Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Yaz 2010, 1(1), 16-21

Test Eşitleme : Aynı Davranışları Ölçen, Farklı Madde Formlarına Sahip Testlerin İstatistiksel Eşitliğinin Sınanması

Adnan KAN*
Gazi Üniversitesi

Özet

Bu araştırma ile aynı davranışları ölçen fakat farklı madde formlarına (sözel ya da sayısal) sahip matematik testlerinin istatistiksel eşitliğini sınamak, bu yolla madde formunun öğrencilerin performansı üzerinde etkili olup olmadığını belirlemek amaçlanmıştır. Çalışma çeşitli ilköğretim kurumlarında öğrenim gören 420 altıncı sınıf öğrencisi üzerinde yürütülmüştür. Bu amaç doğrultusunda klasik eşitleme metotlarından lineer eşitleme ve tek grup düzeneği kullanılmıştır. Aynı zamanda eşitleme işleminin bir parçası olarak tek grup düzeneği için eşitleme hatası (SHE) kestirilmiştir. İki testin eşitliğini değerlendirmek için eşitlemenin standart hatasına dayalı güven aralıkları kullanılmıştır. Sonuç olarak, eşitlik fonksiyonu ve özdeşlik fonksiyonu arasında -0,041 ile 1,159 puan arasında değişen farklılıklar olduğu saptanmıştır. Bu farklılıkların bazı puan ranjlarında 2 SHE'den fazla olması sebebiyle aynı davranışları ölçen sayısal köklü matematik testi ile sözel köklü matematik test versiyonlarının istatistiksel olarak birbirine eşit olmadığı ve birbirinin yerine kullanılamayacağı saptanmıştır.

Anahtar sözcükler: test eşitleme, doğrusal eşitleme, tek grup deseni, test yansızlığı, geçerlik

Abstract

In this study, firstly it was aimed to see whether the tests with different item format create advantage or disadvantage on examinees' performance, in other words, whether the examinees' scores are fairly affected by particular test forms with different formatted items. Second purpose is to make use of test equating procedure as a tool for examining the effect of item formats on test performance and by this way identifying whether the two different item formatted test scores could be used interchangeably. This was questioned by searching whether the test scores obtained from different item formats could be used interchangeably. The data of the study was collected from 402 6th grade students from various secondary schools. Single group design, and linear equating procedure from classical equating methodologies were used in the study. At the same time, as a part of the equating process, SEEs (Standart error of equating) for single group were estimated. Confidence bands based on the SEE was used to assess equivalence of different item formatted mathematic test edition. As a result of the study, differences were found between equating function and identity function, differences ranging from –0,041 to 1,159. Because these differences are more than two SEEs for some score range, the two different item formatted mathematic test edition (numeric formatted and word formatted) can not be considered equivalent and interchangeable.

Test geliştiricilerin veya testi uygulayan kurumların istatistiksel açıdan ve kapsam açısından aynı veya birbirine benzer testler oluşturma çabalarına rağmen her bir oturumda farklı test versiyonları ve farklı soru setleri kullanıldığı sürece özellikle testlerin güçlük düzeylerinde bir miktar farklılıklar olacaktır (Tanguma, 2000). Bu tür geniş test programları çeşitli kanuni, psikometrik ve pratik konuları gözönünde tutmalıdır. Bu konulardan birisi bir testin, aynı özelliği ölçen farklı formlarını oluşturmaktır. İki farklı oturumda aynı testi farklı

ya da aynı bireylere uygulamak pek duyulmuş bir şey değildir. Böyle bir uygulama testi sonra alanların, önce alanlara göre kesin bir şekilde daha avantajlı olmasını sağlayacak ve test güvenliğini tehdit edecektir ve doğal karşılanamaz. Fakat iki testin psikometrik açıdan birbirine eşit (eşdeğer)

* Doç. Dr., Gazi Üniversitesi, Gazi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, adnankan@gazi.edu.tr

olduğu kanıtlandığı sürece, aynı testin farklı versiyonlarının, farklı bireylere uygulanabilmesi kanuni olarak savunulabilir. Bütün bunların yanında test geliştiricilerin veya kurumların tamamıyla birbirine

paralel, fakat farklı sorulardan oluşan ve testi alan her bir birey için aynı veya benzer sonuçlar üreten testler yapılandırmaları oldukça zordur ve beklenemez.

Ülkemizde gerek öğretmen yapımı testlerin gerekse her yıl uygulanan seçme amaçlı bir çok sınavların her birinin farklı soruları içermesi fakat aynı özelliği ölçmesi ölçme aracının yanlı davranmaması için arzu edilen bir durumdur. Bir ölçme aracına ait soruların paralellerini oluşturmanın bir yoluda aynı özelliği ölçen fakat farklı formlarda (sayısal, sözel, sembolik vb.) sorular hazırlamaktır. Şimdi A ve B gibi iki öğrenci olduğunu ve bu iki öğrencininde aynı özelliği ölçen fakat farklı sorulardan oluşan formları aldığını varsayalım. Bu aynı özelliği ölçen, farklı veya aynı zamanlarda uygulanan ve farklı formatta sorulardan oluşan sınavlardan A öğrencisi Ax, B öğrenciside Bx gibi bir puan almış olsun. Bu bir testin farklı veya aynı zamanlarda uygulanmış versiyonlarından elde edilen puanlar birbiriyle karşılaştırılabilir mi ? Testin bir başka versiyonunu aldığı için B öğrencisinin A öğrencisine göre veya A öğrencisinin B öğrencisine göre daha avantajlı olmadığını güvenle söyleyebilirmiyiz ? Bir başka ifade ile farklı zamanlarda uygulanan bu testlerin testi alan bireylere adil davrandığını iddia edebilirmiyiz ? Aynı özelliği ölçen, genellikle seçme amaçlı ve her sene yapılan sınavlarda farklı zamanlarda formatı farklı benzer sorulara yer verilebilmektedir. Aynı soru bazen grafikle, bazen daha sözel veya sayısal ve sembol ağırlıklı sorulabilmektedir. Şüphesiz aynı özellikleri ölçen paralel formlar yaratmak ve testin testi alan bireylere adil davrandığını söyleyebilmek için iyi bir yol gibi gözükmektedir. Acaba aynı davranışı ölçmeye dönük ve aynı amaca yönelik hazırlanan fakat farklı formatta hazırlanan bu maddeler birbirinin yerine kullanılabilir mi? Bu testlerden elde edilen puanlar birbirinin yerine kullanılabilir mi? Bir başka ifadeyle madde formatının bireylerin peformansı üzerinde etkisi varmıdır? Bu gibi sebeplerden dolayı yukarıda belirtilen bir çok test programı her bir oturumda belli özellikleri ölçen bir testin farklı sorulardan oluşan ama mümkün olduğunca aynı güçlük düzeyine ve kapsama sahip bir çok versiyonunu kullanır. Psikometrik olarak ideal olan, bir testin farklı formlarının tamamıyla paralel olması, testi alan tüm grupların random olarak seçilmiş olması, eşit yetenek düzeyine sahip olması ve etik ve kanuni konuların test programını sınırlamamasıdır. Bütün test programlarında karşılaşılabilecek bu ve benzeri konular ve sorunlar test eşitleme olarak bilinen prosedürlerden yararlanmayı gerekli kılar (Woldbeck, 1988). Test eşitleme metodları farklı test formlarından elde edilen puanları birbiriyle karşılaştırılabilir hale getirmek için kullanılır (Davier, Holland & Thayer, 2004 ; Holland, Sinharay, Davier, & Han, 2008; Harris, 2003).

Test Eşitleme

Angoff (1971), test eşitlemeyi bir formun birim sistemini diğer formun birim sistemine dönüştürmek olarak tanımlar. Angof (1982)'ye göre X ve Y gibi birbirine paralel ve eşit güvenirlik düzeyine sahip iki farklı test formundan elde edilen iki puanın birbirine eşit olduğu, eğer bu iki testten elde edilen standart puanlar birbirine eşitse iddia edilebilir. Bir başka ifade ile ; bir testin iki farklı formundan elde edilen iki puan ;

$$\frac{X-\mu_X}{S_X}=\frac{Y-\mu_Y}{S_Y} \qquad (1)$$

koşulunu sağlıyorsa birbirine eşittir. Y eşitlikten çekilerek, gerekli matematiksel düzenlemeler yapılırsa ;

$$Y = \frac{S_Y}{S_X} X + \mu_Y - \frac{S_Y}{S_X} \mu_X \qquad (2)$$

denklemi elde edilir.

Formülde yer alan $\frac{S_Y}{S_X}$ bileşenine "a" ve $\mu_Y - \frac{S_Y}{S_X} \mu_X$ bileşeninede "b" denirse,

Eğimi "a" kesme noktası (sabit) "b" olan Y= ax + b şeklinde bir doğru denklemi elde edilir. Bu aynı zamanda doğrusal eşitlemeninde temel denklemidir. Doğrusal eşitleme, paralel olarak yapılandırılan iki test formuna ait ham puan dağılımlarının özdeş olduğunu fakat eşitlenecek test puanlarının sadece ortalama ve standart sapmalarının farklı olduğunu varsayar (Lord & Novick 1968).

Test eşitlemenin, farklı test formlarını alan bireyler arasında yanlılığı önlemek, farklı formlardan alınan puanları aynı ölçek üzerinde rapor etmek ve rapor edilen puanların anlamını korumak gibi iki önemli amacı vardır (Barnard, 1996).

Farklı formlardan elde edilen puanların eşitlenmesi ile bireylerin gelişimini ölçmek, eğilimlerini belirlemek ve performanslarını karşılaştırmak mümkün olabilir. Üniversite sınavları, kolej giriş sınavları gibi seçme gerektiren ve testin farklı formlarını içeren sınavlara ilişkin ölçme sonuçlarının eşitlenmesi ve eşitleme sonuçları, sınavın geçerliği açısından önemlidir. Çünkü başvuranların hangi formu aldıkları (kolay mı, zor mu) ve herhangi bir gruba avantaj sağlayıp sağlamadığı göz önünde tutulması gereken önemli bir konudur (Angoff, 1971).

Tek grup deseni için eşitlemenin standart hatası ise; N: toplam birey sayısı, z_x: Y* 'ye dönüştürülen X puanlarına ilişkin Z puanları, ρ_{XY}: iki testten elde edilen puanlar arasındaki korelasyon olmak üzere,

$$SE^2_{Y*} = \frac{\sigma^2_Y (1 - \rho_{XY})}{N} \left[z_x^2 (1 + \rho_{XY}) + 2 \right] \qquad (3)$$

şeklinde tanımlanabilir.

Operasyonel olarak eşitlik 3'de tanımlanan eşitlemenin standart hatası eşitleme fonksiyonunun etrafında makul bir güven aralığı tanımlamak içinde kullanılabilir. Birbirine paralel olduğu varsayılan ve aynı beceriyi ölçen fakat farklı sorulardan oluşan OKS test formlarından elde edilen puanları için makul bir güven aralığının sınırları her bir ham puan etrafında ∓ 2 SHE olarak tanımlanabilir (Dorans & Lawrence, 1990). Güven aralığı belirlendikten sonra farklı test versiyonlarından elde edilen puanların istatistiksel eşitliği, özdeşlik fonksiyonunun tanımlanan bu güven aralığı içerisinde yer alıp almadığını belirlemek suretiyle sınanabilir.

Çalışmanın amacı

Bu çalışma teorik ve pratik anlamda iki amaç doğrultusunda yürütülmüştür. Çalışmanın birincil amacı ülkemizde ve dünyada çok fazla kullanılmayan fakat önemli bir prosedür olan test eşitlemenin farklı eğitim problemleri üzerinde nasıl kullanılacağını göstermek, ikincil amacı ise aynı davranışları ölçen fakat farklı madde formlarına (sözel yada sayısal) sahip matematik testlerinin istatistiksel eşitliğini sınamak, bu yolla madde formunun öğrencilerin performansı üzerinde etkili olup olmadığını belirlemektir.

Yöntem

Çalışma grubu

Araştırma ilköğretim 6. sınıfta öğrenim gören 420 öğrenci üzerinde yürütülmüştür.

Veri toplama araçları

Araştırmada veri toplama aracı olarak problem çözme becerisine ilişkin aynı davranışları ölçen 25 adet 5 seçenekli çoktan seçmeli sorudan oluşan sayısal köklü ve sözel köklü matematik testleri

kullanılmıştır. Bu testler alanında uzman matematik öğretmenleri tarafından geliştirilmiştir. Testler doğru cevaba 1 yanlış cevaba ise 0 puan vermek suretiyle puanlanmış ve bireylerin ham puanları doğru cevap sayıları toplanmak suretiyle elde edilmiştir. Testler uygulandıktan sonra test ve maddelere ait betimsel istatistikler hesaplanmıştır. Teste ait bir çok istatistik Tablo 1'de verilmiştir. Sözel köklü matematik testine ait madde ayırıcılık gücü indeksleri 0,39 ile 0,66 arasında, madde güçlük indeksleri ise 0,25 ile 0,78 arasında değişirken, sayısal köklü matematik testine ait madde ayırıcılık gücü indeksleri 0,29 ile 0,67 arasında, madde güçlük indeksleri ise 0,29 ile 0,72 arasında değişmektedir. Gerek Tablo 1'deki istatistikler gerekse madde istatistikleri testin geçerli ve güvenilir olduğunu ve çalışma kapsamında kullanılabileceğine ilişkin kanıt olarak kullanılabilir.

Uygulama

Araştırmada kullanılan problem çözme becerisine ilişkin aynı davranışları ölçen 25 adet çoktan seçmeli sorudan oluşan sayısal köklü ve sözel köklü matematik testleri 3 hafta arayla 420 6. sınıf öğrencisine tek oturumda verilmiş, 25 dakikalık süre tanınmıştır.

Verilerin Analizi

Sayısal köklü matematik test puanları, sözel köklü matematik testi puanlarına eşitlemek üzere, klasik eşitleme metotlarından lineer eşitleme ve tek grup düzeneği kullanılmıştır. Aynı zamanda eşitleme işleminin bir parçası olarak tek grup düzeneği için eşitleme hatası (SHE) kestirilmiştir. İki testin eşitliğini değerlendirmek için eşitlemenin standart hatasına dayalı güven aralıkları kullanılmıştır.

Bulgular

Sayısal köklü matematik test puanları, sözel köklü matematik testi puanlarına eşitlenmeden önce her iki teste ait betimsel istatistikler hesaplanarak Tablo 1'de sunulmuştur.

Tablo 1. Sayısal ve Sözel Köklü Matematik Testlerine Ait Betimsel İstatistikler

Testler	K	N	$\bar{X}$	S^2x	S_X	$KR-20$	Ortalama Güçlük	Çarpıklık Katsayıları	Basıklık Katsayıları
Sayısal Köklü	25	420	12.27	38,22	6,18	0,852	0,51	0,48	-0,80
Sözel Köklü	25	420	12.82	42,23	6,50	0,868	0,49	0.41	-0,91

Tablo 1'de verilen betimsel istatistiklere dayanarak, her iki testten elde edilen puan dağılımlarının birbirine çok benzediği söylenebilir. Sayısal köklü matematik test puanları, sözel köklü matematik testi puanlarına klasik eşitleme metotlarından lineer eşitleme ve tek grup düzeneği

kullanmak suretiyle eşitlenerek eşitleme işleminin bir parçası olarak tek grup düzeneği için eşitleme hatası (SHE) kestirilmiş ve Tablo 2'de sunulmuştur.

Tablo 2'de görülebileceği gibi eşitleme fonksiyonu ile özdeşlik fonksiyonu arasında bir miktar farklılıklar vardır. Bir diğer ifade ile Form 2'den elde edilen ham puanlarla eşitlenmiş puanlar arasında çeşitli puan düzeylerinde farklılıklar vardır. Bu farklılıklar -0,041 ile 1,159 puan arasında değişmekte ve düşük puanlardan yüksek puanlara doğru monotonik bir artış göstermektedir. Özellikle bu farklılıkların bazıları 2 SHE'den fazla olduğu için tanımlanan güven aralığının dışına düşmektedir.

Tablo 2. Sayısal ve sözel köklü matematik test puanlarına ait doğrusal eşitleme sonuçları

Form 2 Ham Puan	Form 1 Eşit puan	Fark	SHE	Oran
1	1,041	-0,041	0,732	-0,056
2	1,991	0,009	0,643	0,014
3	2,941	0,059	0,563	0,105
4	3,891	0,109	0,490	0,222
5	4,841	0,159	0,425	0,374
6	5,791	0,209	0,368	0,568
7	6,741	0,259	0,319	0,813
8	7,691	0,309	0,277	1,116
9	8,641	0,359	0,243	1,476
10	9,591	0,409	0,217	1,882
11	10,541	0,459	0,199	2,304
12	11,491	0,509	0,189	2,694
13	12,441	0,559	0,186	2,998
14	13,391	0,609	0,192	3,176
15	14,341	0,659	0,205	3,217
16	15,291	0,709	0,226	3,140
17	16,241	0,759	0,255	2,982
18	17,191	0,809	0,291	2,780
19	18,141	0,859	0,335	2,561
20	19,091	0,909	0,388	2,346
21	20,041	0,959	0,447	2,143
22	20,991	1,009	0,515	1,958
23	21,941	1,059	0,591	1,792
24	22,891	1,109	0,674	1,645
25	23,841	1,159	0,765	1,514

Sonuç ve Öneriler

Ülkemizde gerek öğretmen yapımı testlerin gerekse her yıl uygulanan seçme amaçlı bir çok sınavların her birinin farklı soruları içermesi fakat aynı özelliği ölçmesi ölçme aracının yanlı davranmaması için arzu edilen bir durumdur. Bir ölçme aracına ait soruların paralellerini oluşturmanın

bir yoluda aynı özelliği ölçen fakat farklı formlarda (sayısal, sözel, sembolik vb.) sorular hazırlamaktır. Bu derece önemli sınavlarda testin yansızlığına ilişkin kanıtların toplanması ve bu sınavların farklı versiyonlarından elde edilen puanların istatistiksel eşitliğinin sınanması gereklidir. Bu çalışma ile öncelikle bu ve bu türden sorunlara yönelik ve geçerliğe kanıt sağlamak amacıyla test eşitleme ve buna bağlı prosedürlerin nasıl kullanılabileceğini açıklamak hedeflenmiştir. Bu kapsamda, aynı davranışları ölçen sayısal köklü matematik testinden elde edilen puanlarla sözel köklü matematik testinden elde edilen puanların istatistiksel olarak birbirine eşit olup olmadığı sınanmış ve bu yolla bu testlerin farklı versiyonlarının bu sınavları alan bireylere adil davranıp davranmadığı bu versiyonların birbirinin yerine kullanılıp kullanılamayacağına (test yansızlığı) ilişkin kanıt sağlamak amaçlanmıştır. Sonuç olarak, elde edilen bulgular doğrultusunda, özdeşlik fonksiyonunun tanımlanan ∓ 2 SHE olarak tanımlanan güven aralığı içerisinde yer almaması sebebi ile , aynı davranışları ölçen sayısal köklü matematik testi ile sözel köklü matematik testinin birbirine eşdeğer olduğunu veya birbirinin yerine kullanılabileceğini söyleyebilmek güçtür. Bir diğer ifade ile Form 2'den elde edilen ham puanlarla eşitlenmiş puanlar arasında bazı puan düzeylerinde farklılıklar ∓ 2 SHE olarak tanımlanan güven aralığı içerisinde yer almamaktadır. Tablo 2 dikkatle incelenecek olursa, Form 2'den elde edilen ham puanlarla eşitlenmiş puanlar arasında çeşitli puan düzeylerinde farklılıklar, 2 SHE'den daha büyük olduğu gözlenebilir. Bu durumda öğrencilerin sayısal köklü matematik testi ile sözel köklü matematik testini almaları bu öğrencilerin performanslarında farklılık yaratabileceği anlamını taşımaktadır. Bu durumda birbirine paralel olduğu varsayılan aynı davranışları ölçen sayısal köklü matematik testi ile sözel köklü matematik testinin birbirine paralel olduğunu ve yansız olduğunu bu çalışmanın sonuçları doğrultusunda söylemek mümkün gözükmemektedir.

Kaynaklar

Angoff, W. H. (1971). Scale, norms and equivalent scores. In R. L. Thorndike (Eds.) *Educational Measurement (2nd. Ed.)* Washington D.C; American Council of Education.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P.W. Holland & D. B. Rubin (Eds.). *Test Equating*. New York: Academic Press.

Barnard, J. J. (1996). "In search for equity in educational measurement: Traditional versus modern equating methods." *Paper presented at ASEESA's national conference at the HRSC Conference Centre, Pretoria, South Africa.*

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. San Diego, C.A: Harcourt Brace Jovanovich College Publisher.

Davier, A. A., Holland, P. W. & Thayer, D.T. (2004). The chain and post stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, 41(1), 15-32.

Dorans, N. J., & Lawrence, I., M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement In Education*. 3(3), 245-254.

Harris, D. J. (2003). Equating the multistate bar examination. *The Bar Examiner*, 72(3), 12-16.

Holland, P. W., Sinharay, S., Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratificationequating methods for the NEAT design. *Journal of Educational Measurement*, 45(1), 17-43.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, M.A: Addison-Wesley.

Tanguma, J. (2000). "Equating test scores using the linear method: A primer." *Paper presented at the annual meeting of the Southwest Educational Research Association. Dallas, TX.*

Wooldbeck, T. (1998). "Basic concept in modern methods of test equating." *Paper presented at the annual meeting of the Southwest Psychological Association New Orleans, L. A.*