptan örneklem çekmektir. Bu strateji değişen ranj daralmasını kontrol altına alır. Tabii ki, ranj daralmasının hafifletilmesi bunu yapabilecek veriye sahipseniz ve uygun düşüyorsa kullanışlıdır.

14. ***Messick (1995) geçerliğin, hem kanıt toplama hem de puan yorumlarının sonuçları ve kullanımı olarak geniş bir biçimde tanımlanabileceğini belirterek kapsam ve ölçüt dayanaklı geçerliği, yapı geçerliğinin alt bölümleri olarak görmüştür. Messick'in birleştirilmiş (unitary) geçerlik kavramı, testlerin geçerliği için toplanan tüm kanıtların testin yapı geçerliğini ortaya koyacağını belirtmektedir. Birleştirilmiş geçerlik kavramı, testin neyi ölçtüğüne dair basit soruları yanıtlamadığı, daha çok nomolojik ağ ile açıklanan karmaşık puan yorumlamalarıyla ilgili olduğu yönünde eleştiriler almıştır. Farklı geçerlik türlerinin farklı amaçlara hizmet ettiği düşünüldüğünde geçerliğe ilişkin (yapı geçerliğinden farklı olarak) bir çatı kavram oluşturmak mümkün müdür?***
15. ***Messick (1995) geçerliğin, hem kanıt toplama hem de puan yorumlarının sonuçları ve kullanımı olarak geniş bir biçimde tanımlanabileceğini belirterek kapsam ve ölçüt dayanaklı***

geçerliği, yapı geçerliğinin alt bölümleri olarak görmüştür. Buna göre kapsam ve ölçüt dayanaklı geçerlikten bağımsız olarak yapı geçerliğini araştırmak ne derece doğru olur? bir ölçümün geçerliği her zaman bir bütün olarak mı araştırılmalıdır? Öncelikle hangi geçerlik türüne bakmak gerekir?

14 ve 15. Lütfen bu iki soru için Sireci (2012)'ye bakınız.

KAYNAKÇA

APA (1952). *Committee on Test Standards. Technical recommendations for psychological tests and diagnostic techniques:* preliminary proposal. American Psychological Association Washington DC US. [American Psychologist, 7(8), 461-475. http://dx.doi.org/10.1037/h0056631]

APA (1954). *Technical recommendations for psychological tests and diagnostic techniques.* American Psychological Association Washington DC US. [Psychological bulletin, 51, 2, pt. 2, March 1954. Supplement.]

APA (1966). *Standards for educational and psychological tests and manuals.* American Psychological Association Washington DC US. [Educational and Psychological Measurement, 26, 3, 751-767 October 1, 1966 DOI: https://doi.org/10.1177/001316446602600328]

APA, AERA, & NCME (1974). *Standardsfor eductiond andpsychological tests and manuals*. Washington, DC: American Psychological Association.

AERA, APA & NCME (1985). *Standards for education and psychological testing*. Washington, DC: American Psychological Association.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Campbell D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin, 56*(2), 81−105.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. http://dx.doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17– 64). Westport: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749

Sireci, S. G. (2012). De-"Constructing" Test Validation. *Center for Educational Assessment Research Report No. 814*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.