

MADDE YANLILIĞI: TEST MADDELERİ YANITLAYICILARA PERFORMANSLARINI GÖSTERMELERİ İÇİN EŞİT ŞANS SAĞLIYOR MU?

H. Eren Suna

hesuna@baskent.edu.tr

Başkent Üniversitesi, Eğitim Bilimleri Bölümü

Günümüzde, eğitim süreci boyunca öğrenciye ve eğitim sisteminin farklı öğelerine yönelik karar verilme durumlarında sıklıkla testler kullanılmaktadır. Testlerin eğitim sisteminde genel kullanım amaçları, öğrencilerin ilgi ve yeteneklerinin belirlenmesi, öğretim programının değerlendirilmesi, öğretimin etkililiğinin belirlenmesi, öğrenme eksikliklerinin belirlenmesi ve öğrenci başarısının değerlendirilmesi olmak üzere beş temel grupta toplanabilir (Baykul, 2000). Testlerin eğitim sürecindeki kullanım amaçlarından görüleceği üzere, farklı amaçlar için geliştirilmiş testlerden elde edilen sonuçlar, eğitim sürecinde alınacak birçok karara dayanak sağlamaktadır. Dolayısıyla, ancak testlerden elde edilen sonuçların güvenilir ve geçerli olması durumunda, bu sonuçlar göz önüne alınarak verilen kararların amaca hizmet edebilmesi mümkündür, aksi takdirde verilen kararların isabetliliği şüphe çekici olacaktır.

Testlerden elde edilen ölçme sonuçlarının güvenilir ve geçerli olması, test geliştirme süreci basamaklarının uygun olarak izlenmesi ile sağlanabilir. Test geliştirme süreci, ölçülmesi amaçlanan özelliğin belirlenmesi amacıyla, sistematik bir yolla, ölçme aracının uygulanabilir hale getirilmesini ifade eder. Test geliştirme sürecinin en önemli basamaklarından birisi de maddelerin yazılmasıdır. Maddelerin yazılabilmesi için öncelikle ölçülmesi amaçlanan özelliğin hangi madde türü ile en uygun şekilde ölçülebileceğine karar verilmelidir. Bu işlem tamamlandıktan sonra, uygun görülen madde türü veya türleri ile maddelerin yazılmasına geçilebilir.

Maddeler, testi oluşturan birimlerdir ve maddelerin kalitesi, içeriğinde bulunduğu testin isabetli kararlara dayanak olacak sonuçlar sağlaması açısından büyük öneme sahiptir. Maddeler ile ölçülmek istenen özelliğin diğer özellikler ile karıştırılmadan ölçülmesi, o maddeden alınan puanların geçerli olmasını sağlar. Bu durumun sağlanması için ölçme yapılacak grup, ölçülecek özellik anlamında tanımlı, grup içindeki olası bilişsel farklılıklar dikkate alınmalıdır. Madde ile ölçülmek istenen özellik dışında, söz konusu özellik ile ilgisiz olan değişkenlerin maddelere dâhil edilmesi ve yanıtlayıcılar arasında bulunan bazı alt gruplara avantaj veya dezavantaj sağlayabilecek durumların maddelerde yer alması, maddelerden elde edilecek puanların geçerliliğini azaltır. Bu şekilde bozucu etkilere sahip maddelerden oluşan bir testin sonuçları ise tartışmaya açıktır.

Testlerin, ölçmeyi amaçladıkları özellikleri, geliştirilme amaçları doğrultusunda ve mümkün olduğunca az hata ile ölçmesi amaçlanır. Bununla beraber, eğitim ve psikolojide kullanılan hiçbir test, ölçmeyi amaçladığı değişkeni hatasız olarak ölçemez. Testlerden ve testleri oluşturan maddelerden elde edilen puanların, testin ölçmeyi amaçladığı yapı dışındaki farklı kaynaklardan etkilenmesi kaçınılmazdır. Bu durum eğitimsel ve psikolojik ölçmelerin bir özelliği olmakla beraber, hatanın yanıtlayıcıların oluş-

turduğu grupları farklı şekilde etkilemesi, diğer bir ifade ile sistematik bir hale gelmesi, ölçme süreçlerinde yanlılık sorununu ortaya çıkarır.

Yanlılık, ölçme süreçlerinde tekil olarak maddeler ve testin bütünü boyutlarında incelenmektedir. Sözlük anlamı ile yanlılık, "Gerçeğin ifadesinden veya göstergelerinden ayrılma" olarak ifade edilir (TDK, 2005). Ölçme alanyazında ise yanlılığın farklı tanımları olmakla beraber, genel ve sık kabul edilen bir tanımı "Bir gruba avantaj veya dezavantaj sağlanmasına yol açan sistematik hata" şeklindedir (Reynolds, Livingston ve Wilson, 2006; Shephard, 1981). Buna göre bir madde, ölçmeyi amaçladığı değişken ile ilgisiz herhangi bir özelliği nedeniyle yanıtlayıcılar arasında bulunan bir gruba, diğer gruplara göre avantaj veya dezavantaj sağlıyorsa, söz konusu maddenin yanlı olduğu ifade edilir.

Test maddelerinin farklı gruplar için (madde yanlılığının belirlenmesi sürecinde karşılaştırılan gruplar referans grubu ve odak grubu olarak isimlendirilir) karşılaştırılabilir sonuçlar sağlaması, söz konusu gruplara maddeleri doğru cevaplamak için eşit şans verilmesi şeklinde yorumlanabilir. Özellikle "High Stakes Tests" adı verilen ve sonuçlarına dayanılarak yanıtlayıcılar hakkında oldukça önemli kararlar verildiği testlerde yer alan maddelerin yanlılık açısından incelemesi oldukça önemli görülmektedir. Amerika Birleşmiş Milletleri başta olmak üzere farklı ülkelerde yüksek eğitim kurumlarına girişte sonucu önem teşkil eden Scholastic Aptitude Test'te (SAT) yer alan kimi sorular özellikle siyahî ve beyaz öğrencilere eşit şans tanımadığı gerekçesi ile eleştirilmiş, söz konusu soruların özellikle daha üst sosyoekonomik düzeyde bulunan beyaz öğrencilere avantaj sağladığı, eşit yetenek düzeyinde bulunan beyaz öğrencilerin incelenen maddelerde manidar olarak daha yüksek başarı gösterdiği bulunmuştur (Holland ve Wainer, 1993).

Madde yanlılığı çalışmaları, günümüzde, özellikle sonuçlarına dayanarak yanıtlayıcılar hakkında önemli kararlar alındığı testler üzerinde yoğunlaşmakla beraber; AERA'nın (1985) yayınladığı "Eğitimde ve Psikolojide Ölçme Standartları"nda, yansız maddelerin tüm eğitimsel ve psikolojik testler için bir ölçüt olduğu vurgulanmıştır. Buna göre, ölçme aracında bulunan maddeler, yanıtlayıcı grubunda bulunan tüm bireylere, bilgi veya becerilerini gösterecek şekilde fırsat sağlamalı ve herhangi bir grubun maddeleri doğru cevaplama olasılığı, diğer gruplara göre manidar farklılık göstermemelidir (Roever, 2005).

Madde yanlılığının belirlenmesi için alanyazında Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı'na (MTK) dayalı çeşitli yöntemler bulunmaktadır. Bu yöntemler genel olarak madde işlev farklılaşmasını (MİF) belirleme yöntemleri olarak adlandırılmaktadır. MİF belirleme yöntem-

leri yapısal olarak farklılaşsa da tüm yöntemler temelde, dezavantajlı durumda olabileceği düşünülen odak grup ile incelenen maddenin avantaj sağladığı düşünülen referans grubunda yer alan yanıtlayıcıların yeteneklerinin eşitlenmesi ve eşit yetenekli yanıtlayıcıların incelenen maddedeki performanslarının karşılaştırılmasına dayanmaktadır. Bu karşılaştırma sonucunda, kullanılan yöntemin gereklerine göre hesaplanan katsayının pozitif-negatif olma durumuna, büyüklüğüne ve manidarlık durumuna göre incelenen maddenin işlev farklılığı gösterip göstermediği, gösteriyorsa hangi düzeyde ve hangi grup lehine farklılık gösterdiği belirlenir. Söz konusu yöntemler kullanılarak farklı gruplarda işlev farklılığı gösterdiği belirlenen maddeler için uzman görüşü alınarak maddenin yanlı ölçme yapıp yapmadığına ilişkin karar verilmektedir. Burada dikkat edilmesi gereken nokta, MİF gösterdiği belirlenen maddelerin yanlı ölçme yapıp yapmadığının belirlenmesinde uzman görüşüne ihtiyaç duyulmasıdır. Diğer bir ifade ile bir maddenin MİF göstermesi yanlılık için önemli bir göstergedir fakat maddenin MİF göstermesi, kendi başına o maddenin yanlı ölçme yaptığının kanıtı olarak görülemez. Uzman görüşünde maddenin neden MİF gösterdiği, karşılaştırılan referans ve odak gruplarındaki eşit yetenekteki yanıtlayıcıların maddeyi doğru cevaplama olasılıklarının neden farklılaştığı incelenir; bu durumun maddenin yapısından kaynaklı olduğuna karar verilirse o maddenin yanlı ölçme yaptığını kanaat getirilir ve yanlı ölçme yaptığı belirlenen maddeler testten çıkarılır.

Test geliştirme, izlenmesi hassasiyet ve özen gerektiren aşamalardan oluşan bir süreçtir. Bu süreci sistematik bir şekilde yapılandırmak için test geliştiricilerine önemli bir sorumluluk düşmektedir. Türkiye'nin mevcut durumu göz önüne alındığında, özellikle yüksek öğretim kurumlarına giriş (YGS-LYS) ve birçok diğer merkezi sınavın bireyin sonraki yaşamı için belirleyici olduğu görülmek-

tedir. Merkezi sınavların tüm Türkiye'de uygulanması dolayısıyla oldukça heterojen bir yanıtlayıcı grubu ile uygulama yapıldığının göz önüne alınması ve yanıtlayıcıların oluşturduğu belirli alt gruplara göre maddelerin yanlı ölçme yapıp yapmadığının sınanması (örn., maddelerin kız ve erkek öğrenci gruplarında hangi düzeyde karşılaştırılabilir sonuçlar sağladığının belirlenmesi) büyük önem teşkil etmektedir. Bu bağlamda, merkezi sınavların büyük kısmını çatısı altında toplamış olan ÖSYM'nin geliştirdiği testlerin yanlılık açısından incelenmesi yeni geliştirilecek testlerin daha isabetli kararlara dayanak olmasını sağlayacak, dahası, son zamanlardaki uygulamaları farklı açılardan eleştirilen ÖSYM tarafından geliştirilen testlerin hangi düzeyde karşılaştırılabilir sonuçlar sağladığını gösterecektir.

Kaynaklar

AERA, (1998). *Eğitimde ve psikolojide ölçme standartları*. (S. Hovardaoğlu ve N. Sezgin, Çev.). Ankara: Türk Psikologlar Derneği ve ÖSYM yayını. (Orijinal çalışma basım tarihi 1985).

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM.

Holland, P. W. ve Wainer, H. (1993). *Differential item functioning*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Reynolds, C. R., Livingston, R. B. ve Wilson, W. (2006). *The problem of bias in educational assessment in measurement and assessment in education*. Boston: Pearson Education.

Roever, C. (2005). "That's not fair!" *Fairness, Bias and differential item functioning in language testing*. 30 Ekim 2012, <http://www2.hawaii.edu/~roever/brownbag.pdf>.

Shepard, L. A. (1981). Identifying bias in test items. B. F. Green, (Ed.), *Issues in testing: Coaching, disclosure and ethnic bias* içinde. San Francisco, CA: Jossey Bass.

