Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
ISSN: 1309 - 6575

Geçmişten Günümüze Geçerlik

Validity from Past to Present

Hülya KELECİOĞLU * Sakine GÖÇER ŞAHİN **

Öz
Psikolojik testler kullanılmaya başlandığından bu yana testler için geçerlik kavramı önem kazanmıştır. Geçerlik, bir testin ölçmek istediği özelliği, testin amacına uygun olarak ölçme derecesi ile ilgilidir. Geçerlik kavramının ortaya çıktığı yıllardan günümüze kadar geçerliğin tanımı, sınıflandırılması, sonuçların yorumlanması üzerine yapılan tartışmalar devam etmektedir. Bu makalede geçmişten günümüze geçerlik ile ilgili yapılan tartışmalar ele alınmıştır. Başlangıçta ölçüt dayanaklı geçerlik temelinde olan geçerlik sınıflamaları, daha sonra psikolojik yapılarla ilişkilendirilerek yapı geçerliği temeline doğru kaymıştır. Bu yaklaşımda tüm geçerlik türleri yapı geçerliği altında birleştirilmiştir. Birleştirilmiş geçerlik olarak adlandırılan bu yaklaşım, testlerin geçerliği için toplanan tüm kanıtların testin yapı geçerliğini ortaya koyacağını belirtmektedir. Bu görüşe karşı çıkanlar ise, daha çok teorik bir çalışma olan yapı geçerliğinin eğitimde kullanılan testlerin geçerliğini tanımlamada yeterli olmayacağını ileri sürmüşlerdir. Günümüzde hem birleştirilmiş geçerlik yaklaşımını savunan hem de buna karşı çıkan görüşler bulunmaktadır ve kapsam geçerliğinin yapı geçerliği altında ele alınmasının yaratacağı sorunlar üzerindeki tartışmalar hala sürmektedir. Ancak nihayetinde geçerliğin bir kanıt toplama süreci olarak ele alınması konusunda uzlaşmaya varıldığı görülmektedir.

Anahtar Kelimeler: geçerlik, ölçüt dayanaklı geçerlik, yapı geçerliği, kapsam geçerliği, birleştirilmiş geçerlik

Abstract
The concept of test validity has gained importance since psychological tests have been used. Validity is related to the degree of measurement capability of a certain test, which intends to measure a desired property. However, from the years which the concept of validity emerged to the present day debates about definition and classification of validity and the interpretation of its results are on-going. In this study all debates about validity from the past to the present day are examined. Originally the classification of validity was the basis of criterion referenced validity, since then, the basic psychological structure has shifted to be associated with construct validity. In this approach all types of validity were unified under construct validity. Unified validity, as this approach is called, indicates that all evidence for validity of test data revealed construct validity. According to opponents of this view, construct validity, which is more a theoretical study, would not be sufficient to describe the actual validity of the tests used in education. Today their are views defending the unified validity approach as well as views against it and debates about the problems of holding onto content validity under the umbrella of construct validity still continue to this day. At the conclusion of all these debates, it seems that there is consensus that validity is an important evidence gathering process.

Key Words: validity, criteria referenced validity, construct validity, content validity, unified validity

GİRİŞ

Psikolojik testlerin kullanılmaya başlamasından bu yana, bu testlerin amaçlarına hizmet etme düzeylerini, yani geçerliğini belirlemek konusu önem kazanmıştır. Geçerlik kavramının, bir testin ölçmek istediği özelliği amacına uygun olarak ölçme derecesi ile ilgili olduğu bilinmektedir. Ancak geçerliğin tanımı, sınıflandırılması, belirlenmesi, geçerliği araştırılan test puanlarının anlamı ve yorumlanması üzerinde henüz bir görüş birliğine varılamamıştır.

* Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-TÜRKİYE, e-posta: hulyaebb@hacettepe.edu.tr
** Araş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-TÜRKİYE, e-posta: sgocersahin@gmail.com

Geçerliğin klasik tanımı, bir testin ölçmek istediği değişkeni onu başka değişkenlerle karıştırmadan ölçebilme derecesidir (Thorndike ve Hagen, 1961; Turgut ve Baykul, 2013). Geçerlik, aynı zamanda test puanlarından çıkarılacak sonuçların doğruluk derecesini de gösterir. Cronbach (1984) geçerliği, test puanlarından elde edilecek çıkarımları desteklemek için kanıt toplama süreci olarak tanımlamaktadır.

Test puanlarından elde edilen çıkarımların neler olabileceğine ilişkin pek çok sınıflama yapılmıştır (Cureton, 1951; Cronbach ve Meehl, 1955; Cronbach, 1984; Hopkins, Stanley ve Hopkins, 1990; Messick, 1995; Murphy ve Davidshofer, 2001; Lissitz ve Samuelsen, 2007b; Embretson, 2007). Amerikan Psikologlar Birliği tarafından 1954 yılında yapılan ve günümüzde de sıkça kullanılan geçerlik sınıflaması, dört tür geçerliği içermektedir: yordama geçerliği, zamandaş geçerliği, kapsam geçerliği ve yapı geçerliği (Cronbach ve Meehl, 1955; Murphy ve Davidshofer, 2001). Bu sınıflama 1966 yılında kapsam geçerliği, ölçüt dayanaklı geçerlik ve yapı geçerliği olmak üzere dörtten üçe indirilmiş ve geçerliğin bugün de tartışılan *sonuçların yorumlanması* boyutu üzerinde ilk tartışmalar başlamıştır (Kane, 2006).

Uzun yıllardan bu yana kullanılan geçerlik sınıflaması üzerinde tartışmalar devam etmektedir. 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 kavramına da itirazlar olmuş, bu yaklaşımın pratik olmadığı, bütünselleştirmeye gerek olmadığı ve eğitimde başarıyı ölçen testler için uygulanmasının güç olduğu iddia edilmiştir (Brennan, 1998; Borsboom ve ark., 2004; Lissitz ve Samuelsen, 2007b; Embretson, 2007).

Bu çalışmada öncelikle geçerlik kavramının kısa bir tarihçesine değinilmiş, daha sonra da günümüzde geçerlik kavramının yapılandırılmasına ilişkin tartışmalara yer verilmiştir.

GEÇERLİK KAVRAMI

Geçerliğe ilişkin ilk tartışmaların çoğu bilimin realist felsefesine dayanmaktadır. Bu yaklaşımda, bireylerin incelenen özelliklerine ilişkin ölçülen değişkenin kesin bir değerinin olduğu kabul edilir. Ölçmenin amacı da bu değeri olabildiğince doğru olarak kestirmektir. Geçerlik ise, bu kestirimlerin doğruluğu olarak tanımlanmaktadır. Geçerlik bu şekilde tanımlanınca, ölçülen değerin o değişkenin *gerçek* değeri olup olmadığını ya da *gerçek* değere ne kadar yakın olduğunu kanıtlama gereği ortaya çıkmıştır. Bu durum, ölçülen değerin kıyaslanabileceği bir ölçüt ihtiyacını doğurmuştur. İlgilenilen özelliğin değeri olarak bir ölçüt ölçüsü belirlenmiş ve bu ölçüte göre doğru bir kestirimde bulunan testin geçerli olduğu kabul edilmiştir (Kane, 2001).

Ölçüt dayanaklı geçerlik

Geçerlik kavramının ilk tartışılmaya başlandığı 1915'li yıllarda formal bir geçerlik tanımı olmamakla beraber, o yıllarda geçerlik, günümüzde *ölçüt dayanaklı geçerlik* adını verdiğimiz kavram üzerinde temellendirilmişti (Lissitz ve Samuelsen, 2007a). Psikolojik testlerden elde edilen puanların, bu puanların yordayıcısı olduğu düşünülen bir ölçüt ile olan ilişkisi geçerlik olarak adlandırılmakta idi.

Cureton (1951) testin amacına hizmet etme derecesi olarak tanımlanan geçerliğin, ölçüt dayanaklı model ile en uygun şekilde belirlenebileceğini öne sürmüştür. Ölçüt dayanaklı modelde, testin uygulandığı gruptan elde edilen puanların yine uygulama yapılan gruptan elde edilen ve testin ölçtüğü özelliklerin bulunduğu (okul performansı, iş performansı vb.) bir başka değişkenle ilişkisine bakılmıştır. Burada test puanlarının geçerliği araştırılmaktadır ve geçerlik kanıtı da iş ya da okul performansının test puanlarıyla olan ilişkisidir. Bu modele

dayanarak Cureton (1951) geçerliği, *toplam test puanları ile gerçek ölçüt puanları arasındaki korelasyon* olarak tanımlamıştır.

Ölçüt dayanaklı model, uygun bir ölçüt ölçüsü bulunması durumunda pratikte oldukça kullanışlıdır. Örneğin bir işveren, uygun bir ölçüt kullanarak işe alacağı elemanları seçebilir. Ölçüt dayanaklı geçerlik, geçerli ve güvenilir bir ölçüt bulmanın zorluklarını taşımakla beraber, uygun bir ölçüt bulunduğunda uygulayıcılara kullanışlı bilgiler vermektedir.

Geçerlik çalışmaları 1930'lu ve 1940'lı yıllarda günümüzde yordama geçerliği adı verilen çalışmalardı. Bu dönemde yeni bir tür ölçüt dayanaklı geçerlik kavramı gündeme gelmiştir: *zamandaş geçerlik*. Zamandaş geçerlik de yordama geçerliği gibi bir ölçüt ile korelasyona dayanmaktadır. Ancak yordama geçerliğinde ölçüt, geçerliği araştırılan testten alınan puanlardan bir süre sonra elde edilirken zamandaş geçerlikte test puanları ile aynı zamanda ya da ondan önce elde edilmektedir. Geçerliğin bu iki formu testin genel amacını da yansıtmaktadır. Testin ölçtüğü özellikler, bazı dışsal davranışlarla ilişkili ise ya da onları yordamak amacıyla geliştirilmiş ise geçerlik çalışması, bu amaçlara ulaşılıp ulaşılmadığını belirlemek için yapılır (Lissitz ve Samuelsen, 2007a). Ölçüt dayanaklı geçerlik, testlerin ilişkili olması beklenen dışsal davranışlarla ilişkisini araştırır. Eğer testin, bu özelliklerle ilişkisini ortaya koymak yeterli ise bu durumda zamandaş geçerlik çalışması yapılır. Ölçüt, yani testin ilişkili olması beklenen davranışlar, test puanları ile aynı anda ya da daha önce elde edilebilir. Ancak testin amacı gelecekteki bazı davranışları tahmin etmek ise bu durumda ölçüt, test puanlarından daha sonra elde edilir. Bu durumda ise testin yordama geçerliğine ilişkin kanıt toplanmış olur.

Ölçüt dayanaklı geçerlik, ancak iyi tanımlanmış ve geçerli bir ölçüt bulunduğunda doğru sonuçlar verir. Bulunan ölçütün geçerli olup olmadığını incelemek, başka bir ölçütün varlığı ihtiyacını doğurmaktadır. Ölçütün geçerliğini irdelemek için bulunan başka ölçütlerle oluşturulan *ölçüt ölçütü* zinciri ile bu model kısır bir döngüye dönüşebilir. Uygun bir ölçütün bulunmaması ya da eldeki ölçütlerin yeterli derecede uygun olduğunun kanıtlanamaması ölçüt dayanaklı geçerliğin pratikteki yararlarını da azaltmaktadır. Ölçüt dayanaklı geçerlik çalışmaları, bir ölçütün geçerli olup olmadığını değerlendirmez ve geçerli bir ölçüt sağlama konusunda da herhangi bir yol sunmaz. Elde edilen sonuçlar, ölçütün geçerliliği oranında genellenebilir. Bu da ölçüt dayanaklı geçerlik çalışmalarının en önemli problemidir.

Kapsam geçerliği

Ölçüt dayanaklı geçerliğin bütün testler için uygun bir geçerlik türü olmadığı 1940'lı yıllarda tartışılmaya başlandı. Rulon (1946), testlerin geçerliği için dışarıdan bir ölçüte gerek olmadığını, bireylerin kazandığı bilgi ve beceri ölçüsünün (test puanlarının) ölçütün kendisi olduğunu düşünüyordu. Testin geçerliğini ortaya koymak için gerekli olan ölçütün yine testin kendi içerisinde olduğunu savunuyordu. Bunun sonucunda, testlerin amaçlanan konu alanı ve bilişsel süreçleri temsil edip etmediğine ilişkin uzman görüşlerinin alınması gereği ortaya çıktı. Cureton (1951) ilk kez *kapsam geçerliği* terimini kullanarak, kapsam geçerliğini ölçüt kavramı ile ilişkilendirerek açıklamıştır. Bu yaklaşımda ölçüt, alan uzmanları tarafından oluşturulan standartları ifade etmektedir. Örneğin, bir meslek alanına ilişkin bilgilerin ölçüleceği bir test hazırlarken, önce uzmanlardan o meslekle ilgili kavramları listelemesi ve bunları önem derecesine göre sıralaması istenir. Daha sonra uzmanlar, mesleğe ilişkin becerileri ölçmek için hazırlanan deneme testini değerlendirir. Uzmanların arasındaki uyum, kapsam geçerliğinin bir ölçüsü olarak alınabilir. Geçerlik, bu biçimi ile okullarda olduğu kadar endüstride ve çeşitli kurumlarda hazırlanan testlerin iş ve görev analizi ile tutarlı olup olmadığını değerlendirmek amacıyla da kullanılmaktadır (Lissizt ve Samuelsen, 2007a).

Cronbach ve Meehl (1955), kapsam geçerliğini, test maddelerinin, ölçülmek istenen alanın (konu ve davranış) bir örneklemi olduğunun gösterilmesi olarak tanımlamıştır. Bu yaklaşıma göre testin kapsam geçerliğini değerlendirmek için (1) her bir maddenin yazıldığı

kapsam alanını ne kadar temsil ettiği ve (2) maddelerin tümünün kapsam alanlarını ne kadar temsil ettiğine ilişkin uzman görüşleri alınır (Thorndike ve Haggen, 1977). Kapsam geçerliği ayrıca, (a) test içeriğinin ve test maddelerine verilen cevapların ölçülen alana uygunluğu, (b) testin kullanılacağı alanın açıkça tanımlanması ve (c) maddelerin teknik özelliklerinin uygunluğunun görgül ve mantıksal süreçlerle ortaya konulması olarak da tanımlanmıştır (Fitzpatrick, 1983). Bu yaklaşımlarda kapsam geçerliğini belirlemek için uzman görüşlerine *mantıksal* ve *görgül* (korelasyonel) tekniklere başvurulmaktadır. Kapsam geçerliğini belirlemede teste değil de test maddelerine verilen cevaplara odaklanmak ise değişen madde fonksiyonu ve faktör analizi gibi kavramların kapsam geçerliği çatısı altında incelenmesine zemin hazırlamıştır (Lissizt ve Samuelsen, 2007a).

Ancak görgül olmayan tanımlar üzerinde bir görüş birliği oluşamamış ve görgül yollarla daha teorik olarak yapılan geçerlik tanımları ağırlık kazanmaya başlamıştır. Örneğin, 1946'da Guilford pratik ve faktöriyel geçerlik olmak üzere iki geçerlik türünden bahsetmiştir. Pratik geçerlik, ölçüt ile ilgili iken faktöriyel geçerlik faktör yükleri ile ilgilidir. Cronbach'da (1984) iki tür geçerlikten söz etmiş ve bu geçerlik türlerini kullanılan analize göre isimlendirmiştir. Analizleri ise *pratik* ya da *yargısal* olarak sınıflandırmıştır.

Yapı kavramı ve yapı geçerliği

Geçerlik tanımı 1950'li yıllarda APA tarafından psikolojik ve klinik çalışmaları da kapsayacak şekilde genişletilmiştir. Yapı geçerliği kavramı ilk olarak Meehl ve Challman'ın da içinde yer aldığı küçük bir çalışma grubunca kullanılmıştır. Bu kullanım tüm çalışma komisyonunca kabul görüp genişletilmiş ve 1954 yılında yayınlanan "*Teknik Öneriler*" kitabında yer almıştır. Daha sonra Cronbach ve Meehl (1955) bu kavramı daha da geliştirmiştir.

Cronbach ve Meehl (1955), özellikle 1950'lerde yoğun bir ilgi gören hipotetik tümdengelim (hypothetico deductive-HD) modelini teorik yapıların açıklanmasında kullanmışlardır. HD modelde teoriler, yorumlanmış aksiyomatik sistemler olarak ele alınmaktadır. Aksiyomlar, gözlenebilir değişkenlerle aksiyomların tanımladığı koşullar arasında bağlantı kurularak yorumlanmaktadır. Aksiyomlar yorumlandıktan sonra, değişkenler arasındaki gözlenen ilişkileri yordamada kullanılabilir. Değişkenler arasındaki bu görgül ilişkiler bir teoriyle açıklanır. İşte bu modele dayanarak oluşturulan teori, *nomolojik ağ* teorisi olup aksiyomatik sistemler ve buradan çıkarılacak tüm görgül kuralları içermektedir. Görgül yasaların verilerle test edilmesiyle teorinin geçerliği sağlanmış olur.

Cronbach ve Meehl (1955) psikolojik yapıyı, *bireylerin var olduğu kabul edilen özellikleri* olarak tanımlamaktadır. Lord ve Novick (1968) doğrudan gözlenemeyen yapılar için iki ayrı tanım yapılması gerektiğini belirtmiştir: (1) Yapıyı ölçmek için gerekli olan *işevuruk* tanımlar ve (2) yapının, kuramsal olarak diğer yapılarla ve dış ölçütlerle ilişkisini ortaya koyan *sözdizimsel (syntactic)* tanımlardır. İşevuruk tanımlar bir yapının tanımlanması için gereklidir, ancak yeterli değildir. Yapının işevuruk tanımının yanı sıra diğer değişkenlerle olan ilişkisini de tanımlayan sözdizimsel tanımlar gereklidir. Böylece bir yapı hem kendine özgü özellikleri, hem de dış dünyada diğer değişkenlerle ya da yapılarla ilişkisi açısından tanımlanmış olur.

Cronbach ve Meehl (1955) yapı geçerliğinin, kapsam ve ölçüt dayanaklı geçerliğe bir alternatif olabileceğini, böyle kabul edilmese bile onlarla eşit düzeyde bir geçerlik türü olduğunu ortaya atmışlardır. Bir testin ölçüt ve kapsam geçerliği sağlandıktan sonra hangi psikolojik yapıyı ölçtüğünün de belirlenerek, testin yapı geçerliğinin de dikkate alınması gerektiğini belirtmişlerdir. Yazarlar, geçerliği araştırılan test için işevuruk olarak tanımlanan bir ölçüt olmadığında, yapı geçerliğinin araştırılabileceğini belirtmiş, yapı geçerliğinde odak noktası olarak ölçüt puanları yerine testin ölçtüğü özelliği almışlardır. Yapı geçerliğini tanımladıktan sonra, yapıların nasıl ölçüleceğini ve yapıların birbirileri ile olan ilişkisinden

oluşan ağ örgüsünün doğasını tanımlayan *nomolojik ağı* açıklamışlardır. Buna göre ölçülen yapı ile diğer yapılar arasındaki ilişkiye dayanan ve teorik olarak tanımlanan nomolojik ağ olmadan yapı geçerliğinin varlığından söz etmek mümkün değildir. Ancak, yazarlar nomolojik ağın işevuruk olarak nasıl tanımlanacağını belirtmemişlerdir. Campbell ve Fiske (1959), değişkenler arasındaki ilişkileri tanımlamak için çoklu yöntem çoklu özellik matrisi yöntemini geliştirmişlerdir. Yapının tanımlanabilmesi için yakınsak ve ıraksak kanıtların birlikte incelenmesi gerektiğini; yöntemden kaynaklanan ve yapıyla ilgisi olan varyansla ölçme işlemleri ile ilişkili varyansın ayrıştırılmasını önermişlerdir.

1966'da yayınlanan *Standartlar*'da (APA, AERA ve NCME, 1966) yapılan yapı geçerliği tanımı ile yapı geçerliği, diğer geçerlik yaklaşımlarından, özellikle ölçüt dayanaklı yaklaşımdan ayrılmıştır (Akt. Kane, 2001). Yapı geçerliği, test ile ölçülmek istenen yapıyı daha iyi anlamak için işe koşulmakta ve kesin bir ölçütün olmadığı durumlarda uygulanmaktadır. Yani, Cronbach ve Meehl'den (1955) on yıl sonra *Standartlar*'da yapı geçerliği, kapsam ve ölçüt dayanaklı geçerliğe ek bir model olarak ele alınmıştır. Bu yaklaşımda, kapsam ve ölçüt dayanaklı modellerin yapı geçerliği altına dahil edilmesi ya da bunların tamamen saf dışı bırakılmasına ilişkin bir öneri sunulmamıştır. Yapı geçerliği altında daha açıklayıcı ve teorik yorumlara odaklanılmıştır.

1974 yılında yayınlanan *Standartlar*'da (APA, AERA ve NCME, 1974) bu fikrin etkileri devam etmiş ve dört bağımsız geçerlik türü tanımlanmıştır: yordama geçerliği, kapsam geçerliği, zamandaş geçerliği ve yapı geçerliği. *Standartlar*'daki (1974) yapı geçerliği tanımı Cronbach ve Meehl'in (1955) tanımına oldukça benzerdir. Buna göre psikolojik yapı, var olan bilginin bazı açılardan organize edilmesi ve açıklanması için geliştirilmiş ya da yapılandırılmış teorik bir imgedir. Örneğin kaygı, okuma hazır bulunuşluğu ya da yazma becerisi gibi kavramlar birer yapıdır ve bunlar değişkenler arasındaki ilişkiler ağından çıkarılacak boyutlardır.

Yapı geçerliği çalışmaları, yapıların doğrudan ya da dolaylı olarak nasıl ölçüleceğinin belirlenmesi, yapıyı açıklayan kurama dayalı hipotezler kurularak bunların test edilmesi, yapıyı tanımlayan ve açıklayan nomolojik ağların kurulması ve bunların yorumlanmasına dayalı süreçlerden oluşmaktadır (Cronbach ve Meehl, 1955). Nomolojik ağ birbirlerine aksiyomlarla bağlı olan ve bir yapıyı tanımlama amacı taşıyan örüntülerdir. Nomolojik ağ, yapının ölçülmesi ve gözlenebilir davranışlar aracılığı ile somut hale getirilebilmesi temeline dayalıdır. Yapının gözlenebilir hale getirilme süreci üç adımı içerir (Murphy ve Davidshofer, 2001):

1. Ölçülen yapı ile ilişkili olan davranışların belirlenmesi,
2. Ölçülen yapı ile ilgili olan ya da olmayan diğer yapıların ortaya konması,
3. Ölçülen yapı ile ilişkili olan diğer yapıları açıklayan davranış dizgesinin ortaya konması.

Birleştirilmiş Model Olarak Yapı Geçerliği

Loevinger'ın (1957) yordama, zamandaş ve kapsam geçerliğinin yapı geçerliğinin çatısı altında toplanması görüşü 1970'lerin sonlarından itibaren kabul görmüştür. Buna göre yapı geçerliği bir geçerlik türü değil, geçerlik için kanıt olabilecek tüm kapsam, yordama, zamandaş geçerliği ve güvenirlik süreçlerini içeren geniş bir modeldir (Kane, 2001). Bu görüşün temelleri Cronbach ve Meehl (1955) tarafından ortaya atılmış olup Loevinger (1957) tarafından açıkça ifade edilmiştir. Messick (1975, 1988, 1989) ise yapı geçerliğinin, geçerlik için genel bir çerçeve olarak ele alınmasını sağlamıştır. Birleştirilmiş geçerlik yaklaşımlarının temeli, ölçmenin gözlenemeyen yapıları gözlenen puanlar aracılığıyla yorumlanmasına dayanmaktadır. Ölçmenin amacı, gözlenemeyen yapıları açıklamak olduğuna göre yapı geçerliği de geçerliğin temeli olmalıdır. Diğer geçerlik türleri yapı geçerliği çatısı altında yer

almalı, kapsam ya da ölçüt dayanaklı incelemeler de yapı geçerliğinin bir parçası olarak değerlendirilmelidir.

Messick'e (1995) göre geçerlik, güvenirlik, karşılaştırılabilirlik ve adil olma sadece ölçmenin temel ilkeleri değil, aynı zamanda ölçmenin dışında karar vericilerin de kullandığı sosyal değerlerdir. Önemli bir sosyal değer olarak geçerliğin hem bilimsel hem de politik önemi vardır. Bu nedenle geçerliğin, sadece ölçüt ve yordayıcı arasındaki basit korelasyonla veya test içeriği hakkındaki uzman kanılarıyla açıklanamayacağını iddia etmektedir.

Messick (1995), geçerliğin hem kanıt toplama hem de puan yorumlarının sonuçları ve kullanımı olarak geniş bir biçimde tanımlanması gerektiğini ifade etmiştir. Geçerliğin bu kapsamlı tanımı; kapsam, ölçüt, puanların anlamı ve kullanımı ile ilgili hipotezleri test eden süreçleri yapı geçerliği altında birleştirmektedir. Yapı geçerliği, test puanlarının anlamı veya yorumlanmasına dayalı kanıtların (kapsam ve ölçüt geçerliği ile ilgili kanıtlar da yapı geçerliğinin alt bölümleridir) bütünleştirilmesinden oluşmaktadır. Messick'in görüşleri, Cronbach ve Meehl (1955) ve Loveinger (1957) tarafından öne sürülen ve yapı geçerliğini merkeze alan görüşlerle benzerlik göstermektedir. Bu yazarlar yapı geçerliğini merkeze alsalar da, geçerliğin farklı türleri, farklı yönleri ya da kategorileri gibi çeşitlenmesine itirazları da olmamıştır (Sireci, 1998). Ancak Messick (1975, 1988, 1989), diğer geçerlik türlerinin tümünü yapı geçerliği altında birleştirerek daha katı bir yaklaşım izlemiştir.

Sireci (1998), Messick (1989) tarafından kapsam geçerliğinin psikometri literatüründe karşılığı olmadığı şeklindeki görüşüne karşı çıkmış ve Messick'in bu görüşünün geniş bir kabul görmediğini belirtmiştir. Sireci'ye (1998) göre, kapsam geçerliğini belirlemek oldukça karmaşık bir süreçtir ve test puanlarının sadece sayısal verilerle değerlendirilmesinin önüne geçmek için de kapsam geçerliğine önem vermek gerekir. Sireci (1998) yapı geçerliğinin kapsam geçerliğine ilişkin bilgi vermediğini ifade etmektedir. Çünkü yapı, gözlenemeyen bir değişkendir ve ancak test puanları aracılığı ile yapının ölçülüp ölçülemediği ortaya konabilir. Öte yandan kapsam gözlenebilir ve test belirtke tablosu ile işevuruk olarak tanımlanabilir. Bu nedenle yapı geçerliği ve kapsam geçerliği ayrı olarak ele alınmalı ve değerlendirilmelidir. Bir testin örtük özellikleri, faktör analizine dayalı yöntemlerle tanımlanarak yapı açıklanabilir. Ancak yapının tanımlanması ve doğrulanması, kapsamın da tanımlanabildiği ve testin kapsam geçerliğinin olduğu anlamına gelmemektedir. Benzer şekilde, kapsam geçerliği için yapılan çalışmalar da testin yapısını açıklama konusunda yeterli olmamakta; özellikle okullardaki öğrenmeleri ölçen testlerin kapsam geçerliği, testin yapısının belirlenmesinin önüne geçmektedir. Kapsamın tanımlanması, sınırlandırılması ve temsil ediciliği okullarda kullanılan testlerin geçerliği için çok önemli kanıtlardır.

Borsboom, Mellenbergh ve Van Heerden (2004) birleştirilmiş geçerlik kavramının, testin neyi ölçtüğüne dair basit soruları yanıtlamadığını, daha çok nomolojik ağ ile açıklanan karmaşık puan yorumlamalarıyla ilgili olduğunu belirtmektedir. Yine bu yazarlar, "birleştirilmiş geçerlik kavramına karşı çıkıyoruz çünkü birleştirilecek bir şeyin olmadığını düşünüyoruz" şeklindeki açıklamaları ile birleştirilmiş geçerlik kavramının gereksizliğine vurgu yapmaktadırlar. Yapılan bu eleştiriler birleştirilmiş geçerlik modelinin eğitim alanında kullanılan testlere uygun düşmediği noktasına doğru genişlemiştir. Eğitimde kullanılan testlerin geçerlik kanıtlarını elde etme sürecinde birleştirilmiş geçerlik kavramının pratik bir yararının olmadığı, yapı geçerliğinin uygulamalarda puanların yorumlamasını kolaylaştırmadığı savunulmaktadır (Brennan, 1998; Lissitz ve Samuelsen, 2007a). Crocker (2003) eğitim alanında kullanılan testler için geçerliğin bel kemiğinin yapı geçerliğinden ziyade kapsam geçerliği olduğunu ifade etmiştir.

Lissistz ve Samuelsen (2007a) tarafından eğitimde kullanılan testlerin değerlendirme sürecine ilişkin bir sınıflandırma sunulmuştur. Bu sınıflandırmada değerlendirme sürecinin odak noktası ve perspektifi incelenmiştir. Değerlendirme sürecinin odak noktası içsel ve dışsal; değerlendirmenin perspektifi de kuramsal ve uygulamalı olarak ele alınmıştır. Bu

sınıflandırmada kapsam geçerliği, testin güvenirliği ve test puanlarının etkisi uygulama boyutunda yer alırken, yapı ve ölçüt geçerliği kuramsal boyutta yer almaktadır. Yazarlar, eğitimdeki ölçmeler için kapsam geçerliğine odaklanmakta ve dışsal özelliklere bağlı olmaksızın testin kapsam geçerliğinin, testin, diğer testlerden, nomolojik ağlardan ve amacından bağımsız olarak tanımlanması gerektiğini savunmaktadırlar. Testin içsel ve dışsal özellikleri ile kuramsal ve uygulamalı çalışmalar farklı sorulara yanıt vermektedir. Eğer bu sorular birbirinden farklı ise bütün bu kavramları birleştirmenin gereksiz olduğu düşüncesiyle, birleştirilmiş geçerlik kavramına itiraz edilmiştir.

Lissistz ve Samuelsen (2007a) tarafından ortaya konan modelde yer alan temel sorun, testin amacından bağımsız olarak kapsam geçerliğinin nasıl değerlendirileceğidir. Sireci (2007), birleştirilmiş geçerlik kavramının, yapı geçerliğini merkezine alması ve yapının çok iyi anlaşılmaması nedeniyle bir kenara bırakılmasını reddetmektedir. Birleştirilmiş geçerlik kavramının bir takım eksikleri bulunmasına rağmen, geçerlik sürecinin yalnızca içsel özelliklere bağlanarak kapsam geçerliği üzerinde durulmasının doğru olmayacağını belirtmektedir. Sireci'ye (2007) göre de testin geçerliğini, testin amacından bağımsız değerlendirmek mümkün değildir. Bir testin kapsamı (örneğin 6. sınıf matematik testi) uzmanlarca değerlendirilip kapsam üzerinde mükemmel bir görüş birliğine varılsa da bu test başka bir amaç doğrultusunda kullanılamaz. Bir testin ya da test puanlarının mutlak anlamda geçerliğinden söz edilemez. Geçerlik, testin amacına bağlıdır ve amaç değiştiğinde geçerlik için önceden toplanan kanıtların da anlamı kalmaz.

Sireci (2007) yapı geçerliği kavramını geçerliğin merkezine koymaya yönelik eleştirilere AERA'nın yapmış olduğu geçerlik tanımı ile karşılık vermektedir. Buna göre geçerlik, belli bir amaç doğrultusunda geliştirilen bir testten elde edilen puanlara dayanarak yapılan yorumları destekleyen teori ve kanıtların derecesidir. Belli bir amaç doğrultusunda geliştirilmiş bir testten elde edilen puanların yorumlanması da çeşitli kanıtlara dayanmaktadır. Bu kanıtların kaynakları 1) testin kapsamı, 2) cevaplama süreçleri, 3) testin içyapısı, 4) diğer değişkenlerle olan ilişki ve 5) testin sonuçları olabilir. Testin amacına göre bu kanıtlardan bazıları ise diğerlerinden daha önemli olabilir. Örneğin bir depresyon ölçeği geliştirme sürecinde testin iç yapısı ve diğer testlerle olan ilişkisi, testin sonuçlarına ilişkin kanıtlardan daha önemli olabilir. Ölçeğin iç yapısına ilişkin kanıtların toplanmasından sonra testin sonuçları, depresyonda olan bireylerle olmayanları ayırma konusundaki başarısı olarak değerlendirilebilir. Yordama amacı taşıyan seçme ve yerleştirme testleri için test puanlarının diğer değişkenlerle ilişkisi ön plandadır. Sınıf içi başarı testlerinde testin kapsamına ilişkin toplanan kanıtlar diğer kanıtlardan önce ele alınmalıdır. Bu arada testteki maddelerin cevaplama süreçleri gerek iç yapı, gerek kapsam gerekse diğer değişkenlerle ilişkiyi belirleyen önemli bir etkendir.

Geçerlik üzerine yapılan tartışmaların bir boyutu da *geçerlik türleri* yerine *geçerlik stratejileri* kavramının kullanılması üzerinedir. Uzun yıllar, farklı geçerlik türlerinin farklı amaçlara hizmet ettiği düşünülmüştür. Örneğin bir başarı testinde aranması gereken geçerlik yönteminin kapsam geçerliği, bir kişilik envanterinde incelenmesi gereken geçerliğin yordama geçerliği olduğu düşünülmekteydi. Murphy ve Davidshofer'e (2001) göre geçerlik türlerini birbirinden bağımsız tutmak doğru değildir. Wood (1961) farklı geçerlik türlerinden ziyade, geçerliği irdelemenin farklı yöntemlerinden söz etmenin doğru olduğunu ve buradaki temel sorunun testin amacına hizmet etme düzeyi olduğunu ifade etmektedir. Geçerliği incelemede birçok yöntem vardır. Geçerliğin farklı yüzleri olan bu yöntemler, testin kullanım amacına bağlı olarak farklı şekillerde adlandırılır (Anastasi ve Urbina, 1988).

Bu tartışmalar ekseninde geçerlik bir küpe benzetilebilir. Küpe ait yüzeylerin her biri bir geçerlik stratejisini temsil etmektedir ve bu altı yüzeyin toplamı küpü, yani geçerliği meydana getirmektedir. Geçerliği bu şekilde tanımlamakla, geçerliği irdeleme yöntemleri

odak noktası olarak alınmış olmaktadır. Bu tanımlama da, birleştirilmiş geçerlik kavramının uzantısı olarak yorumlanabilir.

Journal of Educational Measurement dergisinin 2013 yılındaki bir sayısı tamamen geçerliğe ilişkin tartışmalara ayrılmıştır. Bu sayının temeli Kane'in geçerlik üzerine yaptığı yorumlardır. Kane'e (2013) göre geçerleme çalışmaları iki temele dayanmaktadır: 1) test puanlarının kullanılması ve/veya yorumlanması ve 2) yapılan bu yorumların görgül, mantıklı kanıtlarla değerlendirilmesi. Bunlardan ilki puanların yorumlanması/kullanılması tartışmaları (interpretation/use argument-IUA); diğeri ise geçerlik tartışmaları olarak nitelendirilmiştir. Kane'nin geçerliğe ilişkin düşüncelerinin temelini oluşturan IUA'ya göre bir testten elde edilen puanlar farklı amaçlar için farklı alanlarda kullanılabilir. Buna göre testin geçerliğini incelemek için testten elde edilen puanların yorumlanması, puanların kullanım alanlarının belirli sayıltılar çerçevesinde incelenmesi gerekir. Yapılan yorumların, kullanım alanlarının ardından değerlendirmelerin yapılması gerekmektedir. Brennan (2013) Kane'nin bu geçerlik yorumlamalarını "geçerlik= IUA + değerlendirmeler" şeklinde formüle etmiştir.

SONUÇ

Geçerlik kavramının ne olduğu, nasıl sınıflandırıldığı ya da hangi çatı altında birleştirildiği konusunda tartışmalar halen sürmektedir. Bu konularda fikir ayrılıkları bulunmakla beraber, uzlaşılan noktalar da bulunmaktadır. Geçerliğin, testin neyi ölçtüğüne ilişkin kanıt toplama süreci olduğu görüşü kabul görmektedir. Bir test, ister sınıf düzeyinde bir başarı testi olsun ister psikolojik bir ölçek olsun örtük bir özellik ölçüldüğü için test/ölçek puanları ile doğrudan ölçme yapılamaz. Bu nedenle de testin ölçmek istediği özelliği tam olarak ölçüp ölçmediğini belirleyen kesin kanıtlar yoktur. Bu da bir testin *geçerlidir* ya da *geçerli değildir* şeklinde tanımlanmasını engeller ve geçerliğin kanıtlara dayalı olarak tanımlanması gerektiği gerçeğini ortaya çıkarır. Testin geçerliği kanıt toplamaya dayanan bir dizi yöntem ile mümkün olduğunca berraklaştırılabilir. Ne kadar çok kanıt toplanırsa testin geçerliğine dair o denli bilgi edinilmiş olur. Kanıt toplama süreci durağan olmayıp, zaman ve örneklemlere göre tekrar tekrar incelemeyi gerektirir. Ayrıca geçerlik teste değil, testten elde edilen puanlara dair bir özelliktir. Nitekim zaman içerisinde testin uygulandığı grup değişkenlik gösterebilir. Bu süre zarfında testin kendisi aynı kalmasına rağmen gruptan elde edilen puanların değişmesi, geçerliğin testin değil o testten elde edilen puanların bir ölçüsü olduğunu gösterir.

Geçmişten günümüze kadar süren geçerlik tartışmaları o dönemdeki tartışmaların içeriğine dayanarak çeşitli çağlara ayrılabilir: a)1950'li yılların başında yordama geçerliği baskın olmakla birlikte kapsam geçerliği kavramı da gündeme gelmeye başlamıştır. b) 1950 ile 1990 yılları arasında yapı geçerliği kavramı üzerinde durulmuş ve c) 1990'dan günümüze kadar ise pratikteki geçerlik sorunlarına ilişkin önerilen paradigmalar tartışılmaktadır (Brennan, 2013).

Geçerliğin, ölçüt, kapsam ve yapı geçerliği gibi türlere ayrıştırılması kanıt toplama ekseninde de değerlendirilebilir. Sireci ve Foulkner-Bond (2014), geçerliğin, yapı geçerliği ekseninde tanımlanması ve kanıt toplama süreci olarak ele alınması görüşündedir. Yazarlar kapsam geçerliğini, testin amacıyla uyumlu olma derecesi olarak tanımlamışlar ve *kapsam geçerliği kanıtları* ile *test kapsamına dayalı geçerlik kanıtları* terimlerini aynı anlamda kullanmışlardır.

Geçerlik, ölçmek istediği özelliği ne derece ölçtüğüne ilişkin kanıtlar toplama süreci olarak ele alındığında en azından uygulamada bir noktada uzlaşı sağlanmış olur. Ayrıca, geçerliğin farklı türlerde ele alınmasının pratik yararları da bulunmaktadır. Geçerlik türleri, testin hangi özelliğine ilişkin kanıt toplanacağına bağlı olarak, kanıt toplama yöntemlerinin tanımlanmasını sağlar.

Geçerliğin yapı geçerliği ekseninde ele alınması gerektiğine ilişkin görüşlerin kuramsal dayanakları olsa da bu yaklaşımın özellikle eğitim alanındaki uygulanmasında problemlerle karşılaşılmaktadır. Testin yapı geçerliğine ilişkin kanıtlar, kapsama ilişkin bilgi veremeyebileceği gibi kapsam geçerliği kanıtları da yapı geçerliğini doğrulamayabilir. Örneğin dört işlem becerisini ölçmeye yönelik hazırlanan bir matematik testinin kapsam geçerliği incelenmeden önce faktör analizi ile yapı geçerliğinin incelendiği düşünülsün. Testte sadece toplama işlemine dair soruların sorulması durumunda faktör analizi sonucunda test tek boyutlu çıkabilir. Bu, testin tek boyutu ölçtüğünü gösteren yapı geçerliğine ilişkin bir kanıttır. Ancak dört işlem becerisinden çıkarma, çarpma ve bölme işlemlerine dair soruların sorulmamış olması nedeniyle testin kapsam geçerliği düşük olacaktır. Bu nedenle özellikle eğitim ile ilgili testlerde karşılanması/incelenmesi gereken ilk geçerlik türünün kapsam geçerliği olması gerektiği söylenebilir. Bunun üzerine yapılan her işlem testin geçerliğine katkı sağlayacaktır.

Geçerliğin, kapsam geçerliği ekseninde tanımlanması, geçerlik kanıtlarının öğretmenler tarafından toplanması ve sonuçların yorumlanmasını da kolaylaştıracaktır. Bu durumda "eğitimde yalnızca kapsam geçerliğinin irdelenmesi gerekir" gibi bir sonuca varılmaması önemlidir. Çünkü eğitimde veya psikolojide ilgilenilen özellikler, doğrudan gözlenmeyen yapılardır. Amaç, bu yapıları en az hata ile ve amaca uygun olarak ölçebilmektir. İlgilenilen yapı, doğrudan gözlenemediğinden bu yapının ne kadar hatasız ölçüldüğü de kesin olarak bilinememektedir. Bunun yerine yapının ne kadar az hata ile ölçtüğüne dair kanıtlar sunulmaktadır. Bu kanıt toplama yöntemleri ister kapsam, ister ölçüt veya ister yapı geçerliği olsun tüm bunların birleşimi ile ortaya çıkan resim daha berrak, daha seçilebilir hale gelecektir. Aksi durumda duvara asılan, bulanık bir resim olacaktır.

KAYNAKLAR

Anastasi, A. ve Urbina, S. (1988). *Psychological testing (7th ed.)*. USA: Macmillan Pub. Co. Inc.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, (1974). Standards for educational and psychological tests and manuals. Washington, DC: American Psychological Association.

Borsboom, D., Mellenbergh, G., ve Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17(1),* 5–9.

Brennan, R. L. (2013). Commentary on "validating the ınterpretations and uses of test scores". *Journal of Educational Measurement, 50(1),* 73 – 83.

Campbell, D. T. ve Fiske D. W. (1959). Convergent and discriminant validition by the multitrait-multimethod matrix. *Psychological Bulletin, 56 (2),* 81 – 105.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22(3),* 5–11.

Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper.

Cronbach, L. J. ve Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Cureton, E. E. (1951).Validity.In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621-694).Washington, DC:American Council on Education.

Embretson, S. E. (2007). Construct validity: a universal validity system or just another test evaluation procedure? *Educational research, 36,* 449.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7*, 1, 3-13.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427–439.

Hopkins, K. D., Stanley, J. ve Hopkins B. R. (1990). *Educational and psychological measurement and evaluation.* USA: Prentice Hall.

Kane, M.T. (2001). Current concerns in validity theory, *Journal of Educational Measurement, 38(4),* 319-342.

Kane, M. T. (2006). Validation. In R. Brennan (Ed), *Educational Measurement,* (4th ed.,17-64). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(1),* 1-73.
Loevinger (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694.
Lord, F. M. ve Novick M. R. (1968). *Statistical theories of mental test scores.* New York: Addison- Wesley Publishing Company.
Lissitz, W.R. ve Samuelsen, K. (2007a). A suggested change in terminology and emphasis regarding validity and education, *Educational Researcher, 36(8),* 437-448.
Lissitz, W.R. ve Samuelsen, K. (2007b). Further clarification regarding validity and education, *Educational Researcher, 36(8),* 482-484.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.
Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test Validity* (33–45). Hillsdale, NJ: Lawrence Erlbaum.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., 13–103). New York, NY: American Council on Education and Macmillan.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.
Murphy, K. R. ve Davidshofer, C. O. (2001). *Psychological testing (5th.ed.).* New Jersey: Prentice Hall.
Rulon, P. J.(1946). On the validity of educational tests. *Harvard Educational Review, 16,* 290-296.
Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45,* 83-117.
Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher 36,* 477–481.
Sireci, S. ve Faulkner-Bold, M. (2014). Validity evidence based on test content. *Psicothema 26,* 1, 100-107.
Thorndike R. L.ve Hagen, E. (1961). *Measurement and Evaluation in Psychology and Education.* Newyork: John Wiley and sons.
Turgut, M. F. ve Baykul, Y. (2010). *Eğitimde Ölçme ve Değerlendirme.* Ankara: PegemA Yayıncılık.
Wood, D. A. (1961). *Test construction*. USA: Charles E. Merrill Books.

EXTENDED ABSTRACT

Validity is regarded as a concept demonstrating the degree to which a test measures the property intended to be measured in accordance with the purpose of measurement. Yet, no agreement has yet been reached on the definition, classification and determination of it, or on the significance and interpretation of the researched test scores.

The classification of validity, which was made by American Psychological Association in 1954 and which is frequently used today, involves four types of validity; namely, predictive validity, concurrent validity, content validity, and construct validity (Cronbach and Meehl, 1955). The classification was later reduced to three as content validity, criterion-based validity, and construct validity in 1966. Afterwards, the discussions as to the interpretation of results started (Kane, 2006).

In the 1915s, when the concept was first discussed, validity was based upon what we call criterion-based validity today (Lissitz and Samuelsen, 2007a). Although criterion-based validity has the difficulty of finding a valid and reliable criterion, it provides the user with practical knowledge once an appropriate criterion is found.

Validity studies in the 1930s and 1940s were the studies that we call predictive validity today. A new type of criterion-based validity called *concurrent validity* came into the agenda in that era. Just like predictive validity, concurrent validity is also based on the correlations holding between test scores and a criterion. However, in predictive validity, criterion is obtained after the scores received from the test which is under research for validity whereas in concurrent validity it is obtained simultaneously with the test scores or before the test scores.

That the criterion-based validity was not suitable for all test types was discussed in the 1940s for the first time. Rulon (1946) thought that an external criterion was not needed for the validity of a test, but that the degree of knowledge and skills acquired by individuals (that is to say, the test scores) were the criterion itself.. Cureton (1951) employed the term *content validity* for the first time, and thus explained content validity by relating it with the concept of criterion.

Cronbach and Meehl (1955) define content validity as showing that test items are the sample of the area to be measured (the topic and the behaviour). Yet, no consensus was reached on non-empirical definitions, and the validity definitions made rather theoretically through empirical methods began to gain importance.

In the 1950s, the definition of validity was extended in a manner so as to include the psychological and clinical research conducted by APA. Cronbach and Meehl (1955) used the hypothetico-deductive (HD) model in accounting for the theoretical structures, which received extensive interest especially in the 1950s; and they defined psychological structure as *individuals' properties which are regarded to exist*.

Cronbach and Meehl (1955) also stated that, after attaining the criterion and content validity of a test, the psychological structure which it intends to measure should also be determined. In this way, from that time on, the construct validity of a test was also considered. The authors pointed out that the construct validity could be researched when there was not a criterion which was defined as operational. Having defined construct validity, they explained how to measure the structures and pointed out the *nomological* network describing the nature of the network consisting of the interrelations of the constructs.

Ten years after Cronbach and Meehl (1955), the construct validity was taken as a model in addition to content validity and criterion-based validity in *Standards*. The influences of that thought continued in *Standards* (APA, AERA and NCME, 1974), which was published in 1974, and four independent types of validity were defined: predictive validity, content validity, concurrent validity, and construct validity. The study of construct validity is composed of such processes as determining directly or indirectly how to measure the constructs, forming hypotheses based on the theory explaining the contructs and testing the hypotheses, establishing nomological networks describing and explaining the constructs, and interpreting them.

Loevinger's (1957) view of putting predictive, concurrent and content validities under the roof of construct validity gained acceptance in the late 1970s. Accordingly, construct validity was not a type of validity, but it was an extended model containing all of the content, predictive and concurrent validities and the reliability processes which might be evidence for validity (Kane, 2001). The unified validity studies are based on the interpretation of unobservable constructs through observed scores.

Sireci (1998) and Messick (1989) oppose the view that content validity does not have a counterpart in the literature of psychometrics. Determining the content validity is a complex process, and content validity should be attached importance so as to avoid assessing the test scores only with numerical data. Borsboom, Mellenbergh and Van Heerden (2004) point out that the concept of unified validity does not answer the question of what a test measures, but that it is more related with the complex interpretations of scores which are explained via nomological networks. Those criticisms made later concluded that the unified model of validity was not suitable for the tests used in the field of education.

Another dimension of discussions made in relation to validity is concerned with the *types of validity*. Wood (1961) holds the view that talking of different methods of analysis rather than different types of validity would be correct.

Controversy as to what the concept of validity is, how to classify it, or under what roof it is unified still continues. Although disagreement on these issues exists, there are also issues agreed on. The view that validity is the process of gathering evidence on what a test measures is widely accepted.

Even though there are theoretical bases of the need to address validity on the axis of construct validity, problems exist in the implementation of the approach especially in the field of education. Evidence on the construct validity of a test would be insufficient in providing information on content. In a similar vein; evidence concerning content validity would also be insufficient in confirming construct validity.