EĞİTİM BİLİMLERİ ARAŞTIRMALARINDA GEÇERLİK VE GÜVENİRLİK SORUNSALI

THE PROBLEM OF VALIDITY AND RELIABILITY IN EDUCATIONAL RESEARCH

Yener ÖZEN[*]

Fikret GÜLAÇTI[*]

Mehmet KANDEMİR[**]

ÖZET

Eğitim Bilimleri araştırmacılarının ve öğretmenlerin araştırma için hazırladıkları; anket, bilişsel, duyuşsal ve devinişsel ölçekler (yetenek, tutum, ilgi) vb. ile sınav için hazırlanan ölçme araçlarının belli bir özelliğe yönelik birden fazla ölçüm sonuçları arasında tutarlılık gösterip, göstermediğini belirleyebilmek için güvenirlik özellikleri tartışılmıştır. Ayrıca Eğitim Bilimlerinde kullanılan ölçme araçlarının, tasarlanan ölçümlere ulaşma derecesinde; uygunluk, anlamlılık ve faydalılık özelliklerine bakılarak ölçme aracının geçerlik faktörleri bu çalışmada tartışılmıştır. Eğitim Fakültelerinin yeniden yapılanma sürecinde alandan eğitim araştırmalarına yönelen çalışmalara bir bakış açısı getirmesi için düzenlenmiştir.

Anahtar Kelimeler: Eğitim Bilimi, Geçerlik, Güvenirlik

ABSTRACT

Educational researchers and teachers discuss the reliability characteristics, who are preparing survey, cognitive scales, emotional scales and psicomotor scales (attitude, capabilitiy interest) etc. And the consistent of the measurement results and the consider validity. İn the exams in addition, it was carried out the usage measurement tools in the educational science in terms of the achievement (presenting) the designed measurement and the validity factors of the measurement tools considering appropriate, apprehension and usage properties. The study contributes from the pure science to the educational science in Educatioanl Faculties in Turkey.

Keywords: Educational Science, Validity, Realiability

[*] Atatürk Üniversitesi, Erzincan Eğitim Fakültesi
[**] Gazi Üniversitesi, Gazi Eğitim Fakültesi

GİRİŞ

Eğitimde, davranış görülmedikçe, duyulmadıkça, duyu organları veya benzeri araçlar vasıtasıyla ortaya konulmadıkça öğrenmenin gerçekleşip gerçekleşmediği kestirilemez. Tolman'a göre davranış, organizma ile çevresi arasındaki ilişkiler sistemini yeniden düzenleyen tepkilerdir (Uysal, 1976. s. 24). Bireyin sahip olduğu her davranış bireyin ilişkiler sistemini yeniden düzenlemeyebilir. Hatta bazı durumlarda bireyin davranışları, sistemi düzenlemekten çok, kendisi sistemin gerektiği davranışlara uyum sağlamak için bir takım davranışlarda bulunabilir. Skinner'in "bilimsel araştırmalarda, gözlenip, ölçülemeyen ara değişkenlerin, sonuç üzerindeki kısmi etkisi kabul edilmekle birlikte; asıl olan ölçülemeyen değişkenlerin (faktörlerin) işlevsel (bilimsel) çözümlemede yeri yoktur." ifadesi davranışların ölçülmesinin bilimsel yönden önemini ortaya koymaktadır (Beydoğan, 2003. s.144).

İnsan davranışlarının özellikle eğitim sahasında ölçüp-değerlendirmek için ölçme araçlarının, istenilen ya da sorgulanan problematiğe neden olabilecek değişkenlerin kontrol altına alınabilmesini mümkün kılabiliyor mu? Ölçme aracının belli bir özelliğe yönelik birden fazla ölçüm sonuçları arasında tutarlılık göstermesi ve ölçme araçlarının neyi ne derece ölçtüğünün ortaya konması; ölçme aracının güvenirlik ve geçerlik derecesini gösterir (Fraenkel, 1993. s.146). Eğitim ortamında kullanılan ölçme araçlarının güvenirliğine ve geçerliğine ilişkin iddiaların temelinde de "bireyin niteliklerinin ölçümünde, bireyin niteliklerinin kısa sürede önemli değişiklikler göstermeyeceği varsayımı yatar (Özgüven, 1994, s.87).

Bu savlardan hareketle Eğitim Bilimlerinde araştırma yapanlar için kullandıkları ölçeklerin amaca hizmet edip etmediği ve vargıya ulaşırken yaşananlarla, bulunanların örtüşme sorunsalı bu çalışmada tartışılmaya çalışılmıştır.

Geçerlikte Kritik İncelemeye Duyulan Gereksinim

Testlerin geçerlikleri ölçtükleri değişkenlerin özelliklerine göre farklılık göstermektedir. Örneğin; zihinsel becerileri ölçen testler bu değişkenlere göre geçerliği yüksek olabilir ancak endüstri çalışanlarının mekanik becerilerini ölçmede uygun olmayabilir. Hiçbir test bütün istenen nitelikleri içerisinde sağlayamaz. Testlerin geçerliğini artırıcı en önemli etmenlerden biri yapı geçerliğidir ve bunu sağlamak için gerekirse testin geçerliğine ilişkin diğer niteliklerden vazgeçilebilir (Cronbach, 1990, s. 145).

Geçerlik; bir testin puanlarından çıkartılan yorumların sağlamlığını araştırma işidir. Bu puan gerçekten ölçmek istediği değişkeni ölçtüğünü gös-

teriyor mu? Yapılan yorumun sağlamlığını ve ilişkisini incelemektir. Ölçmek istediğini ölçüp ölçmediğini araştırma işidir. Bir testin şöhretli yada ünlü olması onun değerli olduğunu göstermez. Yeni testler geliştirilmiş ve testlerin yeni kullanımları keşfedilmiştir. Bazı testler uygulamadan kaldırılmıştır. Modern test geliştirme genellikle kendini kritize (eleştirme) ile ilgilidir. Böylece testin niteliği ve test bilgisi gelişir. Buna rağmen bazı kötü testler halen uygulamada mevcuttur. Örneğin 1970'lerde askeri servisler ülke çapındaki milyonlarca öğrenciye mesleki tutum testleri uygulamışlardır. Test puanlarının öğrencilerin kabiliyetlerini belirlemede ve onlara rehberlik hizmetleri vermede yardımcı olduğu düşünülmüştür. 1977'lere gelindiğinde bu testler uzmanlarca inceleme altına alınmış ve kötü bir şekilde yapılandırıldığı anlaşılmıştır. Test puanlarına göre yapılan mesleki önerilerin uygun olmadığı tespit edilmiştir. Ancak buna rağmen testler değiştirilmeden 10 yıl süresince kullanılmıştır (Cronbach, 1990; s. 146).

Bilgi Kaynakları ve Kritizm

Testlerin bazı sınırlılıkları bir gerçeği yansıtmaktadır ki bu hiçbir testin her şeyi yapamayacağıdır. Fakat bazı testler kullanıcıların veya geliştiricilerin kendilerini eleştirmedeki yetersizliğini yansıtır. Yayınlanmış birçok teste ilişkin bu testlerin değerlendirilmesinde temel olabilecek detaylı bilgiler vardır. Testlerin değerlendirilmesinde kılavuzlar ve teknik kitaplar temel bilgi kaynaklarıdır. Testlerin kalite kontrolünün sorgulanmasında kullanılmakta olup bilgiler ilgili herkese açıktır. Testlerin yapılandırılması ve raporlaştırılması O.K. Bunos'un 1934'lerde testleri kritik incelemesiyle (eleştirmesiyle) gelişme kaydetmiştir. 1941–78 yılları arası yaptığı yıllık yayınlarda İngilizce konuşan ülkelerdeki hemen bütün testler eleştirilmiştir ve her bir test iki veya daha fazla uzman tarafından incelenerek öneriler geliştirilmiştir.

Ayrıca İngiltere'de MMY (The Night Mental Measurement Yearbook), Amerika'da Test Corporation tarafından sunulan test kritikleri bu alandaki kaynaklardır. Yine Journal of Psychoeducational Assessment, Measurement of Evulation in Counseling and Developmant ve diğer dergiler test inceleme konusundaki bilgi kaynaklarıdır (Cronbach, 1990, s. 147).

Test Standartları

Geçerliği sağlamada uygun tekniklerin seçilmesi için test kullanıcıların sorumlulukları şöyle ifade edilmiştir (Valois, 2000, s. 287).

Test kullanıcıları; Öncelikle testin amacı ve teste tabi tutulacak evren ve örneklem belirlenmelidir. Daha sonra bu evrene uygun test seçilmelidir. Güvenilir bilgi kaynakları testler tarafından sağlanan bilgiyi kuvvetlendirmek için araştırılmalıdır. Test geliştiriciler tarafından sağlanan materyaller okunmalı ve belirsiz veya eksik söylenen bilgi için testleri kullanmaktan sakınılmalıdır. Testin ne zaman ve nasıl geliştirildiği belirlenmelidir. Test için yapılan bağımsız değerlendirmeler ve olası alternatif ölçme araçları incelenmelidir. Test geliştiricilerin hipotezlerini destekleyecek gerekli deliller aranmalıdır. Testi seçmeden önce örnek setler (aşamalar), soru örnekleri, yönergeler, cevap kağıtları, kılavuzlar ve puan raporları incelenmelidir. Test içeriği ve norm grup/grupları veya kontrol grup/gruplarının hedeflenen amaç için uygun olup olmadıkları ortaya çıkarılmalıdır. Testler seçilip kullanıldıktan sonra puanlar doğru şekilde yorumlanmalıdır.

1. GEÇERLİLİK

1–1. Geçerliği Sorgulama Metotları

Geçerlik, yapılan yorumların sağlamlığını ve ilişkisini incelemektir. Bir test mükemmel bir şekilde hazırlanmış olabilir. Ancak yanlış yorumlandığında fayda sağlamayacaktır. Kullanıcıların şu iki soruya mutlaka cevap vermeleri gerekmektedir; "Verilecek karar için bu test ne kadar geçerlidir" veya "benim yaptığım çeşitli yorumlar ne kadar geçerlidir?" Bu nedenle geçerlik yorumların sağlamlığını sorgulama işi çok önemlidir. Geçerlikte sorgulama türünü gösteren üç anahtar kavram mevcuttur. Bunlar; ölçüte yönelik (criterion-oriented), kapsam ve yapı geçerliğidir. Test standartları yaklaşık otuz beş yıldan beri bu üç kavram etrafında organize edilmiştir.

Ölçütün önemi *(Criterion emphasis)*: Örneğin deniz kuvvetleri, gemicilere gemi motorları ile ilgili bir kursa gönderecektir ve kursta kimin başarılı olacağını araştıracaktır. Personel psikologları gemicilerin aldıkları test puanlarını ölçütlerle karşılaştırarak başarıyı belirlemeye çalışırlar. Burada söz konusu olan alınan test puanlarının ölçütlerle karşılaştırılması ve yorumlanmasıdır.

Kapsamın önemi *(Content emphasis)*: Örneğin Amerika'da bir bölgedeki okullar lise son sınıf öğrencilerinin Amerikan yönetimini ne ölçüde anladıklarını test etmek istiyorlar. Burada test kapsamında bölge yönetimin-

deki görevlilerin test kapsamı içine alınması önemlidir. Test hazırlayıcılar maddeleri ilişkili, açık ve sağlam hazırlamalı, maddeler söz konusu kapsama eşit önem verecek şekilde dağıtılmalıdır.

Yapının önemi *(Construct emphasis):* Örneğin içine kapanıklık puanlarının davranış ve hisleri tanımladığı iddia edilmektedir. Tanımlamanın doğruluğunu değerlendirmek için, bizim öncelikle içine kapanıklığın asıl anlamını anlamamız gerekmektedir. Yorumlayıcının teorisine göre içine kapanıklar bu ya da şu durumda nasıl hareket ederler? Hangi uyarıcılara cevap verirler? Duygusal streslerini nasıl gösterirler? Bu tür sorularla anlam çıkarıldıktan sonra biz testten yüksek puan alan bir kişinin teorideki gibi davrandığını söyleyebiliriz. Bu yapı geçerliğidir. (Cronbach, 1990, s.151-152)

1–2. Ölçüt yönelimli sorgulama

Ölçüt merkezli yorumlama, puanları başka değişkenlere göre ifade etmeye dayanır. İfadeler, eğer ikinci değişken gözlenirse ne olması beklenir gibi tahminlerdir. Bu beklentiler önceki tecrübeleri takip ederler. Örneğin testten 25 puan alan bir kişinin bu puanı için uygun bir tahmin daha önce aynı puanı alan benzer kişilerin tecrübelerinden çıkartılır.

Kriter merkezli yorumlama puanların başka ölçütlerle karşılaştırılmasına dayanır. Örneğin bir tutum ölçeği uygulanmış olsun. Elde edilen puanlar daha önceden uygulanmış ve yorumlanmış tutum puanlarıyla karşılaştırılarak ifade edilebilir. Bu puanların yorumlanmasında geçmiş tecrübelerden yola çıkılarak araştırmacılara, danışmanlara yardımcı olmak amacıyla çeşitli tablolar ve grafikler hazırlanmıştır. (Valois vd, 2000, s. 281–294).

1–3. Kapsam ilişkili sorgulama

İçeriğin incelenmesi özellikle yeterlilik sertifikasına yönelik testler veya eğitim veya terapötik servisler tarafından kullanılan testler açısından önemlidir. Bu tür hizmet servisleri kişilerde belirli bir yeterlilik geliştirmeye, belirli tutumları oluşturmaya veya insanları belirli bir şekilde davranmaya yöneltmeye çalışırlar. Uygulanacak bir final testi başka özelliklerden ziyade gerçekleştirilmeye çalışılan bu tür özellikleri ölçmeye yönelik olmalıdır. Dolayısıyla testin maddeleri kapsam doğrultusunda hazırlanmalıdır. Örneğin bir mesleğe yönelik yeterlilikleri ölçmek amacıyla hazırlanan bir testte kapsam geçerliğine karar vermek için, araştırmacı test içeriğini, meslek yeterlilikleri ile karşılaştırmalıdır. Burada bazı eksiklikler olabilecektir. Ancak bu eksikliklerin testin amacını ne ölçüde etkileyeceğine dikkat edilmelidir. Uygun alanların seçilmesi ve bu alanlar üzerinde anlaşmanın sağlanması ölçüt yöne-

limli sorgulamada olduğu gibi kapsam geçerliğinde de önemlidir (Cronbach. 1990; s.157).

1–4. Yapı Geçerliği: Puanların Açıklanması ve Onların İlişkileri

Test geliştiriciler ve test kullanıcılar sürekli testlerde niçin bazı insanların puanlarının yüksek, bazılarının az olduğunu ve test performanslarının günlük davranışın karşılığı olduğunu veya olmadığını anlamaya çalışırlar. Bu nedenle söz konusu bu durumları açıklamaya yönelik pratik sonuçlara gidilir; yeni testler geliştirilir, kullanılan yollarda değişiklikler yapılır, öğretim programları veya mesleki gereklilikler yeniden düzenlenir.

Test düzenleyiciler elde edilen cevapları yorumlamada ölçüt tabloları ve test içeriğine yönelirler. Her test ima ettiği niteliği veya kavramı ölçmede az da olsa eksikliğe sahiptir. Bu eksikliklerin belirlenmesi açıklama sürecinin bir parçasıdır. Test puanlarının açıklanması bilimsel bir sorgulamadır. Bunun için delillerin toplanması, delillere yönelik kritik incelemelerin gözden geçirilmesi gereklidir. Benzer olaylar, objeler, durumlar veya kişiler dikkate alınmalıdır (Valois. 2000; s. 281–294).

1–5. Geçerlik ve Kriter

Yordama geçerliğinin en zor kısmı uygun ölçüt verilerinin sağlanmasıdır. Örneğin ilk 6 aylık yapılan satışların satış başarısını yansıttığını varsaydığımızda eğer 6 aylık kayıtlar gerçekten 6 aylık satış başarısını yansıtmazsa kullanılan test uygun bir test olmaz. Buradaki önerilen zayıflığına bakalım. "Satış miktarı" başarıya karar vermede uygun bir temel gibi görünse de bazı satıcıların diğerlerinden daha istekli olduğu da ihmal edilmemelidir. Bu durumun kontrol etme yöntemi olarak; satıcıların ortalama satış miktarı ile ölçümdeki satış kayıtlarıyla karşılaştırarak kontrol edebileceğimizi varsayarız. Ancak yine başka değişkenler burada etkili olabilir. Bir bölgedeki kötü ürünler o yılı başarı için olumsuz etkileyebilecektir. (Francis. 2000; s.149–159).

1–6. Korelasyon ve regresyon katsayılarının anlamı

Korelasyon ve regresyon katsayısı iki değişkenin ne kadar örtüştüğünü yansıtır. Test puanlarının ölçütle olan korelasyonu geçerlik katsayısıdır. İki değişken (Rxy) arasındaki ilişki (örtüşme) yüksek olması bu iki değişken arasındaki sebep sonuç ilişkisini göstermez. Bu örtüşme farklı değişkenlere bağlı olabilir. Örneğin, "kelime bilgisi puanları ile okuma becerisi

puanları arasında ilişki vardır." İyi kelime bilgisi kolay okumayı sağlar mı? Kolay okuma iyi kelime bilgisinin göstergesi midir? Bu sorulara verilebilecek genel cevap evet olacaktır. Ancak hem okuma becerisi hem de kelime bilgisi üstün zihinsel fonksiyonların da sonucudur. Dolayısıyla yorumlama yapılırken bu tür faktörlerde dikkate alınmalıdır. Sosyal durumlarda geçerlik katsayısının. .60'ın üzerinde çıkması alışılmamış bir durumdur. Çünkü sosyal durumlar ve buna bağlı olarak da insanlar sürekli değişmekte ve en iyi yordamalar bile yanlış olabilmektedir. "En iyi geçerlik katsayısı nedir?" sorusunun cevabı "Ne kadar alabilirsen o kadar" olacaktır (Attoum ve Khasawneh. 1999; s.3–54).

1–7. Test İçeriğinin Kontrol Edilmesi

İçeriğin seçilmesi: İçeriğin geçerliği "ki onun amaca uygunluğunu gösterir" test dikkatlice planlandığı zaman iyileştirilebilir. İçeriğin geçerliği testin neyi ölçmeye niyetlendiğine bağlıdır. İdeal olarak test geliştiriciler ölçüm alanlarını uygun olarak tanımlarlar ve bunu test içerisinde iyi temsil etmeye çalışırlar. İçerik geçerliği okul sınavlarının sağlamlığına karar vermede önemli rol oynar. Burada testler öğretilecek hedefleri kapsamalıdır. Karar vericiler test içeriğinde neyin önemli olduğuna ilişkin farklı düşüncelere sahip olabilirler. Öğretmenler test içeriğini belirlerken öğretim sürecinde ne öğrettiklerinin çerçevesini çizmeye çalışırlar. Test içeriğini sınırlamada en ideal olan, öğrencilerine gerçekten ne öğretmek istediği ile paralel olmalıdır (buna öğretim geçerliği adı da verilir). Bu durum özellikle mezuniyete yönelik testler için geçerlidir. Bazen öğretim geçerliği ile kapsam (içerik) geçerliği aynı anlamda kullanılmaktadır. Testte, içeriği temsil ederken bunu en iyi, alanların alt bölümlerinin haritasını çizerek ve her bir alt bölüm için istenilen sayıda maddelerin tespit edilmesiyle sağlayabiliriz (Cronbach, 1990; s. 170–172).

Örnek: Belirtke tablosu aşağıdaki şekilde oluşturulur. * işaretliler önem verdiğimiz alanlara işaret etmektedir. Test maddeleri oluşturulurken bu alanlar dikkate alınarak testteki madde sayılarının ağırlıkları buna göre belirlenir.

Hedefler	İçerik				
			*		
	*				*
		*	*		
			*	*	
	*				*

Bu teknik içeriğin dengeli bir şekilde temsil edilmesini sağlayan, içerikte bazı noktalara aşırı önem vermemizi engelleyen bir tekniktir.

İçerik geçerliği, ölçülmek istenen değişkenlerin testte gerçekten temsil edilip edilmediğine karar vermemizi kolaylaştırır.

Burada karar verirken iki noktaya dikkat edilmelidir: Birinci Test maddeleri herhangi bir konuyu çok mu vurgulamaktadır? İkinci Test ölçülmek istenen özelliğin dışında olan ve sonucu etkileyebilecek maddeleri içermekte midir? Öğretmenlere ve test geliştiricilere özellikle başarı testleri geliştirmede içerik seçme konusunda yardımcı olabilecek kontrol listesi aşağıdaki gibidir. Ayrıca maddelerin test geçerliğinin sağlanmasında yardımcı olabilecek kontrol listesidir.

Çoktan Seçmeli Başarı Testleri Geliştirmede Kullanılabilecek Kontrol Listesi Test maddeleri hedef ve hedef davranışlarla uygunluk göstermekte midir? Çeldiriciler materyalden ilgisiz midir? Test madde kökleri ve seçenekleri anlaşılır ve teste uygun mu? Madde kökleri tek bir problemi mi yansıtmaktadır? Negatif soruların altı çizilmiş midir? Gramer kurallarına uygun mudur? Tek bir doğru var mıdır ve cevap açık mıdır? Testteki diğer maddelerle bu maddenin cevabı olabilecek ifadeler var mıdır?

Cevap seçeneklerinin uzunlukları eşit midir? Hiç, tamamı, genellikle, hepsi gibi sözcükler ipucu olabileceğinden kullanmaktan kaçınılmış mıdır? (Albatsh ve Abderhanan. 1990; s.92-136).

1–8. Yapı Geçerliği

İçeriğin analiz edilmesi, ölçülmek istenilen şeyi ölçmek için garanti değildir. Örneğin "düşmanlık davranışı" ile ilgili bir test geliştirilsin ve testin içeriği de dikkate alınarak agresif davranışların düşmanlığın göstergesi olduğu kabul edilsin. Ancak burada bazıları bu göstergeyi kabul ederken özellik-

le klinisyenler bunu kabul etmezler. Çünkü onlara göre stres altındaki insanlar da agresif davranışlar gösterebilirler ancak bu düşmanlığın bir göstergesi değildir. Bu örnekte içerik uygun olmakla birlikte yapının uygunluğu konusunda görüş ayrılığı vardır. Yapı geçerliği süreci, bilimsel kavramlar ve bunların ölçümlerine benzer. Örneğin bir kütle ölçülmek istendiğinde kütlenin ne olduğu ile ilgili bir kuramsal yapı vardır. Siz de bu kuramsal yapıya uygun olarak ölçüm gerçekleştirirsiniz. Kütlenin ne olduğu ile ilgili kuramsal yapı değiştiği zaman sizin ölçümünüz (ölçme aracınız) de değişecektir. Freud, bir insanda düşmanca davranışların olmaması demek o insanda düşmanlık duygularının olmadığı anlamına gelmediğini ispatlamıştır. O halde düşmanca davranışları yada agresif davranışları ölçmekle bir kişinin düşmanlık duygularına sahip olup olmadığına karar veremeyiz. Yapı geçerliğine ilişkin bir karmaşa da başlangıç noktasının belirlenmesine ilişkin bilgidir. Psikologlar zekâyı tanımlamadan zekâyı nasıl ölçebilirler. Bilindiği gibi Zekâyı ölçme çalışmalarından sonra zekâ tanımları bu ölçme işine uygun olarak yapılmaya çalışılmıştır. Daha sonra bu zekâ testleriyle zencilerin beyazlardan daha az zeki olduğu yorumu yapılmıştır. Oysa fen bilimcileri öncelikle uranyumu tanımlamışlar. U235 ve U238 arasındaki farkları ortaya koyduktan sonra U235 ve U238 arasındaki farkları ölçmeye yönelik testler geliştirmişlerdir. Yapı geçerliğinde, test puanlarına dayalı yorumlama ve öneri getirme ikna edici savunmayı gerektirir. Örneğin bir tez jürisinde tezini savunan araştırmacı ölçümler sonunda elde ettiği puanları yorumlarken ve öneriler getirirken, Ölçmek istediği şeyi örneğin zekâyı ölçtüğünü ileri sürer. Bunu yaparken zekânın tanımı budur, zekâya etki eden faktörler bunlardır ve benim ölçeğim de bu tanıma ve etkileyen faktörlere uygun olarak hazırlanmıştır. Öyleyse elde ettiğim ölçümler doğrudur ve bu anlama gelir şeklinde bir savunma yapar ve ikna etmeye çalışır. Örneğin hazırlanan zekâ testinin geçerli olduğuna bir kurumdaki bazı üyeler ikna edilirken, bu konudaki yeterliği daha fazla olan bazıları da ikna edilemeyebilir. Geçerlikte önemli olan bu kurumdaki kişilerin çoğunluk görüşünün alınarak yapı geçerliği konusunda bir karara varabilmektir. Bu cümlede bahsedilen durum yapı geçerliğinin sağlanması çalışmalarında kullanılan tekniklerden "uzman kanısına başvurma" tekniği vurgulanmaktadır. Bir testi kullanabilmek için (ÇN; eğer o testi siz geliştirmemişseniz) o testi geliştiren uzmanın testi geliştirmedeki amacını çok iyi bilmeniz gerekir. Ayrıca testi geliştirenin o alandaki fikirlerini de iyi bilmek gerekir. Fikir uygunluğu içinde misiniz? Bunların yapılması alternatif yorumların yapılmasına katkı sağlar (Cronbach, 1990; s. 183).

1–9. Yordama Geçerliği

Yordama "istatistiksel teknikler kullanılarak bilinenlerden yararlanılarak bilinmeyen durumlar hakkında yapılan geleceğe yönelik tahminlerde bulunma işlemidir" (Arıcı; 1972; s. 146). Testler çoğu kere bireylerin ilerdeki davranışlarının önceden kestirilmesi amacıyla geliştirilirler. Örneğin, lise öğrencilerinin okuma testi puanları onların gelecekteki okuma puanlarını tahminde kullanılabilir. Eğer belirtilen testin puanları lise öğrencilerinin gelecekteki başarı puanlarını kestirmede kullanılabiliyorsa okuma testi puanlarının yordama geçerliği var demektir. Bir testin yordama geçerliği, o testten elde edilen yordayıcı puan ile ölçülmek istenen özellikleri ölçtüğü bilinen bir ölçüt arasındaki korelasyonun hesaplamasıyla elde edilir. Test puanlarına yordayıcı, başarı ölçülerine ölçüt ve bu iki puan kümesi arasındaki korelasyona da yordama gücü ya da geçerlik katsayısı denir. Yordama geçerliği test puanlarının, ölçütü yordamadaki isabet derecesini anlatır (Öncü. 1995; s.73).

1–10. Benzer Ölçekler-Uyum Geçerliği

Geçerliği daha önceden belirlenen test ile aynı değişkenleri ölçtüğü ileri sürülen ve yeni hazırlanan test arasındaki korelasyon yeni testin uyum ya da benzer ölçekler geçerliğidir. Bu geçerlik türünde, ölçüt elde edebilmek için, yordama geçerliğinde olduğu gibi uzunca bir süre beklemeye gerek yoktur(Öncü. 1995; s.76–77).

1–11. Görünüş Geçerliği

Ölçme aracının neyi ölçtüğünü değil de neyi ölçer göründüğünü belirtmektedir. Bir testin görünüş geçerliğine sahip olması için ölçmek istediği özelliği ölçüyor görünmesi gerekir. Ancak, bir testin geçerliğini bazı durumlarda yükseltmek, bazı durumlarda ise gizlemek gerekir. Örneğin ticaret amacıyla geliştirilen testlerin görünüş geçerliği yükseltilirken, kişilik testlerinin görünüş geçerlikleri gizlenmektedir. Çünkü kişiler böyle testlere doğru cevap vermekten kaçınabilir(Öncü. 1995; s.73).

2. GÜVENİRLİK

Güvenirlik çalışmalarının odak noktası şudur: Eğer kişi iki defa teste tabi tutulursa iki testten aldığı puanlar birbiriyle tutarlı mıdır? Birbirine ne kadar yakındır? Bu bölümde değineceğimiz noktalar gözlenen puan, gerçek

puan, hata, ölçmenin standart hatası ve güvenirlik katsayısıdır (Cronbach. 1990; s.191).

2–1. Ölçme Hataları

Mike üç dakikalık bir kelime çalışmasında 162 kelime veya diğer bir ifadeyle dakikada 54 kelime yazmıştır. Bu puan Mike'in becerisini ne ölçüde yansıtmaktadır? Varsayalım ki öğretmen dakikada 50 kelime yazılmasını yeterli kabul etmiştir. Mike bu düzeyin gerçekten üzerinde midir? Mike geçen hafta dakikada 45 kelime yazmıştır. Mike'in bugün ki puanı olan 54, onun kendisini geliştirdiğinin göstergesi midir? Veya değişim dalgalanması mıdır? İki ölçüm arasındaki uyumsuzluğun birçok nedeni olabilir. Bir hareketten diğerine, "dikkat ve çaba" değişebilir. Özellikle uzun periyotlarda puan değişmesi, fiziksel büyüme, öğrenme veya sağlık ve kişilikteki değişimlerden kaynaklanabilir. Yine ikinci ölçümde daha açık soruların kullanılması diğer bir faktör olabilir. Kişi birinci testte soruları kolay bulabilir, ikinci testte farklı bulabilir. Bu iki ölçüm arasındaki puan farklarını yorumlayabilmek için iki konuya dikkat edilmelidir: 1. Gerçek puan teorisi 2. Genellenebilirlik teorisi (Brosnan ve Lee. 1998, s. 559–577).

2–2. Gerçek Puan Teorisi (True Score Theory)

Test teorisinde hata kavramı istenmeyen değişkene işaret eder. Ölçüm hatalar giderilene kadar sürdürülmeli ve böylece gerçek puan elde edilmelidir. Ancak davranış örneği sınırlı olduğu için gözlenen puan gerçek puandan farklılık gösterir. Buradaki farklılık ölçme hatasıdır. Geleneksel olarak hataların varlığı gözlenen puanın gerçek puandan düşük olmasına sebeptir. Örneğin bir koşucunun yarışlara hazırlandığı ve farklı zamanlarda bir mesafeyi, 23.7, 24.0, 24.2,... 25.1, 25.2 saniyelerde koşmuş olsun. Bu ölçümlerin ortalaması 24.7'dir. Koşucunun gerçek ölçüm puanı 24.7'dir. Koşucu bu puana, daha önce 23.7 saniyede koştuğu ölçümü göstererek itiraz edebilir. Bu durumda koşucuya 23.7 ve 25.2 saniyede koştuğu durumlara bir çok faktörün etkisinin olabileceği ve 23.7 saniyede tekrar koşmasının belki de hiç mümkün olamayacağını söyleriz. Koşucuya gerçek puanının 24.7 olduğunu söyleriz. Bunu gerçek puan teorisine göre söyleriz. Standart hata: Aynı obje veya olay üzerindeki bir dizi ölçümün standart sapması onun ölçümdeki standart hatasıdır. (Ç.N. Çünkü ölçümde hata, bir standart sapma kadar olabilir.) Ölçümün standart hatasının büyük olması ölçümün uygunluğunun az olduğunu gösterir. Ölçümün standart hatasının karesi ölçümün hata varyansını verir. Bu tür hesaplamalarda sabit hata ihmal edilir. Çünkü sabit hata her defasında var olan bir hatadır ve değişmemektedir. Karşılaştırma

yaparken bu hatanın varlığı karşılaştırmamızı etkilemez. Çünkü karşılaştırdığımız puanda da aynı hata aynı derecede bulunmaktadır. Gerçek puan teorisine göre bir ölçümün standart hatasının bulunması güven aralığının belirlenmesi ve gerçek puanın ortaya konulması açısından önemlidir. Örneğin bir ölçümde kişinin puanı 63 ve ölçümün standart hatası da 3.5 olsun. Gerçek puan teorisine göre bu kişinin gerçek puanı 59.5 ile 66.5 arasında yer alacaktır (Cronbach, 1990, s. 192–195).

Güvenirlik katsayısı: Varyans, standart sapmanın karesidir. Sabit hata ihmal edildiğinde hata ortalaması sıfır olduğu için şu formül kullanılır.

Gerçek puan varyansı = gözlenen puan varyansı - hata varyansı.

Gerçek puan varyansını tahmin etmek için bir grup kişinin gözlenen puan varyansı bulunur ve hata varyansı çıkartılır.

Güvenirlik katsayısı(r_{xx}) = gerçek puan varyansı / gözlenen puan varyansı

Güvenirlik katsayısı iki ölçüm arasındaki korelasyonu gösterir. Ölçümler hata içermediği zaman korelasyon 1'e ulaşır. .50 katsayısı hatanın çok olduğuna işaret eder. Her iki ölçümde de hatanın çok olması demek ki bu hata iki standart hata kadar olabilir ve bu da .50'ye işaret eder. Bu katsayı ile ölçümün standart hatası arasındaki ilişki ölçümün standart hatasının karesi (hata varyansı)= gözlenen puan varyansı x ($1-r_{xx}$) ile gösterilir. Ölçümün standart hatası genellikle eğer test bir grup için çok kolay veya zor değilse bir gruptan diğerine değişmez fakat grupların farklı varyansı olur. Bu nedenle güvenirlik katsayısı gruplar arasında değişir. Bu gerçek ölçümün standart hatasını katsayıdan daha önemli kılar (Valois vd. 2000, s.288).

2–3. Genellenebilirlik Teorisi

Genellenebilirlik teorisi hata kaynakları arasında ayrıştırılabilir. Yani hata kaynaklarını incelerken genellenebilirlik teorisini görebiliriz. Örneğin, bir kişinin yazı yazma becerisini ölçmek istediğimizde ona Mayıs ayının 3.,5.,7.,12.,23.,30. günlerinde bazı ölçümler yaparız ve bu kişinin aldığı puanlarının ortalamasının o kişinin ortalama puanı olarak belirleriz. Yine bu kişinin gelecekte de aynı puanı alabileceğini söyleriz. Bu bir yordamadır. Ancak bu kişinin örneğin Mayıs ayının 5. Günü aldığı puan sadece o gün için gerçek puanıdır. Bu puanı çeşitli nedenlerden dolayı (gözlemciden ve durumdan kaynaklanan) tekrarlayamayabilir. Genellenebilirlik çalışmaları puanların çeşitli değişkenlerden ne ölçüde etkilendiğini tahmin etmemize yarar.

2–4. Testin Güvenirliğini Tahmin Metotları

Güvenirlik tahminin de temel amaç ölçümlerdeki toplam değişkenliğin ne kadarının gerçekte karşılığı olmayan hata kaynaklı değişkenlik sayılabileceğinin belirlenmesidir (Özçelik. 1992; s. 46).

Testin güvenirlik katsayısının bulunmasında çeşitli yaklaşımlar vardır. Testin güvenirlik tahmini için kullanılan metotlar hata kaynaklarından hangisini dikkate aldığına göre farklılık gösterir. Aşağıda öğretmenlerin hazırladıkları testlerin güvenilir olup olmadıklarını ya da ne derecede güvenilir olduklarını tahmin etmenin değişik metotları vardır. Bunlar (Balcı. 1995; s. 117–118).

Testin tekrarı,

Eşdeğer (paralel) formlar,

İç tutarlılık

Testi yarılama,

Kuder Richardson, (KR-20 ve 21 formülleri),

Cronbach Alfa katsayısı,

Hoyt'un varyans analizi ve

3.Puanlayıcı güvenirliği metotlarıdır.

1, 2. ve 3(a) metotlarının hepsi korelasyon katsayısı kullanılarak hesaplanır (Öncü. 1995; s. 44).

2–5. Testin Tekrarı Metodu

Testin tekrarı yöntemi ölçülen niteliğin kararlı olduğu durumlarda uygulanması gereken bir yöntemdir. Testin ölçtüğü nitelik sürekli değişkenlik gösteriyorsa bu metotla testin güvenirliği hesaplanmamalıdır. Bu metot daha çok iki uygulama arasında kolaylıkla değişmeyen özellikleri ölçen testler için uygundur. Örneğin, genel zihin yetenekleri, kişilik testleri, ilgi envanterleri, tutum ölçekleri vb. gibi testlerin güvenirlikleri bu yaklaşımla hesaplanabilir Testin tekrarı yönteminde, güvenirliğini hesaplayacağımız testi, aynı gruba yaklaşık aynı şartlarda belli bir zaman aralığında iki defa uyguladığımızda elde edilen iki puan dağılımları arasındaki person momentler çarpımı korelasyonu hesaplanır ve yaklaşımdaki sayıltı, iki ölçme arasındaki

zaman diliminde ölçmeye konu olan değişkenin niteliğinin önemli ölçüde değişmediğidir (Öncü. 1995; s. 46).

2–6. Eşdeğer (Paralel) Formlar Metodu

Paralel formlar güvenirliği, bir testin iki paralel formu hazırlandığı zaman uygulanır. Bir testin paralel formlar yoluyla güvenirliğini anlamak için, testin iki eşdeğer formundan aynı gruba aralıksız aynı anda yada aralıklı farklı iki zaman aralığında uygulanması ve uygulamalardan elde edilen puan dağılımları arasındaki korelasyon hesaplanır ve elde edilen değer testin paralel formlar güvenirliğini verir. İki testin eşdeğer olabilmesi için; kapsamları, ölçtükleri davranışları ve maddelerinin sayı ve nitelikleri birbirine denk ve gözlenen ortalamaları ile varyansları eşit olmalıdır. Ayrıca test geliştirme tekniklerine uymalıdır. Bunun için de her iki testteki maddeler eşit güçlük derecesinde olmalı, aynı formatta olmalı ve talimatları (yönergeleri) aynı olmalıdır (Öncü.1995; s. 44).

Eşdeğer olabilecek şekilde hazırlanan iki test, aynı öğrenci grubuna, aynı zamanda uygulanmışsa, bir eşdeğerlik katsayısı verir. Eğer bu iki test farklı iki zaman aralığında uygulanmışsa, hem öğrenci grubunun hem de test formlarının eşdeğerliği kontrol edilmiş olabilir (Gronlond. 1976; s. 129).

Eğer test sonuçlarını uzun süreli olarak kullanmak istiyorsak testin tekrarı yöntemini kullanabiliriz. Bir testin eşdeğer formlarını kullanmak güvenirlik tahminleri dışındaki sebepler için de kullanışlıdır. Program veya öğrencinin başarısının değerlendirilmesi için, ön testte verilen testi uygulamak yerine onun eşdeğeri olan testi kullanmak daha iyidir. Çünkü, iki testin aynı güçlük derecesinde ve farklı sorulardan oluşması hatırlama etkisini azaltır (Öncü. 1995; s. 48).

2–7. İç Tutarlılık Metotları

İç tutarlılık yönteminin dayandığı temel görüş, her ölçme aracının, belli bir amacı gerçekleştirmek üzere oluştuğu ve bunların bilinen ve eşit ağırlıklara sahip olduğu varsayımıdır (Karasar. 1982; s. 157).

2–8. Testi yarılama metodu

Testin güvenirliğini tahminde en çok kullanılan bir metottur (Tekin.1977; s.44; Yıldırım.1983; s. 140; Turgut.1992; s.33). Çünkü bu metot tek bir test formu, tek bir öğrenci grubu ve tek bir test uygulaması gerekir. Bir testi iki kere uygulamanın veya bir testin iki eşdeğer formunun hazırlanmasının güç olduğu ve testin tek bir değişkeni ölçtüğü durumlarda başvurulması gereken bir metottur.

Test yarılama metodu teorik olarak eşformlar metodu ile aynıdır. Bununla birlikte test yarılama metodu genelde bir iç tutarlılık katsayısı vermektedir (Özgüven. 1994; s.88). Çünkü eşdeğer formlar tek bir test için de verilir. Yani testin alternatif formunu uygulamak yerine sadece bir test uygulanır.

Bu yöntemle test güvenirliğini tahmin etmede, uygulanan test iki eşdeğer yarısından alınan puanlar arasındaki korelasyon (r) hesaplanır. Bu korelasyon yarı testin güvenirliğini verir. Testin yarısının güvenirliği belli ise testin tamamının güvenirliği Spearman-Brown formülünden yararlanılarak kestirilir. Bu formül şöyledir:

$$r_x = \frac{2_{r1,2}}{1+_{r1,2}}$$

r_x=Testin tamamının güvenirlik katsayısı $r1_{,2}$=Testin iki yarı arasındaki korelasyon katsayısı

Bu formül, testin iki yarısından elde edilen puanların varyansları eşit kabul edildiğinde kullanılmaktadır. Eğer testin iki yarısından elde edilen puanların varyansları eşit değilse, yukarıdaki formül yerine aşağıdaki formül kullanılmalıdır. Çünkü bu formül, yarı puanların varyanslarının eşit olmayışından doğacak olan hatayı önlemiş olacaktır (Turgut. 1981; s. 17).

$$r_x = \frac{4r_{1,2}S_1S_2}{S_x^2}$$

$r_{1,2}$= Testin iki yarı arasındaki korelasyon S1= Birinci yarı puanlarının standart sapması S_2= İkinci yarı puanlarının standart sapması Sx=Toplam puanların standart sapması r_x= Testin tamamının güvenirlik katsayısı

Bu metotla ilgili problemlerden biri testi eşdeğer iki yarıya nasıl ayırmak gerektiğidir. Bir öğrenci grubuna uygulanmış testi iki eşdeğer yarıya bölmenin yollarından biri, testteki tek numaralı sorularla çift numaralı soruları ayrı ayrı puanlamaktır. Diğeri de güçlük derecelerine göre iki ayrı forma ayırmak ve formların her birini ayrı ayrı puanlamaktır:

Testi yarılama yöntemiyle güvenirlik işleminde Spearman-Brown Düzeltme Formülü kullanıldığından güvenirlik katsayısı olduğundan biraz daha yüksek değerler vermektedir (Özgüven. 1994; s.89). Kaplan, (1986)' a göre bu farklılık testin yarısına ilişkin korelasyonun yüksek, yada düşük olduğu zaman daha az, orta seviyelerde olduğu zaman biraz daha fazladır (Akt: Özgüven. 1994; s.89). Bu özelliği nedeniyle özellikle testin iki yarısına ilişkin varyansın eşit yada birbirine çok yakın olmadığı durumlarda Spearman-Brown eşitliğinin kullanılması tavsiye edilmektedir. Bu durumda "Croncbach Alfa Katsayısı" (1951) kullanılabilir. Ancak, Alfa Katsayısı güvenirlik katsayısını vermez, güvenirliğin alabileceği minimum değeri verir. Testin güvenirlik katsayısı, alfa katsayısından daha yüksek bir değer vermektedir.

2–9. Crocbach Alfa Eşitliği (α)

$$a = \frac{2\left[S^2 t - \left(S_1^2 + S_2^2\right)\right]}{S_t^2}$$

Bu eşitlikte a= Testi yarılama güvenirliği Alfa Katsayısı
S= Testin bütününe ilişkin varyansı
S_2- Testin birinci ve S_2 =Testin ikinci yarısının varyansı

Sadece iki kere uygulanan bir testte güvenirlik kestirmenin bir yolu, iki yarıdan alınan puanların birbiriyle uyumu yerine bütün soruların birbiriyle olan uyumuna bakmaktır. Bu uyuma ilişkin metot aşağıda açıklanacaktır (Öncü. 1995; s.51).

2–10. Kuder Richardson metodu

Test maddelerinin birbirleriyle tutarlılığını esas alan bu metot, test maddelerinin aynı değişkeni ölçtüğü yani testin homojen olduğu varsayımına dayanır. Bu metotla güvenirlik katsayısı kestirmede en çok kullanılan (KR20) ve (KR21) olarak adlandırılan Kuder Richardson formülleridir (Güven, 1990, s.20).

$$KR-20: r_x = \frac{n}{n-1}\left[1 - \frac{\sum p.q}{S_x^2}\right]q$$

Kuder Richardson 20 formülü, doğru cevaplara 1 puan, yanlış ve boş bırakılan maddelere 0 puan vererek puanlama yapılmışsa, yada sorunun cevabının evet veya hayır seçeneklerinden birinin doğru olması durumunda güvenirliği belirlemede kullanılır. Eğer testteki maddeler farklı ağırlıklarla

puanlanmışsa veya test puanları şans başarısı için düzeltilmişse bu formül kullanılmaz (Tekin, 1977, s.48). Çünkü bu formülün elde edilişinde, maddeler arası kovariyansların eşit olduğu varsayılmıştır (Baykul, 1978, s. 16). Bu varsayıma, testteki bütün maddelerin aynı güçlük derecesinde olduğu varsayımının eklenmesiyle K-R20 formülünün özel hali olan K-R21 formülü elde edilmiştir. Bu formül aşağıda verilmiştir:

$$KR-21: r_x = \frac{n}{n-1}\left[1-\frac{n\bar{x}-\left(\bar{x}\right)^2}{S_x^2}\right]$$

n= Testteki soru sayısı S_x= Test puanlarının standart sapması
x= Test puanlarının standart Sapması

K-R21 formülü test puanları ortalaması, standart sapması ve madde sayısına bağlı alarak hesaplandığından madde analizi yapılmamış testlere de uygulanabilmektedir. Oysa K-20 formülü madde analizi yapılmamış testlere uygulanamaz (Turgut 1992; s.34; Adıgüzel. 1985; s.35). KR-20 ile KR-21 arasındaki fark KR-21 eşitliğinin dayandığı önemli varsayımlarından birinin testteki her sorunun güçlük derecesinin aynı olduğu yani, güçlük derecesinin %50 olduğu varsayımıdır. Bu varsayım pratikte nadiren gerçekleşir. Bu nedenle de KR-21 eşitliği ile elde edilen güvenirlik katsayıları test yarılama yöntemi ile elde edilen güvenirlik katsayılarına göre biraz daha düşük çıkmaktadır (Mehrens ve Lehmann 1991, s.256; Özgüven, 1994, s.92).

Testin maddeleri eşit güçlükte değilse KR_21 formülü ile hesaplanan güvenirlik katsayısı, KR-20 ile bulunacak güvenirlik katsayısından genellikle küçük çıkar. Bu sebeple, KR-21 ile hesaplanan değer, testin güvenirlik katsayısının alt sınırı olarak kabul edilir (Turgut. 1983; s.35).

2–11. Cronbach Alfa katsayısı

Kişilik, ilgi ve tutum testleri veya envanterlerinin cevapları derecelendirilmiş ise Cronbach tarafından 1951 yılında geliştirilen ve Cronbach Alfa katsayısı olarak bilinen formülden yararlanılır. Bu formül ve formüldeki sembollerin anlamları şöyledir.

$$a = \frac{n}{n-1}\left[1-\frac{\sum S_i^2}{S_x^2}\right]$$

n=Testteki soru sayısı S^2=Bir maddenin varyansı S=Test puanlarının standart sapması

Bu metot özellikle essey türü sorulardan oluşan bir test için kullanışlıdır (Mehrens ve Lehmann. 1991; s.256).

Hoyt'un varyans analizi

Hoyt (1941)'un varyans analizi yöntemi KR-20 veya α katsayısı ile tamamen aynı sonucu verir. Hoyt'un güvenirlik hesaplama formülü aşağıda verilmiştir:

$$r_{tt} = \frac{Va - Vk}{Va}$$

Vk= Kalan kareler toplamı için varyans

Va= Testi alanlar için varyans

2–12. Puanlayıcı Güvenirliği

Ölçülmek istenen niteliğin değişmesi, tutarlılık tahminleri kullanıldığında tesadüfi bir hata olarak değerlendirilir. Eşdeğer ve testi yarılama tahminleri kullanıldığımızda örnekleme hatası oluşabilir. Eğer aynı testin iki farklı formu uygulanmışsa uygulayıcı hatası olabilir (Öncü. 1995; s.54).

Güvenirliği etkileyen faktörlerde şunlardır;
1 . Testin uzunluğu
2. Test maddelerinin ifadesi
3. Testin açıklaması (yönergesi)
4. Grubun homojenliği
5. Testin güçlüğü
6. Testin süresi
7. Puanlamadaki objektiflik
8. Ölçmenin yapılış şartları
9. Güvenirliği hesaplama yöntemi

SONUÇ

Öğrencilerin öğrenme ortamında edindiklerini ortaya koyabilecek bir faaliyete geçmedikleri sürece onları gözlemek (ölçmek) ve değerlendirmek mümkün değildir. Bazı durumlarda öğrenmeleri gözle görülebilir hale getirmek güçtür. Çünkü bir kimsenin zihinsel aktivitesine dayalı bilgilerin pek çoğunu göremediğimiz gibi bazı durumlarda kısmi yansımalarını bile göremeyebiliriz.

Öğrencilerin düşüncelerini ortaya çıkarmak için, öğrenilenlerin göstergesi olabilecek durumlar işe koşulur. Davranışı ortaya çıkarıcı uyarıcılar sunarak tepkilerini almak en uygun yoldur.

Hiçbir bilim dalı yoktur ki, alanı ile ilgili gelişmeyi sağlayacak, ampirik ilişkiler sistemi ve nitel farklılıkları biçimsel bir matematik sistemi ve özellikle niceliksel farklılıklara dönüştürme çabasına girmesin. Eğitim bilimleri de, bir bilim dalı olarak nitel ve nicel verilerin belli bir matematiksel form içinde, benzerlik, farklılık, denklik gibi özelliklerini tahmin etmek maksadıyla ölçme araçlarından faydalanmaktadır.

Eğitim sahasının ele alıp işlediği materyal insan olduğu için, insan hem kendi iç dinamiklerinin, hem de dış çevresinin etkisi altındadır. Bu nedenle özelliklerinde kararlılık göstermeyen, anlık değişebilen bir karakteristiğe sahiptir. Değişkenlerin çokluğu, ölçme sonuçlarında sapmanın yönünü, derecesini ve miktarını artırmaktadır. Bilindiği gibi ölçmenin temel esprisini "farkın" ortaya konması oluşturur.

İnsan davranışlarının ölçümünde farklılığa neden olabilecek değişkenlerin (varyansların) bilinmesi, değişkenlerin kontrol altına alınabilmesi açısından önem kazanır.

Ve bir başka açıdan ölçme aracında geçerlik derecesinin varlığından veya yokluğundan bahsetmek yerine, belli bir maksatla belli bir özelliği ölçme işine girişildiğinden söz etmek daha uygundur. Böyle bir sonucu ifade ederken, ölçme aracından elde edilen verilerin geçerliği nedir şeklinde dile getirmek durumundayız. Hazırlanan bir ölçme aracı (test) ölçmeyi hedeflediği davranışlara ilişkin sorulardan oluşmasına rağmen, ölçme aracında (testte) yer alan bir veya birkaç maddenin hedeflenen davranışı ortaya çıkaramaması durumunda, ölçme aracının bütününün geçerliğinin düşmesine neden olur. Ancak ölçme aracının (testin), tamamen geçersiz olduğu anlamına gelmez. Bu durumu dikkate alan Wallen bir testin güvenirliğini; toplanan verilerin incelenmesi ile elde edilen "uygunluk", "anlamlılık" ve "faydalılık" olarak dile getirmiştir (Wallen. 1993; s. 139).

Ölçmede kullanılacak bir ölçme aracının, ölçmenin yapılış maksadına hizmet etmesi beklenir. Ölçme aracının ölçmek istediği özellikler arasına başka bir özelliği karıştırmaması, ölçülen özelliğe yönelik değerlendirmeler için yeride ve doğru kararlar alınmasını mümkün kılacaktır. Geçerliği şöyle bir örnekle açıklamak konuyu daha somutlaştıracaktır.

Öğretmen yani biz öğrencilerin matematik seviyelerini tespit etmek için hazırladığı soruların arasında, fizik sorularına da yer vermişse, aynı sınav birkaç kez yenilendiğinde aynı yada benzer sonuçlar almak olasıdır. Ay-

nı ve benzer ölçümler vermesi o testin tutarlılığına dolayısıyla geçerliğine delalet eder. Ancak maksat öğrencinin matematik konularındaki yeterliğini belirlemek olduğundan, test öğrencinin matematik yeterliğini ölçecek geçerlikte değildir. Böyle bir sınavın sonuçlarına bakılarak, öğrencinin matematikteki yeterliğinden söz edilemez.

KAYNAKLAR

Adıgüzel, M. (1985). *"Madde Yapısının Test Güvenirliğine ve Geçerliğine Etkisi"*, (Yayınlanmamış Doktora Tezi) Ankara: Hacettepe Üniversitesi.

Agun, K., Inamo, E., & Olcy, E., (2002). **"**Value domains of Turkish adults and university students students." *Journal of Socual Psychology,* 142(3), 333-353.

Albatsh & Abderhanan. (1990). "Value structure of university of Jordan students." *Dirastat,* 17(3), 92-136.

Alkhaldi, M.A., & Aljabri, I,M**.** (1998). **"**The relationship of attitudes to computer utilization: new evidence from a developing nation**."** *Computers in Human Behavior,* 14(1), 23-42.

Arıcı, H. (1984). İstatistik: Yöntemler ve Uygulama. Ankara

Attoum, A., & Khasawneh**, A.** (1999). "Value matrix of Al-albaty university students." *Al Manar*, 1(1), 3-54 (Arabic Version).

Balcı, A. (1995). Sosyal Bilimlerde Araştırma: Yöntem, Teknik ve İlkeler. Ankara

Beydoğan, Ö. H. (2003). Öğretimde Planlama ve Değerlendirme, Erzurum.

Brosnan, M.,&Lee, W. (1998). **"**A cross-cultural comparison of gender differences in computer attitudes and anxieties: the United Kingdom and Hon Kong." *Computers in Human Behavior,* 14(4), 559-577.

Cronbach, Lee. J. (1990). Essentials of Psychological Testing. Fifth edition, Harper Collins Publishers,.

Fraenkel, J.R. and Norman E. Wallen., How to Design and Evaluate Research in Education (2 nd Ed) Mc. Grow Hill. Inc., 1993.

Francis, L.L., Katz, Y.J., & Jones, S.H**.** (2000). "The reliabity and validity of the Hebrew version of the computer attitude scale*" Computers & Education*, 35(2), 329-340

Gronlund, N.E. (1976). Measurement and Evaluation Teaching. The Macmillan Cem., New York

Güven, Ç. (1990). "Envanter Maddelerinin Analizinde Klasik Test ve Bilgi Kuramı Yöntemlerinin Karşılaştırılması", (Yayınlanmamış Doktora Tezi) Ankara, Hacettepe Üniversitesi

Karasar, N. (1982). Bilimsel Araştırma Yöntemi: Kavramlar, İlkeler, Teknikler. Nobel Yayınları, Ankara

Mehrens, W. A. and Irvin J. L (1991). Measurement and Evalualion in Education and Psychology. Fourth Edition, New York:Holt, Rinehart ve Winston

Öncü, H. (1995). Eğitimde Ölçme ve Değerlendirme, Geliştirilmiş İkinci baskı, Ankara

Özçelik, D. A. (1989). Test Hazırlama Kılavuzu**.** Genişletilmiş ikinci baskı. Ankara: ÖSYM Eğitim Yayınları

Özgüven, İ. E. (1995). Psikolojik Testler. PDREM yayınları, Ankara, 1994.

Robertson, S. I., Calder, J., Fung, P., Jones, A., & O'Shea, T. (1995). "Computer attitudes in an English secondary school." *Computers & Education*, 24(2), 73-81.

Tekin, H. (1977). Eğitimde Ölçme ve Değerlendirme. Ankara,

Uysal, Ş. (1976). "Sosyal Bilim Arşt. Kullanılan Araçların Geçerlik ve Güvenirliği", *Toplum Bilimlerde Araştırma ve Yöntem*, TOBİE, Ankara

Valois, P., Frenette, E., Villeneuve, P., Sabourin, S., & Bordeleau, C. (2000). "Nonparametric item analysis and confirmatory factorial validity of the computer attitude scale for secondary students." *Computers&Education*, 35 (4), 281-294