

MATEGRAFI

Hüseyin Yıldırım* / yildirim.huseyin@ibu.edu.tr



Mategrafi henüz icat edilemedi. Bunun için önce beynimizin bazı sırlarının çözülmesi gerekiyor. Matematik öğrendiğimizde, matematiksel kavramları ve bunlar arasındaki bağlantıları anladığımızda, beynimizde birtakım fiziksel değişimler de oluyordur mutlaka. Beynimizin sırları bu değişimlerin neler olduğunun anlaşılmasına yetecek kadar çözüldüğünde, bu değişimleri tespit edebilecek bir cihaz da elbet icat edilecektir. Mategrafi, işte bu cihazın adı olacak ve icat edildiğinde insanların matematik yeterlik düzeyleri doğrudan ölçülebilecek. Ne var ki o zamana kadar, dolaylı ölçmeyle idare etmek zorundayız.

Öğretmenlerin yaptığı sözlülerden tutun da, üniversite giriş sınavlarındaki testlere kadar, hepsi birer dolaylı ölçme girişimidir. Öğrencilerin, örneğin matematik yeterlik düzeylerini belirlemek üzere, beyinlerine (henüz!) doğrudan bakılmadığı için, bir sınavda yaptıklarına bakılır. Ve bunlara dayanarak, öğrencilerin matematik yeterlik düzeyleri belirlenmeye, kestirilmeye çalışılır. Bu çok karmaşık bir iştir; daha doğrusu karmaşık bir süreçtir. Öğrencilere ne sorulduğu, nasıl sorulduğu ve öğrencilerin cevaplarıyla ölçülmek istenen şey (matematik yeterliği vb.) arasındaki ilişkinin nasıl betimlendiği gibi birçok etmen girer işin içine. Bu dolaylı ölçme süreci sonunda kestirilen yeterlik düzeyiyle, kişinin gerçek (asla bilemediğimiz) yeterlik düzeyi arasındaki fark (yani, ölçme hatası), bu sürecin ne kadar iyi yürütüldüğüne bağlıdır.

Bu karmaşık sürecin yürütülmesini kolaylaştırmak üzere, dolaylı ölçmede izlenmesi gereken yol standartlarla belirlenmiştir. Amerikan Eğitim Araştırmaları Birliği, Amerikan Psikoloji Birliği ve Eğitimde Ölçme Ulusal Konseyi tarafından ortaklaşa bir çalışmayla hazırlanan Eğitimde ve Psikolojide Ölçme Standartları, 1985'te yayımlanmış ve 1999'da güncellenmiştir. Bu standartlar birçok ül-

kede yaygın bir şekilde kullanılır ve özellikle sonuçlarına göre önemli kararlar alınan sınavlarda, bu standartlara uyulduğunu gösteren çok sayıda doküman yayımlanır. Fakat Türkiye'de yayımlanmaz. Bu yazının amacı, ülkemizdeki sınavlarda kullanılan ölçme modelini de özetleyerek, bu eksikliğe dikkat çekmektir.

Sınav sorularına verdikleri cevaplara dayanarak, öğrencilerin matematik yeterlik düzeylerini kestirmek için bir ölçme modeli gerekir. Bu model, bir öğrencinin sınavdaki matematik sorularına verdiği cevaplarla, öğrencinin yeterlik düzeyi arasındaki ilişkiyi betimleyen bir fonksiyondur. Mümkün olduğunca az hatalı bir kestirimde bulunmak için, kullanılan ölçme modelinin karakteristiği, sınırlılıkları, varsayımları bilinmeli ve ölçme araçları (yani sınavlar veya eş anlamlı olarak, testler) bunlara uygun olarak geliştirilmeli ve kullanılmalıdır. Bu yazıda, Klasik Test Kuramı altında yer alan Gerçek Puan Modeli üzerinde durulmaktadır. Türkiye'de yapılan SBS, YGS, KPSS vb. sınavlarda kullanılan ölçme modeli de budur.

Yukarıda sözünü ettiğimiz ölçme standartları, kullanılan ölçme modellerinin gerekleridir aslında. Fakat bu standartlar, kendilerini gerekli kılan matematiksel yapıdan fazlaca söz edilmeden verilmektedir. Böylelikle, test geliştirmek durumunda olan geniş bir kitleye, kolayca takip edebilecekleri bir yol haritası sağlandığı düşünülüyor olsa gerek. Ne var ki, ölçme sürecinin gerisindeki matematik, genel hatlarıyla da olsa anlaşılmasa, bu standartlar yerine getirilmesi gerekli zorunluluklar olarak değil de, "şöyle yapsanız iyi olur" benzeri öneriler olarak da algılanabilir. Hâlbuki, bu standartları karşılamayan bir sınavın ne ölçtüğü belirlenemez. Dahası, sınavın ölçmeye niyetlendiği şeyi ne kadar iyi (hatasız) ölçtüğü de belirlenemez. Ülkemizde, sonuçlarına göre önemli kararlar alınan sınavların böyle sınavlar olmadığını, ancak yeterince kanıt yayımlanırsa görebiliriz.

* A. İ. B. Ü., Eğitim Fakültesi öğretim üyesi.

Tarihten Birkaç Sayfa

Ölçmeyle ilgili tartışmalar, matematik yeterliği gibi soyut zihinsel oluşumları ölçme girişiminden yüzlerce yıl önce, astronomide başlamıştı. Bir gezegenin belirli bir zamandaki konumu ölçülüyor ama farklı sonuçlar elde ediliyordu. Bu sonuçlardan hangisi kullanılmalıydı?

Danimarkalı bir asilzade olan Tycho Brahe 16. yüzyılın sonlarında, elde edilen tüm ölçümlerin aritmetik ortalamasını kullanmayı önerdi. Bu önerisiyle, cevaplanması yaklaşık 300 yıl sürecek bir tartışma başlattığını muhtemelen kendisi de bilmiyordu. İlk eleştiri çağdaşlarından geldi. Hatasız bir ölçümle hatalı bir ölçümün ortalamasının, hatasız olandan daha kötü bir ölçüm olacağı açıkken, ölçümlerin aritmetik ortalamasını kullanmanın iyi bir fikir olduğu nasıl savunulurdu?

Bunu savunmak gerçekten de kolay olmadı. Tatmin edici bir savunma, ölçme hatasının modellenmesiyle mümkün oldu. Gauss tarafından inşa edilen hata eğrisi (normal dağılım veya çan eğrisi diye de bilinen), tekrarlı ölçmelerde yapılabilecek ölçme hatalarını hesaplamanın yolunu açtı. Bu arada, hata eğrisini kuran Gauss olsa da, ona yol gösteren çalışmaların Laplace'dan geldiğini belirtmeden geçmeyelim. Gauss hata eğrisini yayımladıktan bir yıl sonra, bu eğriye dayanarak, bugün *merkezî limit teoremi* diye bildiğimiz teoremi ispatlayan da yine Laplace oldu.

Bu çalışmalar sayesinde, tesadüfi ölçme hatalarının normal dağılıma eğiliminde olduğunu ve bu dağılımın beklentisinin sıfır olduğunu biliyoruz artık. Diğer bir deyişle, aynı şeyin tekrarlı ölçümlerindeki ölçme hatalarının simetrik olduğunu (belirli bir büyüklükteki pozitif hata kadar, yaklaşık aynı büyüklükte ve sayıda negatif hata olduğunu), küçük hataların daha sık görüldüğünü, hata büyüdükçe görülmeye ihtimalinin azaldığını biliyoruz. Ayrıca, merkezî limit teoremi sayesinde, bunların ölçüm hatalarının ortalaması için de geçerli olduğunu biliyoruz.

Brahe de bunları bilseydi, ölçme sonuçlarının aritmetik ortalamasını alma önerisini şöyle savunabilirdi: Gezegenin belirli bir zamandaki konumunu i 'nci ölçme girişiminden elde ettiğimiz sonuç X_i , bu girişimdeki ölçme hatası H_i olsun. Gezegenin bizim bilmediğimiz gerçek konumu da G olsun. Bu durumda

$$X_i = G + H_i$$

Yani, tekrarlı ölçümler sadece tesadüfi bir ölçme

hatasından ötürü değişiklik göstermektedir. Bu şekilde n ölçüm yapılmışsa, bunların ortalaması

$$\sum \frac{X_i}{n} = \sum \frac{G}{n} + \sum \frac{H_i}{n} = G + \sum \frac{H_i}{n}$$

Ölçme hataları ortalaması, beklentinin sıfır olduğu bir normal dağılıma sahip olduğundan,

$$\lim_{n \rightarrow \infty} \left(\sum \frac{H_i}{n} \right) = 0$$

Dolayısıyla,

$$\lim_{n \rightarrow \infty} \left(\sum \frac{X_i}{n} \right) = G$$

Diğer bir deyişle, ölçüm sayısı arttıkça bunların ortalaması, gezegenin gerçek konumuna yakınsamaktadır.

Dikkat edilirse, ancak sonsuz sayıda ölçme hatasının ortalaması sıfır olmaktadır. Dolayısıyla, pratikte

$$\sum \frac{H_i}{n}$$

sıfır olmaz. Üstüne üstlük, bu ortalama ölçme hatasının kesin değeri de hesaplanamaz. Ama *merkezî limit teoremi* sayesinde, bu değer belirlenir bir ihtimalle hangi aralıkta olduğu hesaplanabilir. Örneğin, ortalama hatanın diyelim ki, %95 ihtimalle (-2, 2) aralığında olduğu gibi bir hesaplama yapılabilir. Böyle bir durumda, gerçek değer de yine aynı ihtimalle, ölçme sonuçları ortalamasının 2 komşuluğunda olacağı görülür.

Merkezî limit teoremi (veya o zamanki adıyla *hataların frekansı kanunu*) bilimsel çalışmalara büyük bir soluk aldırdı. Fiziki olgular gözlenirken yapılmış olması muhtemel ölçme hatası hesaplanabiliyordu artık. Örneğin, İngiliz bilim adamı Lord Rayleigh, atmosferden oksijen ve karbondioksit gibi gazları ayırıştırarak elde edilen ve nitrojen olduğu düşünülen gazla, laboratuvarında üretilen nitrojen gazı arasında gözlediği kütle farkını, ölçme hatası olarak açıklamanın zor olduğunu bu sayede görmüştü. Böylelikle, atmosferden elde edilen gazın içinde henüz bilinmeyen bir element olduğundan şüphelenmiş, bu yöndeki çalışmaları sonunda argon gazını bulmuş ve Nobel Ödülü almıştı. Bu ve benzeri uygulamalar sayesinde bilim insanları kanunun güzelliğinin ve işe yararlığının daha çok farkına varmaya başladı. Francis Galton, "Eski Yunanlılar bu kanunu bilseydi, onu ete kemiğe bürür, tanrı yapıp tapardı" diyordu. Kanunun toplum bilimde ve nihayet zihinsel oluşumların ölçülmesinde kullanılmaya başlanması da çok zaman almadı.

Matematik Yeterliğini Ölçmek

Hataların frekansı kanununun matematik yeterliğinin ölçülmesinde kullanılmasını, yukarıdaki gezegen örneğiyle karşılaştırarak ele alalım. Gezegenin konumu yerine bir öğrencinin matematik yeterlik düzeyini ve farklı ölçme girişimleri yerine de sınav sorularını koyduğumuzda, kanunun zihinsel oluşumların ölçülmesinde nasıl kullanıldığını da genel hatlarıyla görmüş oluruz. Eğitimde kullanılan ölçme teorilerinin dayandığı temel prensip budur. Bu durumda, sınavdaki her bir sorunun, öğrencinin matematik yeterlik düzeyini ölçmek üzere yapılan yeni ve diğerlerine özdeş bir ölçme girişimi olarak ele alınması gerektiği dikkatlerinizden kaçmamıştır. İleride görüleceği gibi, tam olarak sağlanması mümkün olmayan bu varsayım, eğitimde ölçmenin Aşıl topuğudur.

Ölçme Modelinin Üç Varsayımı

Temel güçlük, ölçülen şeyin insanla ilgili soyut bir zihinsel oluşum olmasıdır. İnsan, gezegen örneğinin aksine, önceki ölçme girişimlerinden az veya çok etkilenir. Örneğin, zor bir sorudan sonra morali bozulabilir. Çözdüğü soru diğer soruyu çözerken kullanabileceği bir ipucu içeriyor olabilir. Bir soruyla uğraşırken zihni önceki soruda takılı kalabilir veya en azından bir soruyu çözdükten sonra, diğer soruya biraz daha yorulmuş olarak başlar. Bu durumda, bir sorudan diğerine, ölçmek istediğimiz şeyin kendisi de değişecektir. Ölçme modeli, bu değişimin ihmal edilebilecek bir düzeyde olduğunu varsayar. Bu varsayım test teorisinde, *yerel bağımsızlık varsayımı* olarak bilinir ve “kişinin bir soruya vereceği cevabın, başka bir soruya vereceği cevaptan etkilenmemesi ve cevapların istatistiksel olarak bağımsız olması gereklidir” diye tanımlanır.

İkinci sorun, sınavdaki soruların özdeş birer ölçme girişimi olmasının, diğer bir deyişle, her sorunun aynı zihinsel yeterliği ölçmesinin nasıl sağlanacağıdır. Ölçme girişimleri özdeş olsun diye aynı soruyu tekrar tekrar soramayız bir sınavda. Öncelikle, *matematik yeterliği* tek bir soruyla belirlenebilecek kadar dar değildir. Diğer yandan, böyle kapsamlı bir soru olsa bile, *yerel bağımsızlık varsayımından* ötürü, bir sınavda birden fazla kullanılamayacaktır. Dolayısıyla, bir yandan farklı sorular yazmak, diğer yandan bu farklı soruların aynı şeyi ölçtüğünü garantilemek gerekmektedir. Test teorisinde bu gereklilik, *tek boyutluluk varsayımı* olarak bilinir.

Tahmin edeceğimiz gibi, bu iki varsayımın tamamen sağlanması imkansızdır. Buna rağmen, bu varsayımlar kabul edilebilir bir oranda sağlandığı müddetçe, sınavlardan anlamlı (yani, ölçülmek istenen şeyin düzeyiyle ilgili fikir veren) sonuçlar elde edilebilmektedir. Dolayısıyla bu varsayımları sağlamak için özel bir gayret gösterilmelidir. Aksi hâlde, test sonuçlarının geçerliğinin azalacağı göz önünde bulundurulmalıdır. Peki, nelerdir bu gayretler?

Tek boyutluluk varsayımının sağlanabilmesi için öncelikle, ölçülmek istenen zihinsel oluşum detaylı ve işlevsel bir şekilde tanımlanmalıdır. Hem de o kadar iyi tanımlanmalıdır ki, bir sorunun bu tanımda belirtilen özelliklere sahip olup olmadığına rahatlıkla karar verilebilmelidir. **Eğitimde ve Psikolojide Ölçme Standartları**'nda testin amacının, ölçülmek istenen zihinsel oluşumun, test sonuçlarının nasıl yorumlanacağını, testin içeriğinin ve testteki soruları cevaplamak için gerekli zihinsel faaliyetlerin detaylı bir şekilde tanımlanması gerektiğine dair bir çok madde esasında bu sebeple yazılmıştır.

Türkiye'deki sınavlardaysa bu tanımlamalar ya hiç yapılmamakta, ya da usulen yapılmaktadır. Örneğin ALES'te (Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı) soruların sayısal ve mantıksal akıl yürütme becerilerini ölçtüğünü ve bunların farklı alanlardan gelen yükseköğretim kurumu mezunlarının cevaplayabilecekleri nitelikte olacağı açıklaması var sadece. Mevzuatla ilgili açıklamaların arasında gözümüzden kaçmadıysa, YGS'de (Yükseköğretime Geçiş Sınavı) bu kadarı da yok. SBS'de (Seviye Belirleme Sınavı) ise testlerin, öğrencilerin öğretim programlarında belirtilen kazanımları elde etme seviyesini ölçtüğünü söyleyen bir cümleden başka bir tanımlama yok.

Böyle genel tanımlar, “biz gezegeni ölçüyoruz” benzeri ifadelerdir. Bunun gibi eksik bir tanımla yola çıkıldığında, astronomlardan biri gezegenin diyelim ki, konumunu ölçerken diğeri dünyaya uzaklığını, bir diğeri de çevresini ölçebilir. Bu durumda da bulunan sonuçların ortalaması alınabilir elbet. Ancak, bu ortalamanın neyi gösterdiği belirsizdir artık.

Farklı şeyler ölçmeye yönelik sorulardan oluşan bir sınavda da durum aynıdır. Örneğin, sadece ezberlenmiş alışkanlıklarla bile çözülebilecek bir soru, akıl yürütme ile çözülebilecek bir soru ve çözülebilmesi için uzunca bir metnin okunup anlaşıl-

masını gerektiren başka bir soru, özdeş birer ölçme girişimi olmaktan uzaklaşmıştır. Bu tür soruların aynı testte yer almasının *tek boyutluluk* varsayımını tehdit ettiği göz önünde bulundurulmalıdır. Bir sınavın ölçmeye niyetlendiği zihinsel oluşumun oldukça işlevsel olarak tanımlanması bu sebeple çok önemlidir. Böylece her bir soru, bu tanıma uygunluğu açısından değerlendirilebilir ve muhtemel tehditler önemli oranlarda azaltılabilir.

Bu konu çok önemli olduğu için birkaç iyi örnek de verelim. PISA 2003 (Uluslararası Öğrenci Değerlendirme Programı) çalışmasında ne ölçüldüğü 194 sayfalık bir dokümanla tanımlanmıştır (The PISA 2003 Assessment Framework). Bu dokümanda, matematik testiyle ölçülen zihinsel oluşum, *matematik okuryazarlığı* olarak adlandırılır. Bu, kişinin gerçek yaşamda karşılaşılabileceği bazı durumlarda matematiği kullanarak çözümler üretebilme, söylediklerini destekleyecek kanıtlar sunabilme becerisi olarak tanımlanmaktadır. Öğrencilerden, sorularda verilen gerçek hayat durumunu soyutlayarak, çözüm için gerekli matematiği tespit etmesi beklenmektedir. Dolayısıyla sorularda çoğunlukla, nasıl bir matematiksel işlem yapılması gerektiği açıkça görülmez. Dokümanda ayrıca, bu genel amaca yönelik soruların içeriği, nasıl bir gerçek yaşam durumunda verileceği, cevaplanması için gerekli zihinsel faaliyetlerin (süreçlerin) neler olduğu ve bunların derinliği gibi tanımlamalar teorik alt yapılarıyla birlikte oldukça detaylı bir şekilde verilmektedir. Bu zihinsel süreçlerin tanımlanması da oldukça önemlidir. Üst düzeyde bir zihinsel çaba gerektirecek sorunun nasıl bir soru olması gerektiği böylelikle belirlenebilmekte, soruların zorluğu veya kolaylığına bu sayede karar verilebilmektedir.

TIMSS 2007 (Uluslararası Matematik ve Fen Eğilimleri Araştırması) ise ne ölçtüğünü 182 sayfalık bir dokümanla tanımlamaktadır. TIMSS, öğrencilerin okulda ne öğrendiğine odaklanmaktadır. Dolayısıyla, ne ölçtüğünü öğretim programlarında yer alan kazanımlar (örn., kesirleri karşılaştırır, denk olanları belirler) üzerinden tanımlar. Bu amaçla öğretim programlarına dayanılarak bir *test programı* geliştirilmiştir. Tanımlamalar bu program üzerinden yapılır.

Burada özetleyebildiğimiz kadar tanımla bile, aşağıdaki ilk sorunun bir TIMSS sorusu, ikincisinin ise bir PISA sorusu olabileceği, dolayısıyla bu

iki sorunun aynı testte yer almasının *tek boyutluluk varsayımını* tehdit edebileceği rahatlıkla görülecektir.

Soru 1. *Bir araba yarışında, iki kontrol noktası arasındaki mesafe 160 km'dir. Sürücüler, maksimum puanı almak için, bu mesafeyi tam olarak 2,5 saatte gitmelidir. Bunun için ortalama hızları ne olmalıdır?*

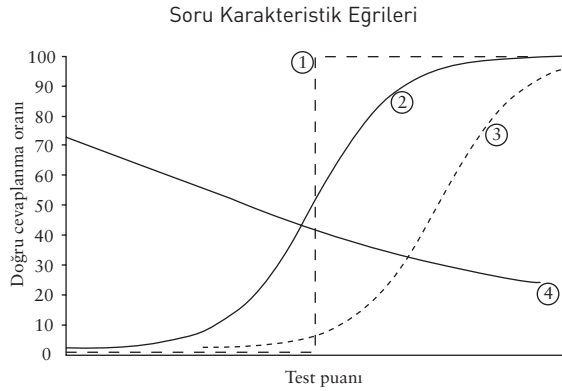
Soru 2. *Belediye meclisi, üçgen şeklindeki küçük bir parkın tümünü aydınlatacak bir sokak lambası dikecektir. Lamba nereye yerleştirilmelidir?*

Bu örneklerin ardından, SBS'de yapılan, testlerin öğrencilerin öğretim programlarında belirtilen kazanımları elde etme seviyesini ölçtüğü tanımının niçin eksik olduğunu biraz daha açalım. Eksiktir çünkü öğretim programları test için değil, adı üstünde, öğretim için geliştirilmiştir. Öğrencilere farklı boyutlarda beceriler kazandırmayı hedefler, dolayısıyla çok boyutludur. Sorular bu öğretim programına göre hazırlandığında, farklı boyutlarda becerilere yönelik olacağından *tek boyutluluk varsayımından* uzaklaşılacaktır. Yapılması gereken öncelikle, eldeki teoriler ışığında temel bir *matematik yeterliği* tanımlamak ve öğretim programındaki kazanımları bu temel tanım çerçevesinde ele alarak düzenlemek (bazı kazanımları birleştirmek, bazılarını almamak vb.) ve bir test programı oluşturmaktır. Amerika'da yürütülen NAEP (National Assessment of Educational Progress, <http://nces.ed.gov/>) bunun yapıldığı iyi bir örnektir.

Yeri gelmişken sıkça karşılaştığımız bir tepkiden de söz edelim. *Tek boyutluluk varsayımına* karşılık olarak, bir sınavda farklı boyutların ölçülmesinin daha iyi olduğu iddiasını sıklıkla duyarız. Elbette öyledir. Sorun şu ki farklı boyutları ölçmeye yönelik sorulardan oluşan bir sınavda kullanılacak ölçme modeli henüz geliştirilemedi (daha doğrusu, yeterince geliştirilemedi). Dolayısıyla, tek boyutlu bir sınav hazırlamak, en iyi ölçme yolu değildir; kullandığımız ölçme modelinin bir sınırlamasıdır. Kaldı ki, *matematik yeterliği* gibi karmaşık bir zihinsel oluşumu ölçmeye yönelik sınavlarda, farklı konulardan ve farklı zorluklarda sorular sorulduğu için, tek boyutluluk varsayımı, istenirse de, bir düzeye kadar ihlal edilir. Bir de bu varsayımı sağlamak için özel bir gayret gösterilmezse,

bu ihlal sınav puanlarının geçerliğini düşürecek bir düzeye çıkacaktır.

Son olarak, ölçme modelinin gerektirdiği üçüncü varsayımı tanıyalım. Bu varsayım, grafikte örneklerini gördüğümüz ve *madde (soru) karakteristik eğrisi* olarak bilinen eğrilerle ilgilidir. Sınavdaki her bir sorunun, 2 veya 3 numaralıya benzer bir karakteristik eğrisine sahip olması gerektiğini söyler. Bu eğri, ilgilenilen sorunun belirli bir toplam test puanına sahip öğrencilerin kaçta kaç tarafından doğru cevaplandığını gösterir. Bir numaralı eğriden başlayarak, konuyu biraz açalım.



Bir sınavın yukarıda açıkladığımız iki varsayımı yüzde yüz sağladığını, tesadüfi hata dahil herhangi bir hatanın söz konusu olmadığını ve sınavın ölçülen zihinsel oluşumu tamamen kapsayacak kadar çok sayıda sorudan oluştuğunu varsayalım. Böylece, toplam test puanının öğrencilerin yeterli düzeyini hatasız bir şekilde gösterdiği zorunlu sonucuna ulaşırız. Bu durumda, bir soruyu doğru cevaplamak için diyelim ki en az X kadar bir yeterli düzeyine sahip olmak gerekiyorsa, bu düzeyin üzerinde bulunan herkes bu soruyu doğru cevaplarken, bu düzeyin altında bulunanlar doğru cevaplayamayacaktır. Sonuçta böyle bir sorunun 1 numaralıdaki gibi bir karakteristik eğrisi olacaktır.

Pratikte, ölçme hatasından ve varsayımların kusursuz bir şekilde sağlanamamasından ötürü, 1 numaralıdaki gibi bir eğri ortaya çıkmaz. Sınavda, X 'ten fazla puan aldığı hâlde bu soruyu doğru cevaplayamayan öğrenciler olacağı gibi, X 'ten az puan aldığı hâlde doğru cevaplayanlar da olacaktır. Ancak, varsayımlar kabul edilemeyecek bir şekilde ihlal edilmediği ve tesadüfi ölçme hatasının dışında kayda değer bir hata söz konusu olmadığı müddetçe, eğrideki bozulma sınırlı olacak ve 2 numaralıya benzer bir eğri elde edilecektir. Tesadüfi ölçme ha-

talarının normal dağılmasından ötürü, küçük hataların daha sık görüldüğünü, hata büyüdükçe görülme ihtimalinin azaldığını yukarıda belirtmiştik. Buna paralel olarak, karakteristik eğrisindeki bozulma da uç noktalarda nispeten daha az olacaktır.

Sınavdaki bazı sorular, diğerlerine göre daha fazla veya daha az zihinsel faaliyet gerektireceği için, böyle sorulara ait madde karakteristik eğrilerinin konumları da farklı olur. Örneğin, doğru cevaplanması için X 'e karşılık gelen düzeyden daha fazla yeterlik gerektiren bir sorunun karakteristik eğrisi, 3 numaralı eğriye benzer şekilde daha sağda bir yerlerde olacaktır. Ancak eğrinin S harfini andıran şekli aşağı yukarı aynı kalacaktır.

Karakteristik eğrisi, örneğin, 4 numaralıdaki benzer bir sorunun ise, testteki diğer sorulardan farklı bir şey ölçtüğü ve/veya bu soruda, tesadüfi ölçme hatasından başka hataların da söz konusu olduğu kolaylıkla görülecektir.

Özetle, madde karakteristik eğrileri sayesinde yukarıda açıkladığımız iki varsayımın kabul edilebilir düzeyde sağlanıp sağlanmadığı kolaylıkla görülebilir. Dolayısıyla önemi büyüktür. Bu sebeple sınavların teknik raporlarında, madde karakteristik eğrileri veya bu eğrilerin elde edilebileceği veriler yayımlanır. Ancak, ülkemizde yapılan sınavların teknik raporları yayımlanmadığı için soruların karakteristik eğrilerini de göremiyoruz.

Teknik Rapor

Türkiye'de her sınavdan sonra, sınav sorularının tamamı yayımlandığı için başka bir şey yayımlamaya gerek olmadığı düşünülüyor olabilir. Oysa, sadece nitelikli soruların bir araya getirilmesi nitelikli bir sınavın garantisi olmaz. Bir sınavın kalitesi, soruların nitelikli sorular olmasının yanı sıra, soruların birbiriyle ilişkisine de bağlıdır. Yukarıda açıklanan varsayımlar ihmal edilirse, örneğin, sorular nitelikli ancak farklı yeterlikleri ölçüyorsa sınav sonuçları sorgulanabilir olacaktır. Teknik rapor, soruların tek tek değil, sistematik bir bütün olarak değerlendirilmesi sonucu elde edilen bilgilerden oluşan bir dokümandır.

Eğitimde ve Psikolojide Ölçme Standartları, yapılan bir testin teknik raporunun yayımlanması gerektiğini belirtir ve böyle bir raporda olması gerekenlerle ilgili de bilgi verir. Bu raporda test geliştirme sürecinden, testin nasıl ölçüklendiğine; hangi ölçme modeli kullanıldığından, bu modelin varsa-

yımlarını sağlamak için neler yapıldığına; bu varsayımların sağlandığını gösteren kanıtlardan, testin geçerli ve güvenilir bir test olduğunu gösteren diğer kanıtlara kadar neler bulunması gerektiği standartlarla belirlenmiştir.

Teknik rapor bu açıdan, binalar için gerekli yapı ruhsatlarına benzer. Bu ruhsatların binaların teknik eksikliği olmadığını, gerekli inşaat şartlarını sağladığının bir göstergesi olması gibi, teknik raporlar da sınavların test teorisinin gereklerini sağladığının bir kanıtıdır. Dolayısıyla, ruhsatsız binaların kaçak sayılması gibi, teknik raporu yayımlanmamış sınavlar da kaçak sayılmalıdır. Bu teknik raporların ne kadar ciddiye alındığı ve ne kadar kapsamlı hazırlandığıyla ilgili bir fikir vermesi açısından PISA 2003 çalışmasının 426 sayfa, TIMSS 2007 çalışmasının 630 sayfa, NAEP 1998 çalışmasının 970 sayfa teknik raporu olduğunu belirtelim.

Sonuç Yerine

Görüldüğü üzere, zihinsel oluşumların ölçülmesinde hata riski yüksektir. Hatta, tesadüfi hatanın dışında, sabit veya sistematik hatalar da söz konusu olabilir. Örneğin öğrencinin cinsiyeti, ailesinin sosyoekonomik düzeyi, okulunun türü gibi değişkenler de ölçme sonuçlarını etkileyebilir. Böyle bir durumda, bir ölçme sonucunun içinde gerçek (G) ve tesadüfi hata (H) değerlerine ek olarak sistematik bir hata da (S) bulunacaktır. Sembolle gösterirsek,

$$X_i = G + S + H_i$$

olacaktır. Bu durumda gözlem (ölçme) sonuçları ortalamasında tesadüfi hatanın etkisi azalsa da, sistematik hatanın etkisi kalacaktır.

Fakat bunların artık çok büyük problemler olmadığını da belirtelim. Çünkü gerek tesadüfi gerekse sistematik hatayı en aza indirmek için gerekli teknik bilgi (standartlar) oluşmuştur. Ayrıca bu hataların muhtemel etkileri hesaplanabilmektedir. Yeter ki yapılması gerekenler hakkıyla yapılsın. Aksi hâlde hatalardan kaynaklanan bütün fatura öğrenciye ve böylelikle toplumun kendisine kesilecektir. Örneğin, standartları karşılamayan bir sınavdan 50 alan bir öğrencinin, 40 alan bir öğrenciden nesinin daha iyi olduğu bilinmeyecektir. Bunun gerçekten kayda değer bir fark olup olmadığı, bu farkın öğrenci hakkında alınacak kararda (örn., hangi okula gidebileceği, öğretmen olarak atanıp atanamayacağı) bir önemi olup olmadığı bilinme-

yecektir. Sonuçta bu tür sınavlar, öğrencilerin elenmesine değil, harcanmasına yol açabileceğinden ülkenin geleceği de harcanmış olacaktır. Dahası ve belki de en önemlisi, böyle sınavlarla esir alınmış bir eğitim sisteminde öğrenciler gerçekten nitelikli bireyler olma ihtiyacını da hissetmeyecektir.

Özetle, ülkemizde öğrencilerle ilgili hemen her karar bir sınavla alındığı hâlde bu sınavlar yeterince sorgulanmamaktadır. Oysa, son zamanlarda sıkça patlak veren kopya, şifre ve benzerlerinden çok daha vahim sorunlarımız var muhtemelen. Örneğin, Gündüz Vassaf (1 Mayıs 2011, Radikal) geçmiş yıllarda yaptığı bir araştırmada, üniversite giriş sınavı sonuçlarının geçerli olduğunu gösteren bir kanıt bulamadığını yazdı. Test standartlarını karşılamayan sınavların geçersiz sonuçlar üretmesi şaşırtıcı değildir. Gelgelelim, böyle sınavlara dayanarak önemli kararlar almak ve bu kararları yeterince sorgulamamak oldukça şaşırtıcıdır ve acıklıdır.

Ne diyelim! Mategrafinin icadı çok gecikmese bari... ♣

Kaynakça

- [1] Osterlind, S., J. Constructing Test Items, Boston: Kluwer Academic Publishers 2002.
- [2] Stigler, S., M., The History of Statistics, Cambridge: The Belknap Press 1986.
- [3] Thorndike, R. L., Applied Psychometrics, Boston: Houghton Mifflin Company 1982.

