

Tıbbi arařtırmalarda güvenilirlik ve geçerlilik

Reliability and validity in medical research

Zekeriya Aktürk¹, Hamit Acemođlu²

¹ Atatürk Üniversitesi Tıp Fakültesi Aile Hekimliği AD, Erzurum, Türkiye

² Atatürk Üniversitesi Tıp Fakültesi Tıp Eğitimi AD, Erzurum, Türkiye

Geliş Tarihi / Received: 30.09.2011, Kabul Tarihi / Accepted: 05.03.2012

ÖZET

Tıbbi arařtırmalarda sıklıkla arařtırma ölçeklerine başvuru-
ruruz. Sadece psikometrik ölçümler için deđil, her çeřit
ölçümde güvenilirlik ve geçerlilik kavramlarını gündeme
getirebiliriz. Bu yazıda güvenilirlik ve geçerlilik kavram-
larının örneklendirilerek açıklanması amaçlanmıştır. Bir
ölçeđin/ölçümün güvenilirliğini ve geçerliliđini bilimsel
yöntemlerle deđerlendirmek mümkündür. Bir ölçeđin gü-
venilirliğinden bahsedince akla stabilliliđi (aynı örnekleme
yapılan tekrarlayan ölçümlerden aynı sonucun alınması
, eřdeđerliliđi ve homojenliliđi gelir. Homojenlik açasından
güvenilirlik ölçeđin iç özelliđiyle (iç tutarlılık "internal con-
sistency") ilgilidir; aynı yapıyı ölçen maddelerin ne kadar
benzer sonuçlar verdiđinin ölçülmesidir. Yani aynı özelliđi
ölçmede farklı maddelerin ne kadar tutarlı olduđunun
incelenmesidir. Bu kavramlara göre ařađıdaki güvenilirlik
çeřitleri ile ölçebiliriz: paralel formlar, test/tekrar test (test/
retest), gözlemciler arası güvenilirlik, yarıya bölme (split
half) ve Cronbach alfa. Geçerlilik açasından içerik geçer-
liliđi (Content validity), tahmin ettirici geçerlilik (Predictive
validity), yapı geçerliliđi (Construct validity), eř zamanlı
geçerlilik (Concurrent validity) ve Görünüm geçerliliđin-
den (Face validity) bahsedilebilir. Bilim ölçüm demektir.
Yanlıř ölçümler yapmak kanıtlarımızın ve tedavilerimizin
güvenilmez olmasına yol açar. Bu nedenle ölçüm ara-
larımızın geçerli ve güvenilir sonuçlar vermesine azami
önem göstermeliyiz.

Anahtar kelimeler: ölçme, geçerlilik, güvenilirlik

GİRİř

Ölçme, "Bir niteliđin gözlenip, gözlem sonucunun
sayı ve sembollerle gösterilmesidir"¹ İstatistikte gü-
venilirlik, ölçme aracının kendi içinde kararlılıđı ve
tutarlılıđı olup, standart hatanın az olması demek-

ABSTRACT

Scientists commonly refer to study instruments during
medical research. In fact, the reliability and validity issues
go beyond psychometric studies and can be linked with
any kind measurements. In this study we aimed to explain
the reliability and validity concepts by giving examples.
It is possible to evaluate the reliability and validity of an
instrument by scientific methods. If we speak of reliability,
we have to mention stability (having the same results in
repeated measurements from the same sample), equiva-
lence, and homogeneity. Homogeneity is related with
internal consistency; it measures how close results are
obtained from items intending to measure the same struc-
ture. In other words, how consistent are the different items
in measuring the same feature? Accordingly, the following
types of reliability can be measured: parallel forms, test/
retest, inter-observer reliability, split half, and Cronbach
alfa. From the point of validity, we will discuss content
validity, predictive validity, construct validity, concurrent
validity, and face validity. Science means measurement.
Wrong measurements will make our evidence and thus
the treatments unreliable. Therefore, medical researchers
have to give utmost importance in receiving valid and reli-
able results from the instruments they use.

Key words: measurement, validity, reliability

tir.² Geçerlilik ise bir ölçme aracının ölçmek istediđi
deđiřkeni ölçüp ölçmediđi, ölçüyorsa onu bařka de-
điřkenlerden ne derece ayırarak ölçtüđüdür.

Bir testin söz konusu bir durumu ölçebilmesi
için (a) bahse konu edilen durum var olmalıdır ve
(b) ölçümü hedeflenen durumdaki deđiřimler ölçüm

Yazıřma Adresi /Correspondence: Dr. Hamit Acemođlu

Atatürk Üniversitesi Tıp Fakültesi Tıp Eğitimi AD, Erzurum Eposta: hamitacemoglu@yahoo.com
Copyright © Dicle Tıp Dergisi 2012, Her hakkı saklıdır / All rights reserved

sonuçlarını da deęiřtirmelidir.³ Tıbbi arařtırmalarda sıklıkla arařtırma ölçeklerine başvurulur. Yařam kalitesi için SF-36,⁴ ağrı řiddeti için WOMAC,⁵ depresyon için Beck depresyon ölçeęi,⁶ benlik saygısı için Rosenberg⁷ kullanıldıęı gibi, daha birçok saęlık durumunu deęerlendirmek için ölçekleri kullanırız.

Aslında sadece psikometrik ölçümler için deęil, her çeřit ölçümde güvenilirlik ve geçerlilik kavramlarını gündeme getirebiliriz. Örneęin biyokimyasal analiz yapan cihazların da doęru ölçüp ölçmediklerinden ve her defasında aynı sonucu vermelerinden emin olmak gerekir.

Bu yazıda güvenilirlik ve geçerlilik kavramlarının örneklendirilerek açıklanması amaçlanmıřtır.

Güvenilirlik ve Geçerlilik Kavramları

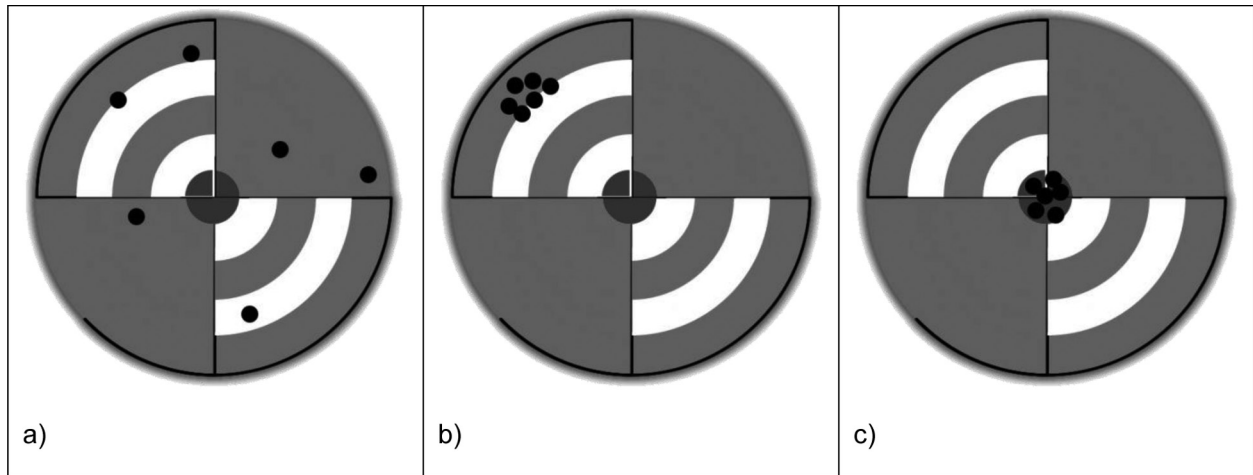
Özellikle sözlü sınavlarda řikayet edilen bir durum vardır: eęitici iyi bir sınav yaptıęını düşünür ama öęrenciler aynı görüşte olmayabilir. Sınavın eęitimin içerięini ne kadar ölçtüęü sorgulanabileceęi

gibi, aynı sınava tekrar girilse geçme/kalma durumunun çok farklı olacaęı da iddia edilebilir.

Bu durumda alt ekstremite kemiklerinin anlatıldıęı anatomi dersinin sınavında birkaç çeřit eęitici den bahsedebiliriz. Birinci eęitici tipi sınavda hem konuyla alakasız sorular sorar, hem de her öęrenciye farklı içerikte sorular sorar. Hedefi tuturmaktan uzak olan ve hem de her ölçümde farklı sonuçlar veren bu ölçüme “Hem güvenilirmez hem de geçersiz” ölçüm denilebilir (Şekil 1a).

İkinci eęitici tipi ise yine anlatılan içerikle ilgili sorular sormaz ama tüm öęrencilere örneęin üst ekstremite kemikleriyle ilgili sorular sorar. Bu eęiticinin ölçümü için “Güvenilir ama geçersiz” bir ölçüm diyebiliriz (Şekil 1b).

Üçüncü eęitici tipi ise sorularını alt ekstremitte kemikleri hakkında hazırlamıřtır ve tüm öęrencilere de aynı içerikte sorular sormaktadır. Bu eęiticinin ölçümü için ise “Hem güvenilir, hem de ama geçerli” bir ölçüm diyebiliriz (Şekil 1c).



Şekil 1. a) Güvenilmez ve geçersiz ölçüm, b) Güvenilir ama geçersiz ölçüm, c) Güvenilir ve geçerli ölçüm.

Şüphesiz her ölçüm aracından hem güvenilir, hem de geçerli olması beklenir. Ancak, tamamen tutarsız bir eęiticideyse, en azından yanlış yerden de sınav yapsa hep aynı “ters köşeye yatıran” eęiticinin tercih edileceęi gibi, güvenilir olmayan bir ölçek için geçerlilięin bir deęerinin olmayacaęını söyleyebiliriz.

GÜVENİLİRLİK ÇEŞİTLERİ

Bir ölçeęin güvenilirliğinden bahsedince akla stabillięi (Aynı örneklemede yapılan tekrarlayan ölçümlerden aynı sonucun alınması halinde ölçeęin stabil olduęu söylenir.), eşdeęerlilięi ve homojenlilięi gelir.⁸ Homojenlik açısından güvenilirlik ölçeęin iç özellięiyle ilgilidir (iç tutarlılık da “internal consistency” denir). Aynı yapıyı ölçen maddelerin ne kadar benzer sonuçlar verdięinin ölçülmesidir. Yani aynı özellięi ölçmede farklı maddelerin ne kadar tutarlı olduęunun incelenmesidir. Bu kavramları açacak olursak ařaęıdaki güvenilirlik çeřitlerinden bahsedebiliriz:

1. Paralel formlar

Paralel form güvenilirliğine bakmak için aynı kişilere iki farklı ölçek uygulanıp aralarındaki korelasyona bakılabilir. Korelasyon ne kadar yüksekse ölçüklerin o kadar eşdeğer olduğunu söyleriz. Ölçülen aynı şeydir. Sadece ifade tarzları ve/veya soruların tasarımı farklıdır.

Bunu yapmanın bir yolu, soruları hazırlamak ve rastgele ikiye ayırarak uygulamaktır. Paralel formlar uygulaması ölçeğin stabilliğini de ölçer.

2. Test/tekrar test (Test/retest)

Aynı araştırma aracı aynı deneklere benzer şartlar altında 2 veya daha fazla kez uygulanır. Burada iki ölçüm arasında fazla bir farklılık olmamalıdır. Tabii ki, ölçüm yapılan zaman aralığının ne kadar olduğu da önemlidir. Uygulama ne kadar erken tekrarlanırsa o kadar benzer sonuçlar elde edilir. Diğer taraftan çok geciktirilmesi ölçülen durumun değişmesine yol açabilir. 2-4 hafta gibi zaman aralıkları genel anlamda uygun kabul edilse de (9) hafıza faktörü, bireyin konuya duyarlaşması ve zaman içerisinde oluşabilecek değişiklikler de dikkate alınarak tekrar test yapılmalıdır.

3. Gözlemciler arası güvenilirlik

Uygulayıcılar arası (Interrater) güvenilirliği tek bir formun iki uygulayıcı tarafından uygulanması ve aralarındaki korelasyona bakılması ile ölçülür.

Ölçek kategorik bir ölçüm yapıyorsa (evet/hayır gibi) iki araştırmacının uygulamasında ne kadar uyum olduğuna bakılır. Aralarındaki uyum (örn. %82) rapor edilir. Ölçek nümerik bir ölçüm yapıyorsa iki araştırmacının uygulamasının ne kadar korelasyon (Intraclass Correlation Coefficient - ICC) gösterdiğine bakılır. Phi (basit korelasyon), Kappa (rastlantı açısından düzeltme yapılmış) ve Kendall's tau (sıralı veriler için) katsayıları hesaplanabilir.

4. Gözlemci içi güvenilirlik

Gözlemci içi (intra-rater) güvenilirlik aynı değerlendiricinin yaptığı birden fazla ölçümün arasındaki uyum derecesidir. Aynı gözlemcinin aynı ölçüm araç ve gereçlerini kullanarak yaptığı ölçümler birbirinden farklılık gösterebilir. Anlaşılacağı gibi bu kullanılan ölçüm araçlarına değil, araştırmacıya bağlı bir durumdur. Ölçümün nümerik olduğu du-

rumlarda sınıf içi korelasyona (intra-class correlation), kategorik olduğu durumda ise Cohen'in kappa katsayısına bakılarak değerlendirme yapılabilir.

5. Yarıya bölme (Split half)

Bir özelliği ölçmek için kullanılan tüm maddeler rastgele ikiye ayrılır. Ölçek bir grup bireye uygulanır ve her iki yarımın puanları hesaplanır. Bu iki yarımın karşılaştırılmasıyla (Guttman Split-Half katsayısı) güvenilirliğin derecesi belirlenir.

6. Cronbach alfa

Cronbach alfa yarıya bölmenin (matematiksel anlamda) eşdeğeridir. Güvenilirlik hesaplarında sıkça kullanılan bir katsayıdır. Maddeler arası korelasyon ortalamasını da dikkate alarak iç güvenilirliği hesaplar.

Cronbach alfa hesaplamasında ölçek maddeleri rastgele ikiye ayrılarak karşılaştırılır. Bu rastgele ikiye ayırma işlemi tüm ihtimaller için tekrarlanır. Benzer bir ölçüm de Kuder-Richardson'dur.

Özetleyecek olursak, güvenilirlik açısından paralel formlar ve uygulayıcılar arası güvenilirlik testin eşdeğerliliğini, yine uygulayıcılar arası güvenilirlik ve test/tekrar testin stabilliğini, yarıya bölme, Kuder-Richardson ve Cronbach alfa gibi ölçümler ise homojenliğini belirler.

GEÇERLİLİK ÇEŞİTLERİ

Geçerlilik açısından içerik geçerliliği (Content validity), tahmin ettirici geçerlilik (Predictive validity), yapı geçerliliği (Construct validity), eş zamanlı geçerlilik (Concurrent validity) ve Görünüm geçerliliğinden (Face validity) bahsedilebilir.

1. İçerik geçerliliği

İçerik geçerliğinden anlaşılan ölçeğin içeriğinin gerçekten ölçülmesi hedeflenen durumla ilgili olup olmamasıdır. Depresyonu taramak için oluşturduğumuz bir ölçekte keyifsizlik, suçluluk hissi, intihar düşüncesi gibi maddeler bekleriz; gastrointestinal kanamayla ilgili soruların olması içerik açısından geçersiz olduğunu düşündürür.

İçeriğin boyutlarının belirlenmesini belki ölçek geliştirmedeki en zor kısımdır. Bu amaçla bir uzmanlar grubundan yararlanılabilir ve literatür desteği gerekir.

2. Tahmin ettirici (=criterion= predictive =ölçüt) geçerlilik

Araştırma aracının gerçek yaşamda durumları ne kadar tahmin ettirici olduğuyla ilgilidir. Depresyon ölçeğinde intihar riski saptananların ne kadar intihar ediyor? Ya da trafik sınavında yüksek puan alanlar trafikte ne kadar iyi araç kullanıyor?

3. Yapı geçerliliği

Aracın ölçülmeye çalışılan teorik psiko sosyal yapı ile ne kadar korelasyon gösterdiği ile ilgilidir. “Bu ölçek ölçmeye çalıştığımız fenomeni ne kadar ölçüyor?” sorusuna cevap aranmasıdır. Altta yatan fenomenle ilgili farklı konseptleri ölçmeye çalışır. Bu amaçla madde analizi yapılabilir.

4. Eşzamanlı geçerlilik

Eş zamanlı geçerliliği test etmek için ölçek aynı veya ilişkili bir yapıyı inceleyen ve daha önce geçerliliği ispat edilmiş başka bir ölçekle eşzamanlı olarak uygulanır. Bu da tahmin ettirici geçerlilik gibi bir ölçütü tahmin etmeye ne kadar yaradığını gösterir. Yeni geliştirilen depresyon ölçeğinin Beck depresyon ölçeği ile birlikte uygulanmasını örnek olarak verebiliriz.

5. Görünüm geçerliliği

Bir arabanın hızının dış görünüşünden tahmin edilmesi gibidir. Maddelerin görünüşü, okunabilirliği, uygulama kolaylığı gibi konular açısından değerlendirme yapılır. Bu amaçla Tablo 1’deki soruların sorulması faydalı olabilir.

Tablo 1. Bir ölçeğin görünüm geçerliliğini değerlendirmek için katılımcılara sorulabilecek sorular

- 1 Anket hakkındaki genel görüşleriniz nelerdir?
- 2 Açıklamalarla ilgili görüşleriniz nelerdir?
- 3 Bu anketi doldurmanız ne kadar sürdü?
- 4 Soruların sayısı kabul edilebilir mi?
- 5 Soruların sırası mantıklı mı?
- 6 Cevap vermede zorlandığınız veya anlayamadığınız sorular oldu mu?
- 7 Genel olarak anketin anlaşılabilirliği ve sadeliği nasıldır?
- 8 Soruların derecelendirmesiyle ilgili problem yaşadınız mı, bu konuda başka bir öneriniz var mı?
- 9 Anketle ilgili herhangi başka bir öneriniz var mı?
- 10 Anketi doldurmak için yardıma ihtiyaç duydunuz mu? Kim yardım etti?

SONUÇ

“Bilim ölçüm demektir. Yanlış ölçümler yapmak kanıtlarımızın ve tedavilerimizin güvenilmez olmasına yol açar”¹⁰ Bu nedenle ölçüm araçlarımızın geçerli ve güvenilir sonuçlar vermesine azami önem göstermeliyiz.

Nasıl bir tansiyon aletinin veya laboratuvarında kullandığımız bir cihazın kalibre edilmiş olması gerekiyorsa, psikometrik ölçümler için kullandığımız ölçeklerin de geçerli ve güvenilir olması önemlidir. Araştırmalarda kullanacağımız uluslararası ölçeklerin Türkçe geçerliliğinin olup olmamasının yanında kendi hazırladığımız sınav ve soruların da geçerli ve güvenilir olmasına dikkat etmeliyiz.

KAYNAKLAR

1. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Fort Worth: Holt, Rinehart and Winston; 1986.
2. Meeker WQ, Escobar LA. Statistical Methods for Reliability Data. Hoboken, New Jersey: Wiley; 1998.
3. Borsboom D, Mellenbergh GJ, Heerden Jv. The concept of validity. *Psychological Review* 2004;111(4):1061-1.
4. Filiz TM, Topsever P, Uludağ C, Görpelioglu S, Çınar N. Türk kadınlarında üriner inkontinans şiddeti ve yaşın jenerik yaşam kalitesi sf-36 üzerine etkileri. *Türkiye Klinikleri J Med Sci* 2007;27(2):189-4.
5. Paker N, Buğdaycı D, Sabırlı F, Özel S, Ersoy S. Diz incinme ve osteoartrit sonuç skoru: Türkçe sürümünün güvenilirlik ve geçerlilik çalışması. *Türkiye Klinikleri J Med Sci* 2007;27(3):350-6.
6. Aktürk Z, Dağdeviren N, Türe M, Tuğlu C. Birinci basamak için beck depresyon tarama ölçeği'nin türkçe çevriminin geçerlik ve güvenilirliği. *Türkiye Aile Hekimliği Dergisi* 2005;9(3):117-2.
7. Sayar K, Bilen A, Arıkan M. Kronik ağrı hastalarında öfke, benlik saygısı ve aleksitimi. *T Klin J Psychiatry* 2001;2:36-42.
8. Bannigan K, Watson R. Reliability and validity in a nutshell. *J Clin Nursing* 2009;18(23):3237-3.
9. Carmines EG, Zeller RA. Reliability and validity assessment. California: Sage Publications; 1979.
10. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med* 2006;119(2):e7-6.