Elementary Education Online, 11(2), dy: 1-6, 2012.
İlköğretim Online, 11(2), dy:1-6, 2012. [Online]: http://ilkogretim-online.org.tr

DÜŞÜNCE YAZISI-OPINION PAPER

EDİTÖRE MEKTUP II:

DEĞERLENDİRME, DEĞERLEME, NORM, NORM GELİŞTİRME, NORMALLEŞTİRME NEDİR NE DEĞİLDİR?*

Adnan Erkuş[1]

Bir sözcüğün günlük dildeki karşılığı ile bir alana özgü terim olduğundaki anlamı birbirinden çok farklı olabilir. Bu bölümde; aşağıda ele alınacak terimler konusunda da ne yazık ki, ülkemizde yanlış adlandırmalarla karşılaşmak olasıdır. ***Bu yanlış adlandırmalar yanlış anlamlandırmalara yol açarak, özellikle ölçme ve değerlendirme alanında, yanlış işlemler yapılmasına neden olabilmektedir***. Terim karşılıklarının adlandırılması bu nedenle önemlidir.

"Değerlendirme" deyince, ülkemizdeki, özellikle "ısmarlama" ölçme ve değerlendirme kitaplarına ve bunlar ile bağlantılı KPSS vb. kitaplara ve hatta testlerindeki maddelerine bakıldığında, sanki bir "amentü" gibi, "*mutlak ve bağıl değerlendirme*" tanım ve ayırımlarına rastlanmaktadır. Bu ayırım, işlevsel midir, değerlendirmenin "nasıl" yapılacağı hakkında bir fikir vermekte midir; işlevsel değilse, işlevsel olan nedir? Peki, bir ölçüm sonucu hakkında karar vermek ile bir etkinliğin (örneğin eğitimin) değerlendirilmesi arasında "nüans farkının difrınsı" var mıdır? "Değerlendirme"yi mutlak ve bağıl diye ikiye ayırırsak, "tanı-izleme-nihai değerlendirme"yi nereye koyacağız? Ya, otantik değerlendirme, performans değerlendirmesi,…?

Dilimizde ne yazık ki, hem "assessment" hem de "evaluation" için "değerlendirme" karşılığı kullanılageldiğinden ciddi bir karmaşa ("kargaşa" değil, kargaşa, ne yaptığı anlaşılamayan insan veya hayvan çokluğu ile ilgilidir) yaşanmaktadır. Bu iki terim için de aynı karşılık kullanılınca ne olur? Neler olmaz ki?! Öte yandan, bu terimlerin teknik işlem farklılıklarının yanında, "performans değerlendirme", "iş içinde ve olay anında (authentic) değerlendirme" gibi "nasıl" değil, "ne" sorusuna karşılık olarak kullanılan "değerlendirme" sözcüğü de işin içine girdiğinde terim karmaşası daha da artmaktadır. Bu durum, "analiz" sözcüğünde yaşanan karmaşaya benzemektedir: Analizi teknik anlamda, istatistiksel işlemler olarak da, bir olayın ayrıntılarını irdeleme anlamında da kullanmaktayız. "Test" sözcüğünü, sınav anlamında da, ölçme aracı anlamında da, istatistiksel analiz anlamında da kullanmaktayız. Terim anlamını bozan Türkçe karşılıklara ne yazık ki ülkemizde çok sık rastlanmaktadır (Erkuş, 2000). Ancak, bu durum alanımızda işlem yanlışlarına yol açtığı için biz psikometrist ve ölçme-değerlendirmecilerin, terim karşılıkları konusunda çok titiz davranmamız gerektiğini göstermektedir (Erkuş, 2010). Umarız ki, aşağıdaki kısımlarda, hiç değilse "değerlendirme" konusundaki karmaşayı çözümlemiş oluruz. Bu amaçla ve aradaki ayırımı ayırt etmek amacıyla, daha uygun bir karşılık bulununcaya kadar "assessment"a "***değerleme***", "evaluation"a da "***değerlendirme***" diyeceğiz. Aşağıdaki açıklama ve örnekler, yakında basılacak olan bir kitabımın temelini oluşturan "*Erkuş, 2011 Basılmamış Ders Notları*"ndan değiştirilerek aktarılmıştır (Bu açıklama, daha önceki yazımda da olduğu gibi, *"potansiyel" bir intihali önlemek* amacı taşımaktadır; ne yazık ki bu güzel ülkede çok çirkin davranışlarla karşılaşmak -her an- olasıdır). Aşağıda, tartışmaya açık olmakla birlikte, sözü edilen karmaşayı çözümlemek amacıyla yeni tanım ve sınıflamalara yer verilecektir.

Değerleme (assessment): *Bir tek ölçümü, uygun ölçüt ya da ölçütlerle karşılaştırarak, o ölçüme sahip nesne ya da birey hakkında bir karara, bir yargıya varma işlemi ve süreci.* Değerleme, *somut* ölçümlere ve ölçütlere dayanarak *tamamen nesnel* gerçekleştirilir. *Ölçütün ne olduğuna bağlı olarak* 4 tür *işlemsel* değerleme vardır:

1. Maksimum puana dayalı (alan-dayanaklı)
2. Kesme puanına dayalı
3. Sıralamaya dayalı
4. Grup değerlerine dayalı

Maksimum puana dayalı değerleme: Örneğin, Ali, güvenilir ve geçerli bir ölçekten 60 puan almıştır. Ali (ona ait ölçme sonucu), 60 üstünden (ölçekten alınabilecek maksimum puan) 60 almışsa başka, 600 üstünden 60 almışsa başka değerlenir. İlkinde, Ali o özelliğin %100'üne sahipken, ikincisinde %10'una sahip demektir. Görüldüğü gibi, *Ali'nin puanına değer biçmek için maksimum puandan başka bir şeye gereksinim yoktur.* Ülkemizdeki ölçme ve değerlendirme kitaplarında "mutlak" değerleme için verilen örneklerin çoğu da buna dayanmaktadır: "Ali, 100 maddelik bir başarı testinden 60'ını doğru yanıtlamıştır." gibi. Alanyazında bu tür değerleme alan-dayanaklı değerleme (domain-referenced assessment) olarak da geçmektedir (Berk, 1980;

[1] Prof. Dr., Mersin Üniversitesi Eğitim Fakültesi, adnanerkuspsi@gmail.com

Hambleton, Swaminathan, Algina & Coulson, 1978; Huynh, 1976). Maksimum puana dayalı değerleme için, ölçeğin tekboyutlu olma zorunluluğu olduğu açıktır.

Kesme puanına dayalı değerleme: Aslında, ne düzeyde ölçme yapılırsa yapılsın ve ne tür değerlemeye başvurulursa başvurulsun, sonuçta toplam puanlara göre nihai olarak bir sınıflamaya başvurulur. Bu sınıflama işlemi de, önceden belirlenmiş bir standart (ölçüt-kesme puanı) gerektirir; standart belirleme işlemleri psikometri alanyazınında geniş bir yere sahiptir (Behuniak, Archambault & Gable, 1982). Bu standart, bir bireyin eğitiminde bir üst eğitime geçebilmesi için geçmesi gereken *minimum yeterlik düzeyi* (örneğin, "60 ve üstü alan geçer" gibi); bir duyuşsal özellik için (örn., depresyon) de o duyuşsal özelliğe sahip olanlar ile olmayanları *ayırma düzeyi* (örneğin, "30 ve üstü olan ağır depresyonludur" gibi) olarak tanımlanabilir. Örneğimizdeki, Ali, bir eğitimde geçme puanı 59 ise "başarılı", 61 ise "başarısız" olarak nitelendirilecektir. Bu tür değerlemede de, Ali'nin puanı hakkında bir yargıda bulunmak için kesme puanını bilmek yeterli olacaktır.

Sıralamaya dayalı değerleme: Sıralamaya dayalı değerleme ile sıralama düzeyinde ölçme birbirine karıştırılmamalıdır. Sıralama düzeyinde ölçmede, ölçmenin kendisi sıralamaya dayanır ve sıralamalar arasında bir birim söz konusu değildir. Sıralamaya dayalı değerleme ise, bir *ölçümün* başka *ölçümlerle* karşılaştırılmasına dayanır. Puanlar, büyükten küçüğe veya küçükten büyüğe göre sıralandıktan sonra, ilgili ölçüme, diğer bireylerin puanlarına göre bir değer biçilmiş olur. Ali'nin 60 puanı, diğer bireylerin puanına göre bir anlam kazanabilir: Ali, 65 puanlı Veli'ye göre daha başarısız, 58 puanlı Selami'ye göre daha başarılıdır. Sıralamaya dayalı değerleme, özellikle "kota" söz konusu olduğunda ve/veya ayrık ve genellikle de normal dağılmayan psikolojik özelliklerde başvurulan bir değerleme yöntemidir. Görüldüğü gibi, Ali'nin puanı diğerlerinin puanlarının sırasına göre değer kazanır. Ali, 1. sırada yer alıyorsa ne ala! Öte yandan, alanyazındaki *yüzdelik normlar* da sıralamaya dayalı değerlemenin ta kendisidir. Yüzdelik normlarda, küçükten büyüğe sıralanan puan dizisinde bir birey, grubun yüzde kaçından daha üstte ya da aşağıda olduğuna bağlı olarak irdelenir. Bu tür değerleme daha çok "kota"nın yer aldığı öğrenci veya eleman seçiminde ve ayrık psikolojik özelliklerin (ilgiler, kişilik vb) değerlemesinde önemli hale gelmektedir. Ancak, Ali, diğer bireylerin puanlarına göre kota veya yüzdelik kesme alanı dışında da kalabilir. Örneğin Kuder İlgi Envanteri bu tür değerlemeye bir örnektir.

Grup değerlerine dayalı değerleme: Bu tür değerleme, değerlenecek olan ölçümün içinde yer aldığı grubun ham puan aritmetik ortalaması başlangıç, standart sapması da bir birim şeklinde ele alınarak yapılır. Örneğin, bir yılsonu sınavındaki 100 üzerinden puanlanan bir testten kimse 50 ve üzeri puan almamış; grubun aritmetik ortalaması 35, standart sapması 5 bulunmuş olsun. Bu testten 40 puan almış bir öğrenci için "grubun ortalamasının bir standart sapma üstünde yer alır" değerlemesi yapılabilir. Elbette, bu ham puanların amaca bağlı olarak çeşitli standart puanlara çevrilmesi de söz konusu olabilir. Bir başka grup değerlerine dayalı değerleme, *ölçümler başka değişkenlere göre değiştiğinde* de yapılır. Örneğin "Ali 70cm'dir", şeklinde bir ölçme sonucu, Ali'nin uzun mu kısa mı olduğunu değerlemede bir işe yaramaz. Ali'nin ölçümü, ancak içinde bulunduğu zaman ve mekandaki *yaş* grubunun değerlerine göre bir anlam kazanabilir. Ali, eğer bir aylık bir bebekse başka, bir yetişkinse başka bir değerleme söz konusu olacaktır. Yine, bireylerin reaksiyon süreleri de yaşa bağlı olarak değişiyorsa, fark yaratan yaş grupları için ham puanların standartlaştırılmasına gitmek gerekir. Toplumsal cinsiyet (gender) bilindiği gibi eğitim düzeyine bağlı olarak değişir; bu durumda da, bir toplumsal cinsiyet ölçeğinden alınacak puan da içinde bulunduğu eğitim düzeyine bağlı olarak değerlenir. Bu durum, bireylerin puanlarını yorumlama, değerleme ve karşılaştırmada karşımıza ***norm geliştirme*** gerekliliğini çıkarır. Ancak ne yazık ki, "norm geliştirme" denince de çoğunlukla grup değerlerine dayalı değerleme anlaşılmaktadır.

Yukarıdaki değerleme türlerinden ilk ikisi, başka ölçümlere dayanmadığı için "**mutlak değerleme**"; son ikisi ise, ölçümün değeri ve yeri başka ölçümlere dayandığı için "**bağıl değerleme ("değerlendirme**") olarak da adlandırılmaktadır. Ancak bu ayırım, işlemsel olmadığı için değerlemenin ***nasıl*** yapılacağı konusunda bir fikir vermemektedir. Öte yandan uluslar arası alanyazında ilk ikisi "criterion-referenced assessment", son ikisi de "norm-referenced assessment" olarak da adlandırılmaktadır ki, bu da doğru görünmemektedir; çünkü, tüm değerlemeler (assessment) zaten *ölçüte (criterion) bağlı* olarak yapılmaktadır (Erkuş, 2004). Öte yandan, değerlemeye göre testleri ayırmak ise (ölçüt-dayanaklı 'testler', norm-dayanaklı 'testler' gibi) çok daha fazla yanlıştır. Çünkü, değerleme zaten bir ölçekten elde edilen ölçüm için yapılır; bir ölçekten elde edilen ölçümler ise, amaca bağlı olarak çok çeşitli şekillerde değerlenebilir; hatta birkaçı bir arada da kullanılabilir; bu nedenle, ölçekleri yapılan değerlemeye göre adlandırmak doğru görünmemektedir. Örneğin, WISC için öncelikle kronolojik yaş ham puanları standart puanlara çevrilir (grup değerlerine dayalı), ancak sonra da bu zeka testinin ortalaması ve standart sapmasına göre de kesme puanları (85-115 arası "normal" gibi) belirlenerek bireyin normal, normalin altı vb olup olmadığına karar verilir (kesme puanına dayalı değerleme). Şimdi, zeka testine ne diyeceğiz? Kotanın belli olduğu bir lisanüstü programa öğrenci alınacağında, öncelikle kesme puanına dayalı bir değerleme ("Ağırlıklı toplam puanları 65 ve üstü olanlar…"), sonra da, bu puanlar sıraya dizilerek en yüksek

puan alanların içinden "ilk beşini…" programa kabul etmede (sıralamaya dayalı değerleme) olduğu gibi birden çok değerleme yöntemi kullanılabilir.

Norm geliştirme deyince, ne yazık ki çoğunlukla grup değerlerine dayalı değerleme anlaşılmaktadır. Oysaki yukarıdaki örneklerden de anlaşılacağı gibi, maksimum (alan-dayanaklı) puana dayalı değerleme dışındaki tüm değerlemeler için bir "norm" oluşturma söz konusudur. Bu nedenle, "norm geliştirme-oluşturma" yerine, ***amaca bağlı "ölçüt oluşturma"*** adlandırması, işlemleri daha iyi tanımlamaktadır.

Değerlendirme (Evaluation): Değerlendirme terimi de kullanıldığı yere bağlı olarak çok çeşitli anlamlara gelebilmektedir. Bloom'un (1956) bilişsel davranışların aşamalı sınıflamasında en üstte yer alan değerlendirme, analitik ve eleştirel düşünmenin üzerinde yükselen ve bireyin kendine özgü bir ürün ortaya çıkarmasını ifade eden yaratıcı düşünmeye karşılık gelen bir bilişsel süreçtir. Günlük yaşamda da daha çok, zihinsel irdeleme anlamında ("appraisal") da değerlendirme sözcüğü kullanılmaktadır. Öte yandan, "performans değerlendirme" gibi adlandırmalar da, "neyi" sorusuna karşılık gelmekte, "nasıl" sorusuna yanıt olmamaktadır. *Bu kısımda ele alacağımız değerlendirme ise*, *belirli bir amaç için planlanmış bir etkinlik sürecini ve ürünlerinin değerlendirilmesini kapsamaktadır*: Örneğin bir eğitim programının veya bir psikoterapi programının değerlendirilmesi. *Buradaki teknik anlamıyla değerlendirme, aslında bir araştırma sürecinin ta kendisidir ve görgül verilere dayanarak amaca bağlı olarak yapılacak çok çeşitli istatistiksel analizleri ve işlemleri içerir.* Belirli bir amaç için planlanmış bir etkinlik sürecini ve ürünlerini değerlendirme; bu etkinlikte yer alan aktörleri ve fiziksel koşulları da kapsayabilir. Bu bir eğitim etkinliği ise, eğitim alan birey ya da bireyleri, eğitim veren birey ya da bireyleri, eğitimi planlayıp yürüten yönetici veya yöneticileri; eğitim yapılacak dersliklerin sayısı, büyüklüğü, ısı-ışık durumu, eğitim araç-gereçleri gibi pek çok konuda değerlendirme yapılabilir. Bu anlamda değerlendirme, *tek tek ve pek çok değerlemeleri de içermek* durumundadır. Öte yandan, planlı bir etkinlik söz konusu ise, bu etkinliğin başlangıcında, etkinlik yürütülürken ve etkinlik bitiminde (hatta sonrasında) bir zamansal boyutta değerlendirmeler yapılabilir. Bu bir eğitim programı için de bir psikoterapi programı için de veya başka bir planlı etkinlik için de geçerlidir. Bu boyutuyla bir eğitim programı değerlendirme işi eğitimde ölçme ve değerlendirmecilerin, bir psikoterapi programı değerlendirme işi de klinik psikologların işidir. Endüstri psikologları için de çok çeşitli değerlendirme biçimleri vardır. Burada sadece eğitim programı değerlendirme süreci ve işlemi ele alınacak, ancak yer yer diğer alanlardaki değerlendirmelere de değinilecek veya örnekler verilecektir.

1. Programın başında; tanı amaçlı değerlendirme (diagnostic evaluation):

Bir programa başlayacak olan öğreniciler (learner), eğiticiler, yöneticiler ve fiziksel koşulların değerlendirilmesi tek tek bireyler bazında yapılacağı gibi, bireyin içinde bulunduğu grup ve gruplar arası da yapılabilir. Herhangi bir etkinliğin başlangıcında, söz konusu etkinliğin sağlıklı yürütülebilmesi için tüm bu etkenlerin veya edilgenlerin tanılanması gerekir: Durum ne, ne veya nelerle karşı karşıyayız, bunları nasıl düzenlersek programın yürütülmesini en iyi gerçekleştirebiliriz, gibi sorulara yanıt bulunması gerekir. Burada, değerlemeden (assessment) farklı olarak, birden çok ölçümün (isterse bir tek birey olsun) irdelenmesi (istatistiksel işlemler temelinde, isterse bir grafik çizilsin) söz konusudur. En klasik örneklerden birisi, yabancı dil kursuna başvuran öğrenicilerin yabancı dil bilme düzeylerinin saptanarak ona uygun eğitim verilebilmesi için gruplanmalarıdır: A, B, C düzeyleri gibi. Çünkü eğitim belirli bir hazır bulunuşluk düzeyi gerektirir; heterojen bir gruba eğitim vermek uygun değildir. Oysaki bu klasik örneğin dışında pek çok farklı durumla karşılaşmak ve ona uygun tanı amaçlı değerlendirme yapmak olasıdır. Gerek eğitimde (örneğin YGS ve LYS; SBS gibi) gerekse işletmelerde her türlü öğrenci ve eleman seçimi ile yerleştirme işlemleri tanı amaçlı değerlendirmeleri kapsar. Bir klinik psikolog veya danışman da öncelikle kendisine başvuran bir bireyi tanımak için çeşitli teknik ve işlemlere başvurur: Bireyin hangi özellikleri sorunlu vb. Buradaki amaç, bu planlı etkinliğe başvuran birey veya bireyler ile bu etkinliği yürütecek olanların *durumlarının saptanması*dır. Bir eğitim programı başlamadan önce eldeki eğitmenlerin çeşitli özelliklerini saptayıp ona uygun sınıflara yöneltmek, fiziksel koşulları saptayıp ona uygun düzenlemeler yapmak, vb hep tanı amaçlı değerlendirmeye birer önektir. Bu amaçla ne tür istatistiksel işlemler yapılacağı pek çok etkene bağlı olduğu için burada tümünü ele alabilmenin olanağı yoktur. Ancak birkaç örnek vermek en azından değerleme ile değerlendirme arasındaki farkı görebilmek açısından yararlı olacaktır. Bir lise son sınıf öğrencisinin ilgi ve yetenek alanlarını saptayıp bu öğrencinin ilgi ve yeteneğine uygun alanlara yönlendirilmesi başlı başına tanı amaçlı bir değerlendirme için araştırma niteliğindedir. İlgili eğitim programına başvuranların önceki eğitim, sosyo-ekonomik vb durumlarına ilişkin betimsel istatistikler veya karşılaştırmalar tanı amaçlı değerlendirmeye örnektir. Bunu, yöneticiler, fiziksel koşullar vb.'ne uyarlayabilirsiniz. Bir eğitim programına başvuranların çeşitli demografik veya eğitsel özelliklerine ilişkin yapılabilecek betimsel (descriptive) veya vardamsal (inferential) istatistikler; eğitmenlerin alanlarına veya önceki eğitim alanları veya düzeylerine göre sınıflanması ve ona uygun işekoşulması tanı amaçlı değerlendirmeye örneklerdir. Bir klinik psikoloğa başvuranların çeşitli özelliklerinin saptanması ve ondan sonra yapılabilecek uygulamalara göre sınıflanmaları vs tanı amaçlı değerlendirmeye birer örnektir. Görüldüğü gibi, değerlemeden

farklı olarak bir tek birey bile olsa, birden çok ölçüm elde etmek ve bir şekilde onları istatistiksel olarak irdelemek, tanı amaçlı değerlendirmenin ayırıcı özelliklerini göstermektedir.

2. Program sürerken; biçimlendirme ve izleme amaçlı değerlendirme (formative evaluation):

Eğer bir etkinlik söz konusuysa, bu etkinlik bir süreç oluşturur. Bu etkinliğin nasıl yürü(tül)mekte olduğu, eksik ya da güçlü yanlarının neler olduğu, ona uygun ne gibi düzenlemelerin yapılması gerektiğinin, aktörler ve koşullar açısından saptanması ile ona uygun düzenlemelerin yapılması, izleme ve biçimlendirme amaçlı değerlendirmeye girer. Konulan tanı ya da tanılara göre yürütülmekte olan etkinlik nasıl gidiyor (aksaklıkları vb neler), varsa aksaklık veya eksiklikler bunların saptanması ve nasıl giderileceği bu tür değerlendirmeye dayanır. Bu bir anlamda "süreç değerlendirme" olarak da adlandırılmaktadır. Herhangi bir eğitim birikici özelliğe sahiptir ve önceki öğrenmelerde eksiklikler varsa, bu eksiklik katlanarak sonraki eğitimlere de aktarılır; bu nedenle, varolan eksikliklerin saptanıp giderilmesi son derece önemlidir; aksi halde yeni öğrenmeler inşa edilemez. Peki bu nasıl yapılır? Birey, bireyler vb için…

DÜŞÜNELİM

Süreç ve işlem (process) dilimizde İngilizcedekinden çok daha zengindir. Nedendir bilinmez (!) son yıllarda "süreç *değerlendirme*" diye bir şey çıktı; sanki eskiden süreç diye bir şey yokmuş ve değerlendirilmiyormuş gibi! Süreç veya işlem ne dersek diyelim, *neyi kastediyoruz*? İlköğretim 1. sınıftan 8. sınıfa mı, 2. sınıfın başından sonuna mı, 2. sınıfın bir ünitesinden sonuna mı, bir ünite içindeki bir konunun başından sonuna mı, yoksa yoksa bir etkinliğin (proje, sunum vb) başından sonuna mı? Peki, bunların hangisi *yeni*? Peki bu süreçte ne, nasıl ölçülüp değerlendirilecek, bildiklerimiz ve uyguladıklarımızdan farklılar mı? Tüm bunlar ayrı (veya *alternatif*) *değerlendirme* biçimi olabilir mi? Ne yazık ki, bugün bu anlamda "süreç izleme ve değerlendirme" dersleri bile açılmaktadır! Tüm ölçme tarihi boyunca, hangi kitap vs. "süreç değerlendirilmez" dedi ki?! Arasınavları, "quiz"ler, popquize"ler neydi ki? Maddelerin güçlük düzeyleri (p) niçin belirleniyordu ki, "laf" olsun diye mi? Süreç mi, işlem mi, hangisi? Ne zaman, ne için? Giyimde moda olabilir, ama bilimde?!

Birkaç farklı aktör ve süreç için birkaç örnekle izleme-biçimlendirme amaçlı değerlendirmeyi açıklamaya çalışalım:

Ünite 1			Ünite 2				Ünite 3		
a	b	c	d	e	f	g	h	i	j
0,98	0,78	0,88	0,45	0,15	0,43	0,38	0,12	0,09	0,05

Bir arasınavında kullanılan 10 ayrı kazanımı yoklayan 10 maddelik testin yukarıdaki madde güçlükleri incelendiğinde, sınıfın 1. ünitedeki kazanımların tümüne ulaştığı, 2. ünitede sınıf başarısının düştüğü ve e kazanımına ulaşılamadığı, 2. ünitedeki bu öğrenme eksikliklerinin 3. üniteye etkisinin çok daha olumsuz olduğu söylenebilir. Bu değerlendirme, öğretmene öncelikle 2. ünitedeki konuları ve kazanımları tekrar etmesi gerektiğini göstermektedir.

İki farklı öğretmenin 2 saatlik derslerini haftalara göre ne sürede tamamladıklarına ilişkin veriler şu şekilde olsun;

	1.	2.	3.	4.	5.	6.	7.	8.	9.
A	60dk	61dk	59dk	60dk	60dk	61dk	58dk	60dk	61dk
B	60dk	60dk	55dk	50dk	48dk	45dk	45dk	40dk	27dk

Tablodan da anlaşılacağı gibi A öğretmeni, ders sürelerini etkili ve kararlı bir şekilde kullanırken, B öğretmeni haftalar ilerledikçe "ipe un sermektedir".

Yine eğitim sürerken, dişi ve erkek öğrencilerin arasınav ortalamalarını karşılaştırıp, hangi grubun daha başarılı olduğu veya farklı 3-4 sınıfta aynı derse giren bir öğretmenin sınıflar arası başarıları karşılaştırıp, hangi sınıflarda başarının düşük olduğunu saptaması (ve sonra ona göre ek düzenlemeler yapması ve önlemler alması) izleme biçimlendirme amaçlı değerlendirmelere birer örnek olabilir.

Ayrıca, bir performans sergilenmesinin gerekli olduğu bir durumda, örneğin bir motor sarma işinde öğretmen, öğrencinin motor sarma aşamalarının hangisinde hata yaptığını gözlem veya değerlendirme ölçekleriyle saptayabilir; bu da izleme-biçimlendirme amaçlı bir değerlendirmeye girer.

3. Program bitiminde; bitim amaçlı bütünsel değerlendirme (summative evaluation):

Bu tür değerlendirme, eğitim veya bir başka etkinliğin sonunda yapılır. Bu aşamada da hem bir tek birey (öğrenici, öğretmen, okul, müdür, vb) hem de grupların (farklı öğretmen, öğrenci, sınıf, okul vb) durumları değerlendirilebilir. 'Etkinliğin başında öne konulan kazanımlara-hedeflere kim ne kadar ulaştı?' sorusuna bu aşamada yanıt aranır. Elbette, bu aşamadaki değerlendirme, gelecek etkinlikler için veri kabul edilip ona göre düzenlemeler yapmaya olanak sağlayacaktır.

Bu tür bir değerlendirmede bir bireyin, eğitim sonunda hangi kazanımlara ulaştığı, hangilerine ulaşamadığı saptanabilir ve buradan hareketle bu bireyin hangi alanlarda çalışırsa daha uygun olacağı yordanabilir. Sınıflar, başka altgruplar, okullar ve okul türleri arasındaki başarı oranları karşılaştırılıp, ona uygun çıkarımlar yapılabilir. Elbette, merak edilen bu sorulara yanıt bulabilmek için veri ve grupların özelliklerine göre çok çeşitli istatistiksel işlemlere başvurulabilir. Bitim amaçlı değerlendirmenin asıl amacı *geriye ve ileriye doğru çıkarımlar*da bulunmaktır: Verilen eğitimin aksayan ve güçlü yanlarını görmek ve gelecek için yönlendirme yapmaya ve önlemler almaya hizmet etmesidir.

4. Takip amaçlı çalışma ve değerlendirmeler (follow-up evaluation):

Eğitimin temel amacı, insanları yaşama hazırlamaktır. Bu nedenle, bundan önceki aşamaların tümü bu amaca yöneliktir. Ancak, ne yazık ki, özellikle eğitimde değerlendirme ile ilgili alanyazın ve işlemlerde, en zayıf kalınan hatta görmezden gelinen değerlendirme aşaması bu aşamadır. Oysa ki, verilen eğitime ilişkin en iyi geribildirim alınacak aşama budur; bu aşama, verilen eğitimin en iyi ölçütüdür; bu anlamda, bu tür değerlendirme, verilen eğitimin bir anlamda yordama geçerliğine kanıt bulma aşamasıdır. Eğitimden sonra "saldım çayıra mevlam kayıra" anlayışının bireysel ve toplumsal sonuçları ortadadır. Bu sonuçlarda, siyasal-ekonomik ve yönetsel pek çok temel etken etkili olmasına rağmen, verdiğimiz eğitimin özelinde de bir işe yararlılıktan söz edebilmek neredeyse olanaksızdır; bunu itiraf etmek ve özeleştiri yapmak durumundayız. Sanayi, meslek lisesi veya meslek yüksekokulu çıkışlıların aldıkları eğitimin çok eksik, ilkel ve işeyaramaz olduğu konusunda hemen hemen hemfikirdir. Aldığı eğitim doğrultusunda (diğer makro etkenleri göz ardı edersek) kendi önünü açamayan, alan bilgi ve becerisiyle donanık olmayan, ne yazık ki pek çok mühendis vs yetiştirmeye devam etmekteyiz. Oysa ki, her eğitim kurumu yetiştirdiği bireyleri takip ederek eksikliklerini saptayabilir ve ona uygun düzenlemeler yapabilir.

Takip amaçlı çalışmalarda ve değerlendirmelerde şu sorulara yanıt aranabilir:

- Mezunlarımızın kaçı, aldığı eğitim ile ilişkili bir işe girdi?
- İşe girenler ile giremeyenlerin, aldıkları eğitimdeki başarıları (ve hangi kazanımlar için) farklı mıdır?
- Girdikleri işlerde, aldıkları eğitimin yetersiz veya eksik olduğu yerler nelerdir?
- Girilen işlerdeki değişim ve gelişmeler nelerdir? gibi daha pek çok soruya yanıt bulmak amacıyla;

takip çalışmalarının yapılması zorunludur. Aksi halde, ne tür çalışmalar yaparsak yapalım, ne tür süslü düzenlemeler yaparsak yapalım; eğitim sonrası takip çalışması yapılmadığında hepsi pratik yaşamın gerisinde kalacaktır.

Yukarıda ele alındığı gibi, elde edilen bir ölçüm hakkında bir karar verebilmek için o ölçümü bir ölçüte göre irdelemek gerekir ki bu sürece "değerleme" (assessment) demiştik. Değerleme yapılabilecek bu ölçütler de, ölçekten alınabilecek maksimum puan, kesme puanı, bireylerin birbirine göre sıralanmaları ve grup değerleri olabilir. Öte yandan değerlemede bu ölçütlerin biri kullanılabileceği gibi, birkaçı bir arada da kullanılabilir. Maksimum puan dışındaki ölçütleri oluşturma sürecine *norm geliştirme* diyebiliriz. Bu anlamda, yüzdelik normlar, standart puan normları ve kesme puanı normlarından söz edebiliriz. Norm, sosyolojide, "yazılı olan veya olmayan grup kuralları" olarak ele alınırken, psikolojide "*bir grubun davranış özeti*" olarak ele alınabilir. Psikolojik anlamda normların, mutlaka görgül işlemler sonucunda elde edilen istatistiksel sonuçlara dayanması zorunluluğu vardır. Bu anlamda psikometrik değerleme, değerleyiciden bağımsız, nesnel bir ölçüte dayanır ve öznellik içermez.

Öte yandan, ne yazık ki, *norm geliştirme* ile *normalleştirme* de birbirine karıştırılan iki terim ve işlemdir. Bilindiği gibi, pek çok insan özelliği çok sayıda bireyden toplandığında frekans dağılımı normal dağılır; ancak bazı özelliklerimizin de (özellikle patolojik özelliklerde) normal dağılmadığı gözden kaçırılmamalıdır. İstatistiksel dağılımlar ve bunlara dayalı analizler de bu gerçeğe dayanır. Bazı durumlarda veri toplama sürecindeki örneklemden kaynaklı olarak normal dağılması *beklenen* veriler normal dağılmayabilir. Örneğin, örneklemin yapısının veya ölçeğin yapısının ölçülen psikolojik özelliğin ranjını ve dağılımını yansıtmaması söz konusu olabilir: Dil gelişimini ölçen bir ölçeğin normları için veri toplanırken, hemen el altındaki kolay ulaşılabilen anaokullarındaki öğrencilerden veri toplandığında (convenience sampling), dezavantajlı gruplar temsil edilmediğinden, verilerin dağılımı sola kayışlı olabilir ya da ölçeğin içinde zor maddeler bulunmadığı zaman verilerin yine sağa yığmalı (tavan etkisi) olabilir. Bu tür durumlarda, ya ölçek tekrar gözden geçirilir ya yeniden dezavantajlı gruplardan da veri toplanır. Elbette, örneklemdeki verilerin norm geliştirme açısından az olması da verileri çarpıklaştırır ki, bu daha ciddi bir hatadır. Yine verilerin ölçek düzeyleri (milimetrelerle, oranlarla veya makro büyüklüklerle) de çarpıklığa yol açabilir. Eğer, tüm bu sakıncalar yok ve başka da yapacak bir işlem kalmamışsa ve de ölçülen özelliğin normal dağılması *bekleniyor da dağılmıyorsa*, verilerin veya norm değerlerinin *normalleştirilmesi* yoluna gidilir. Veriler normallik gerektiren analizler yapmadan önce, ölçek düzeylerine ve dağılıma bağlı olarak 1/X, Karekök X, logX, arcsinX gibi yöntemlerle normalleştirilir. Norm geliştirirken ise, özellikle bağıllık ve özelliğin normal dağılmasının *beklenmesi* söz konusu olduğunda, norm değerlerinin yığmalı dağılımını dikkate alarak ve daha çok da grafik yöntemle normalleştirilir. Görüldüğü gibi, norm geliştirme ile normalleştirme aynı işlemler değildir. Öte yandan,

norm geliştirme, yukarıda ele alındığı gibi, sadece verileri standart puanlara çevirmek de değildir. Kesme puanına, sıralamaya ve grup değerlerine dayalı norm geliştirmelerden hangisinin (veya hangilerinin) kullanılacağı, ölçeğin puanlama biçiminden bağımsızdır, bir parça ölçeğin ölçeceği özelliğe bağlıdır; ancak önemli olan ölçeğin kullanılma amacıdır.

Şimdi isterseniz tekrar başa dönelim: "*Değerlendirme, mutlak ve bağıl olmak üzere ikiye ayrılır*" sözü, yukarıdaki açıklamalardan sonra ne kadar anlamlı? Tüm "ısmarlama" EÖD kitapları ve bunlara bağlı olarak 'ilgili testleri hazırlayanlara' "yanlışa devam mı" diye sormak hakkımız değil mi?

Kaynaklar

Behuniak, P., Jr. Archambault, F. X. & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretattion. *Educational and Psychological Measurement, 42*, 247-252.

Berk, R. A. (1980). A consumer's guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17(4),* 323-349.

Bloom, B. S. (1956). *Taxonomy of educational objectives handbook I: The cognitive domain*. New York: McKay.

Erkuş, A. (2000). Bazı psikometrik terimlerin türkçe karşılıklarında yaşanan sorunlar. *Türk Psikoloji Yazıları, 3(6),* 31-40.

Erkuş, A. (2004). The proposal of a new conceptualization for validity and criterion-referenced assessment. *Eurasian Journal of Educational Research, 16*, 113-117.

Erkuş, A. (2010). Psikometrik terimlerin Türkçe karşılıklarının anlamları ile yapılan işlemlerin uyuşmazlığı. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1(2), 72-77.*

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48(1),* 1-47.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.