

## ÖLÇÜT DAYANAKLI TESTLERİN MADDE ANALİZİNDE UYUŞMA İNDEKSLERİNİN KULLANIMI

Vahit BADEMCI (\*)

Eğitimde başarının ölçülmesinde ve değerlendirilmesinde, ölçüt dayanaklı (criterion referenced) ölçme ve değerlendirme yöntemi belirgin bir şekilde kullanılmaktadır. Ölçüt dayanaklı ölçme, bireyin testteki başarısını belirlenen bir ölçüte göre açıklar. Bu ölçme yönteminde grubun özellikleri değil, ölçülmek istenen özelliklerin tümü ön plana çıkmaktadır. Eğitimde kullanılan ölçme araçları içinde önemli bir yer tutan çoktan seçmeli testler, ölçüt dayanaklı değerlendirmelerde de sıkça kullanılmaktadır. Ölçüt dayanaklı testlerde öğrenme birimiyle ilgili davranışların bütünü esas alındığından, ölçmeyi konu alan öğrenme birimini en iyi yansıtacak olan test maddelerinin seçimi, bu değerlendirme yönteminin temelini oluşturmaktadır. Bu testlerde madde güçlükleri gruba göre değil, ölçülmesi amaçlanan davranışların güçlüğüne göre seçilir.

Ölçüt dayanaklı testler ölçme alanında ilk kez Glaser (1963) ve sonra Popham ve Husck (1969) tarafından hedef dayanaklı öğretim programlarına yönelik öğrenmelerin ölçüldüğü testlerde kullanılmıştır (Keeves, 1988: 277-282). Daha sonra ölçüt dayanaklı testlerin kullanım alanları giderek artmıştır. Bu testler hedef dayanaklı öğretim programları içinde biriyi izleme öğrenme güçlüklerini ve eksiklerini teşhis etme ve öğretim programlarını değerlendirme amacıyla kullanılmaktadır. Ölçüt dayanaklı bir test, iyi tanımlanmış bir öğrenme birimi için ulaşılan öğrenme düzeyinin, önceden tanımlanan bir öğrenme düzeyine göre yeterli olup olmadığına karar vermede kullanılır. Burda sözü edilen, önceden belirlenen öğrenme düzeyi, tam öğrenme modelinde yer alan ve tam öğretmenin gerçekleştiğine karar vermede kullanılan kesme puanı olarak tanımlanabilir (Hambleton ve Swaminathan, 1985: 257-262). Tam öğrenme modelinde birey iyi tanımlanmış davranışların minimum %80'ini kazanmak durumundadır (Thorndike, 1982: 1-3; 181-183). Öğrencinin ölçüt dayanaklı bir testten aldığı puan, öğrencinin istenen tam öğrenme düzeyine ulaşmış, ulaşmadığını gösterir (Ornstein ve Hunkins, 1988: 271; Cronbach, 1984: 87-89).

(\*) Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Araştırma Görevlisi

İyi bir ölçüt dayanaklı test maddesi, tam öğrenme düzeyine ulaşanlar tarafından doğru, ulaşamayanlar tarafından yanlış cevaplandırılmalıdır. İşte bu aşamada madde seçimi önem kazanmaktadır. Ölçüt dayanaklı testlere madde seçimi amacıyla geliştirilen değişik yöntemler ve istatistikler vardır (Thorndike, 1982: 93-96; Hambleton ve Swaminathan, 1985: 27-262; Crocker ve Algina, 1986: 329-335; Keeves, 1988: 382-383).

Ölçüt dayanaklı testler geliştirilirken maddeler hakkında cevaplanması istenilen bazı sorulara, değişik madde istatistikleriyle yanıtlar bulunulmasına çalışılmıştır. Kullanılan istatistiklerle, madde güçlüğü (item difficulty), öğretimin duyarlılığı (instructional sensitivity) ve uyuşma indeksleri (indices of agreement) bulunarak, maddeler hakkında değişik yorumlamalara gidilmiştir (Crocker ve Algina, 1986: 329-335).

Ölçüt dayanaklı testler (ÖDT)'in geliştirilmesinde öğretimin duyarlılığı ve madde güçlüğü, üzerinde en çok çalışılan ve bilinen madde istatistikleridir (Roid ve Haladyna, 1982: 215-224). Ancak bu istatistikler aynı özellikleri ölçmek için hazırlanan maddelerde, bir takım sorulara cevap vermede yetersiz kalmaktadır. Son zamanlarda, iki madde aynı şeyi ölçüyor mu?, maddelerin her ikisini de doğru ya da yanlış cevaplamış cevaplayıcıların oranı nedir? vb. sorulara, uyuşma indeksleri adı altında, değişik istatistiklerle cevaplar bulunmaya çalışılmıştır (Crocker ve Algina, 1986: 331-335).

ÖDT'in geliştirilmesinde ve maddelerin seçiminde, artarak önem kazanan uyuşma indeksleri burada açıklanmaya çalışılacaktır.

Ölçüt dayanaklı testler geliştirilirken test geliştiriciler aynı davranışı ölçen maddeler için, bir grup cevaplayıcının verdiği cevapların benzer olup olmadığını bilmek isterler. Bir çift maddeye verilen cevaplar birbirine benzerse, maddeler birbirinin yerini tutabilir demektir. Bu bilgi, test oluşturmada ve geliştirmede büyük kolaylıklar sağlayacaktır. Zira daha sonra bu tür maddelerden oluşmuş setlerden, maddeler random yolla seçilerek, yeni testler oluşturulabilecek ve geliştirilebilecektir. Test geliştiriciler, aynı özelliği ölçmek için yazılmış olan madde çiftlerinin benzer olup olmadığını, o madde çiftine cevaplayıcıların verdikleri cevaplardan elde edilen verilere göre bulabilir. Bunun için cevaplayıcılardan elde edilen veriler, Harris ve Pearlman (1977) tarafında geliştirilen Tablo 1a'daki gibi dört gözeli bir tabloda gösterilir ve ki-kare istatistiğiyle bazı sorulara cevap aranır (Crocker ve Algina, 1986: 332). Bu yöntemle, sorulan bazı sorulara bulunan cevapları bir örnekle gösterebiliriz.

Örnek: Bir test 100 kişilik bir öğrenci grubuna verilmiştir. Bu testte bulunan iki maddeye (Madde 17 ve Madde 18) verilen cevaplardan elde edilen veriler Tablo 1b'de gösterilmiştir.

Tablo 1  
Maddelerin Bir Çiftine Verilen Yanıtlar  
Madde 17

		+	-	
Madde 18	+	a	b	(a)
	-	c	d	

		+	-	
Madde 18	+	40	10	(b)
	-	10	40	

100 cevaplayıcı için frekanslar

**Cevaplanması istenilen soru:** İki madde aynı şeyi mi ölçüyor? Bu sorunun cevabını bulabilmek için ki-kare istatistiğinden faydalanılır.

$$\chi^2 = \frac{n (ad - bc)^2}{(a + b) (c + d) (b + d) (a + c)}$$

n : Örneklemdaki kişilerin sayısı

a, b, c, d : Cevaplayıcıların verdikleri yanıtlar doğrultusunda frekansların yazıldığı hücreler

Tablo 1b'de ki veriler formüldeki yerlerine konulur.

$$\chi^2 = \frac{100 (1600 - 100)^2}{(50) (50) (50) (50)} = 36$$

Hesaplanmış ki-kare istatistiği istendik alfa düzeyinde (örnekte  $\alpha = .05$  alınmıştır) ve 1 serbestlik derecesinde, tablodan okunan ki-kare değeriyle karşılaştırılır. Hesaplanan ki-kare değeri, tablodan okunan ki-kare değerinden büyüktür ve manidar kabul edilir.

**Yorumlanması:** Eğer test istatistiğinin hesaplanmış değeri anlamlıysa, iki maddenin yanıtları benzerdir ve iki maddenin aynı beceriyi ölçtüğü kabul edilir. Eğer aynı özelliği ölçmek için yazılmış iki maddenin yanıtları arasında anlamlılık yoksa, iki madde birbirine benzemiyor ve aynı şeyi ölçmüyor denebilir. Bu durumda maddelerden birisinin (veya her ikisinin) içeriği veya teknik kalitesi veya yazıldığı özellik açısından yetersizlikleri vardır, yorumu yapılabilir. Test geliştirici bu durumda maddelerin birisini ya da her ikisini tekrar kontrol etmeli ve düzeltmelidir.

**Cevaplanması istenilen soru:** Maddelerin her ikisini de doğru ya da yanlış cevaplamış cevaplayıcıların oranı nedir?

Burada cevaplayıcıların maddelere verdikleri cevaplardaki benzerlik, şans yoluyla verebilecekleri benzerlikten daha büyüktür. Test geliştiriciler benzerliğin derecesini bulan bir istatistiği hesaplamak isterler. Harris ve Perlman (1977) tarafından ileri sürülen aşağıdaki istatistikle, uyuşmanın oranı kolaylıkla ve basitçe hesaplanmaktadır. İki maddeye de tutarlı cevaplar vermiş cevaplayıcıların oranı, bu formülle hesaplanır.

$$P = \frac{a+d}{n}$$

- P : İki maddeye de tutarlı cevaplar veren cevaplayıcıların oranı.  
a, d : Her iki maddeyi de birlikte doğru ya da birlikte yanlış cevaplayan cevaplayıcıların frekanslarının yazıldığı hücreler.  
n : Örneklemdaki kişilerin sayısı.

Tablo 1 b'de ki veriler formüldeki yerine yerleştirilir.

$$P = \frac{40 + 40}{100} = 0.80$$

**Yorumlanması:** Cevaplayıcıların %80'i iki madde üzerinde tutarlı bir performans göstermişlerdir. Bu sonuç aynı zamanda maddelerin benzer olduğu şeklinde de yorumlanabilir. Zira elde edilen bu oran uyuşma oranıdır. Bu oran düştükçe yani sifıra doğru yaklaştıkça cevaplayıcıların birlikte cevaplandıkları iki madde üzerinde tutarlı bir performans göstermedikleri ortaya çıkacaktır. Bu durumda cevaplayıcıların, cevaplarını şansa cevaplandırmış olmaları muhtemeldir ya da iki madde de benzerlik yoktur, yorumlamaları getirilebilir. Bunları yaratan nedenler kontrol edilmeli ve düzeltilmelidir. İki maddeye verilen yanıtların uyuşma derecelerinin hesaplanması phi katsayısı (Roid ve Haladyna, 1982: 218-220) ve kappa katsayısı, Yule'nin Q'su ile de yapılmaktadır (Crocker ve Algina, 1986:334).

**Cevaplanması istenilen soru:** Cevaplayıcıların iki maddeden biri üzerindeki performansları, diğeri üzerindeki daha mı iyidir?

Test geliştiriciler cevaplayıcılara sorulan iki maddenin eşit güçlükte olup olmadığını ve öğretim programının uygun olup olmadığını bilmek isterler. İki maddenin içeriğinin eşit şekilde öğrenilip öğrenilmediği, Harris ve Perlman'ın ileri sürdüğü ki-kare istatistiğinin bir formuyla hesaplanabilir (Crocker ve Algina, 1986: 334).

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

Tablo 1 b'de ki veriler formüldeki yerlerine konulur.

$$\chi^2 = \frac{(10 - 10| - 1)^2}{10 + 10} = 0.05$$

Hesaplanmış ki-kare değeri istendik alfa (örnekte  $\alpha = .05$  alınmıştır) düzeyinde ve 1 serbestlik derecesinde tablodan okunan ki-kare değeriyle karşılaştırılır. Hesaplanan ki-kare değeri, tablodan okunan ki-kare değerinden küçüktür ve manidar değildir.

**Yorumlanması:** Eğer test istatistiğinin hesaplanmış değeri manidar değilse, madde güçlükleri arasında anlamlı bir farklılık yok demektir. Yani cevaplayıcıların her iki madde üzerindeki performanslarında bir farklılık yok denilebilir. Ya da cevaplayıcılar bir maddeyle ölçülen içeriği diğer maddeyle ölçülen içerikten daha fazla öğrenmemişlerdir veya her iki maddenin içeriğini de eşit şekilde öğrenmişlerdir, diye yorumlar yapılabilir. Eğer hesaplanmış ki-kare değeri manidarsa, madde güçlükleri arasında bir farklılık var demektir. Ya da iki maddenin içeriği eşit şekilde öğrenilmemiştir denilebilir. Bu durumda şans faktörünün etkisi çok büyüktür, gibi bir yoruma da gidilebilir. Böyle bir sonuçla karşılaşıldığında gerekli düzeltme çalışmaları yapılmalıdır.

Madde uyuşma indeksleri, maddelerin belirli çiftleri üzerinde cevaplayıcıların nasıl davrandıklarına ilişkin bazı sorulara cevap arandığında sıklıkla kullanılır. Bir test üzerinde madde uyuşma indeksleri adı altında toplanan istatistiklerden bir veya bir kaçının birden uygulanması test üzerinde çok önemli ve anlamlı bilgiler verecektir. Az maddeli alt testler veya kapsam geçerliği çalışmaları için veya test puanlarının yorumlanmasında, uyuşma indeksleri büyük kolaylıklar sağlar.

#### KAYNAKÇA

- CROCKER, L. ve J. ALGINA 1986, *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston.
- CRONBACH, L.J. 1984 *Essentials of Psychological Testing*. (Fourth Edition). New York: Harper and Row.
- HAMBLETON, R.K. ve H. SWAMINATHAN 1985, *Item Response Theory*. Boston: Kluwer-Nijhoff.
- KEEVES, J.P. 1988 *Educational Research, Methodology, and Measurement. An International Handbook*. Oxford: Pergamon.
- ORNSTEIN, A.L. ve F.P. HUNKINS 1988, *Curriculum: Foundations, Principles, and Issues*. New Jersey: Prentice Hall.
- ROID, G.H. ve T.M. HALADYNA 1982 *A Technology for Test-Item Writing*. New York: Academic Press
- SPIEGEL, M.R. 1988 *Theory and Problems of Statistics*. (Second Edition). New York: Mc Graw Hill
- THORNDIKE, R.L. 1982, *Applied Psychometrics*. Boston: Houghton Mifflin.