Türk Psikoloji Yazıları
2000, 3 (5) 1-21

Ölçüt-Dayanaklı Değerlendirme Bağlamında Geliştirilen Sınıflama Tutarlığı İndeksleri ve Bazı Sorunları

Adnan Erkuş*
Mersin Üniversitesi

Özet

Psikolojik ölçme araçlarıyla, bireyler hakkında pek çok yaşamsal karar verilmektedir: seçme, yerleştirme, tanı koyma gibi. Bu kararların çoğu da, ölçek puanları üzerinde belirlenen bir ölçüte göre, bireylerin sınıflanmasına dayanarak verilmektedir. Bu sınıflamanın ne kadar tutarlı yapıldığına ilişkin ise çeşitli indeksler geliştirilmiştir. Ancak bu indekslerin büyük ölçüde keyfi olarak belirlenen bir kesme puanına dayanmasından dolayı, indeks değerlerinin kesme puanının, puan dağılımı üzerindeki yerine göre radikal değişiklikler gösterdiği çeşitli görgül çalışmalarla gösterilmiştir. Ayrıca, bu indekslerin adlandırılmasında bir kavram karmaşası da yaşanmış ve bu alandaki çalışmaların sürekliliğini olumsuz yönde etkilemiştir. Yeni bir kavramlaştırmaya gidilerek bu sorunun aşılabileceği öne sürülebilir.

Anahtar sözcükler: *Ölçüt-Dayanaklı Değerlendirme, Sınıflama Tutarlığı*

Consistency of Classification Indices and Common Problems Within the Context of Criterion-Referenced Assessment

Abstract

Generally, important decisions about individuals, such as selection, replacement, and diagnosis, are taken on the bases of their test score. These decisions are usually taken considering test score on a given test depending on the classification of individuals. A number of indices have been developed to assess the consistency of these classifications. However, since the majority of these indices are based on an arbitrarily determined cut-off score, as demonstrated in recent empirical studies, the values of these indices may show dramatic changes according to its place in the distribution of scores. Furthermore, there has been a conceptual confusion in the definition of these indices that influences the continuation of studies on this issue negatively. It can be suggested that the conceptual confusions regarding criterion-referenced assessment can be overcome by agreeing on a reconceptualization of the basic indices.

Key words: *Criterion-referenced Assessment, Classification Consistency*

* Yazışma Adresi: Yrd. Doç. Dr. Adnan Erkuş, Mersin Üniversitesi Eğitim Fakültesi, Yenişehir Yeşleşkesi, Mersin.
E-posta: adnanerkus@yahoo.com

Bir ölçme işleminin sonunda elde edilen bir ölçüm tek başına bir şey ifade etmez. Herhangi bir ölçme sonucu, bir ölçüte göre ele alındığında ve buna göre bir yargıya varıldığında anlam kazanır (Turgut ve Baykul, 1992) ve bu sürece de "değerlendirme" denir. Neyin ölçüt olarak alındığı ise birkaç tür değerlendirme biçimini ortaya çıkarır. Herhangi bir ölçüm, o ölçümün alındığı grubun normları dikkate alınarak değerlendiriliyorsa *"norm-dayanaklı (norm-referenced) değerlendirme"* olarak adlandırılır. Glaser'la (1963) birlikte, herhangi bir ölçme sonucunun grubun değerlerinden bağımsız olarak değerlendirilebileceği öne sürülmüş ve bu değerlendirme biçimi de *"ölçüt-dayanaklı (criterion-referenced) değerlendirme"* olarak adlandırılmıştır. Süreç içinde ölçüt-dayanaklı değerlendirme, sadece kesme puanının (cutoff score) ölçüt alındığı durumlar için kullanılmış (bu nedenle, *"kesme puanına dayalı değerlendirme"* olarak adlandırılması daha uygun olabilir); ölçütün, maksimum test puanının olduğu ya da bir alandaki davranış hedeflerinden ne kadarının kazanıldığının ele alındığı değerlendirme biçimi ise *"alan-dayanaklı (domain-referenced) değerlendirme"* olarak adlandırılmıştır.

Psikolojik ölçme araçları sınıflama, seçme ve yerleştirme gibi, insanlar hakkında verilecek pek çok karar için geliştirilir. Bu kararların verilmesi ise, ister istemez, o kararın dayanacağı bir ölçütün belirlenmesini gerektirmektedir. Örneğin bir öğrencinin herhangi bir dersten geçip geçmediği kararı, öğrencinin çoğu kez 100 üzerinden en az 50 puan alması gerektiği gibi bir kesme puanını gerektirir. Bu durumda, bu başarının ölçüldüğü ya da verilecek kararın dayandığı ölçme aracının, kesme puanının üstündeki ve altındaki bireyleri ne denli tutarlı sınıfladığı ile ilgilenilir. Bu tutarlığın sayısal olarak saptanması yönünde çok sayıda indeks geliştirilmiştir (ayrıntılı bilgi için bkz. Berk, 1980). Ancak, yapılan görgül çalışmalar bu indekslerin hemen tümünün kesme puanına dayanmasından kaynaklanan sakıncaları ortaya çıkarmıştır (Crocker ve Algina, 1986; Huynh, 1976; Huynh, 1978; Subkoviak, 1976). Bu durum, bu indeksler üzerinde yapılan çalışmaların sorgulanmasını da beraberinde getirmiştir (Glaser, 1994; Linn, 1994; Millman, 1994).

Bir ölçme sonucunu değerlendirmek (ya da bir birey hakkında karar vermek) için herhangi bir ölçüte başvurulması kaçınılmazdır. Kesme puanı kullanmak da çoğu durumda bu bakımdan gerekli, hatta zorunludur. Ancak, verilecek sınıflama kararının tutarlığını belirlemeye yönelik geliştirilen indekslerin mutlaka kesme puanına dayanması gerekip gerekmediği ise tartışmaya açıktır. İlerde de görüleceği gibi, indekslerin kesme puanına dayanması çeşitli sıkıntıları da beraberinde getirmektedir. Bu bakımdan, kesme puanını bertaraf edecek yeni yöntem ve indekslerin geliştirilmesi bir gerekliliktir. Bu makalede, kullanılagelen sınıflama tutarlığı indeksleri ve bunların zaaflı yanları ele alınacaktır.

Üç Değerlendirme Türü Arasındaki Farklılıklar

Ölçüt-dayanaklı ve norm-dayanaklı değerlendirme ayırımı, Glaser'a (1963) dayanmaktadır. Ona göre, başarı testlerindeki puanlar iki tür bilgi verir. Birincisi, bireyin puanının, tüm puan dağılımındaki göreli (bağıl) pozisyonudur ve standart puanlarla rapor edilen test performansı bu bilgiyi sağlar. Test performansının yorumunda kullanılan bu standart puan görelidir ve bireyin puanı norm-dayanaklı ölçü olarak adlandırılır. İkincisi, bireyin, eğitim hedeflerinin ne kadarını kazandığı hakkında bilgi verir ki, bu durumda bireyin puanı, test edilen diğer bireylerin

performansını dikkate almaksızın yorumlanabilir. Bu durumda bireyin performans düzeyi (genellikle, doğru cevap puanının oranı) ölçüt-dayanaklı ölçü olarak adlandırılır. Glaser'a (1963) göre, iki bilgi arasındaki temel fark, test performansının dayandığı standarttır. Ölçüt-dayanaklı ölçü terimi, ölçümlerin, bir öğrencinin sergilediği ölçüt davranışlar anlamında yorumlanmasını; norm terimi ise, bu gibi ölçümlerin grup normları anlamında yorumlanmasını ima eder. Bir örnekle açıklanacak olursa, en yüksek puanı 100 olan bir testten bir birey 60 almışsa, bu bireyin testin ölçtüğü görevler alanından %60'ını başardığı anlamı çıkarılabilir ve bu da diğer bireylerin performansı dikkate alınmadan yorumlanabilir. Aynı bireyin içinde bulunduğu grubun bu testteki puanlarının ortalaması 50, standart kayması 5 ise, bu bireyin, grubun ortalamasından 2 birim standart kayma üstte yer aldığı yorumu yapılabilir. Bu bakımdan, norm-dayanaklı değerlendirme grubun performansına göre yapıldığından bağıl değerlendirme; yukarıdaki örneğin ilk durumunda ise, bireyin puanı grubun performansından bağımsız yapıldığından mutlak değerlendirme olarak da anılmaktadır (Crocker ve Algina, 1986; Ornstein ve Gilman, 1991).

Norm-dayanaklı ve ölçüt-dayanaklı değerlendirmeler arasındaki farklılıkları ele alan Popham ve Husek'e (1969) göre, "... bir teste bakarak onun norm-dayanaklı ya da ölçüt-dayanaklı test olduğunu söylemek mümkün değildir. Tersini düşünmek kolay olmasına rağmen, aslında bir ölçüt-dayanaklı test, aynı zamanda norm-dayanaklı test olarak da kullanılabilir. Ancak, bu iki yaklaşım arasında önemli farklılıklar da vardır" (s.2). Yazarlara göre, norm-dayanaklı yaklaşımda, bireyin puanının anlamlılığı, diğer bireylerin (normatif grup) puanıyla karşılaştırıldığında ortaya çıkar; bu bakımdan çoğu standart test norm-dayanaklı olarak sınıflanabilir. Ölçüt-dayanaklı yaklaşım ise, bireyin durumunu bir ölçüte (yani performans standartına) göre belirlemek için kullanılır. Bu durumda, bireyin puanının anlamlılığı, diğer bireylerin puanlarıyla karşılaştırılmasına bağlı değildir. Yine Popham ve Husek'e (1969) göre, norm-dayanaklı test puanının anlamlılığı diğer puanlarla karşılaştırıldığındaki göreli duruma bağlı olduğundan, puanların değişkenliği (varyansı) ne kadar büyükse, o kadar tercih edilir. Oysa, ölçüt-dayanaklı testlerde puanın anlamı diğer puanlarla karşılaştırılmasına bağlı olmadığından, puan değişkenliğinden çok, ölçüt ile maddeler arasındaki bağlantı önemli hale gelir. Bu bakımdan klasik güvenirlik belirleme yöntemleri ile madde seçme yöntemleri ölçüt-dayanaklı testler için uygun değildir; klasik iç tutarlık yöntemleri yerine ne konulacağı ise açık değildir. Yine ölçüt-dayanaklı yaklaşımda puanların yapısı "geçer-kalır" biçiminde olduğundan, yani birey için kullanılan ölçüt ya geçmiş ya da geçememiş olduğundan, norm-dayanaklı yaklaşımda kullanılan yüzdelik sıralar ya da standart puanlar gibi grup betimleyicileri uygun değildir.

Bir bireyin, bir eğitim hedefleri alanındaki beklenen görevlerden ne kadarını başardığı ile ilgileniliyorsa (böyle bir durum, Glaser ile Popham ve Husek'in ilk ölçüt-dayanaklı kavramsallaştırmasına uygun düşmektedir), bireyin alan puanının ne kadar güvenilir kestirileceği önemli olur. Bunun için alan ya da evrendeki davranış hedeflerinin tanımı çok önemlidir. Test, bu alandaki hedefler örneklenerek oluşturulur. Bireyin alan puanı, bir alanda doğru cevaplandırdığı maddelerin tüm maddelerin sayısına oranıdır. Bu alan puanı, aynı zamanda genellenebilirlik kuramının evren puanıdır. Buradaki genelleme evreni, alandaki tüm maddeleri içerir ve bireyin alan ya da evren puanı, genelleme

evrenindeki tüm maddeler üzerinden ortalama puanı olarak tanımlanır. Dolayısıyla sorun, gözlenen cevap oranından, beklenen evren (ya da alan) puanının kestirimine dönüşür. Bunun yanıtı genellenebilirlik kuramıyla bulunmaya çalışılmıştır (Gleser, Cronbach ve Rajaratnam, 1965; Cronbach, Gleser, Nanda ve Rajaratnam, 1972). Glaser (1963) ile Popham ve Husek'in (1969) makaleleri incelendiğinde, "ölçüt" sözcüğünü "davranış alanı" (behavior domain) anlamında kullandıkları görülür (Hambleton, Swaminathan, Algina ve Coulson, 1978). Bu anlamda ölçüt-dayanaklı "test", iyi tanımlanmış bir davranış alanına göre bireyin alan puanı olarak ifade edilen pozisyonunu belirlemek için kullanılmaktadır (Popham, 1975). Bu bağlamda Millman (1974), "alan-dayanaklı" (domain-referenced); Hambleton ve arkadaşları (1978) "hedeflere-dayalı" (objectives-based), "performansa-dayalı" (performance-based), "becerilere-dayalı" (skills-based) ya da "yeterliğe-dayalı" (competency-based) terimlerini kullanmışlardır.

Hambleton ve arkadaşlarına göre (1978), ölçüt-dayanaklı, alan-dayanaklı ve hedeflere-dayalı "testler" arasında önemli farklar vardır ve bu da karmaşaya neden olmaktadır. Karmaşanın önemli kaynaklarından biri olarak da, "ölçüt" sözcüğü gösterilmektedir. Bu saptama doğru görünmektedir, çünkü, ölçüt-dayanaklı değerlendirmede ölçüt, kesme puanı; alan-dayanaklı değerlendirmede testten alınabilecek maksimum puan; norm-dayanaklı değerlendirmede de grup performansıdır. Değerlendirme, tanımı gereği bir ölçüte dayanır. Bu bakımdan, ölçüt-dayanaklı, alan-dayanaklı ve norm-dayanaklı ayırımı, değerlendirme boyutundaki ayırımdır. Ancak, daha sonra ayrıntılı bir şekilde ele alınacak olmasına rağmen, ilgili tüm yazarların değerlendirme boyutundaki ayırımı, "testler", "ölçme" gibi kavramlarla ölçme boyutuna taşıdığı, bunun da karmaşaya yol açtığı görülmektedir. Bu durum, Hambleton ve arkadaşlarında (1978) örtük (implicit) bir şekilde ele alınmaktadır:

"Norm-dayanaklı ölçmelere uygun ölçme araçları yapımı konusunda bildiğimiz işlemler mevcuttur. Peki, hedeflere-dayalı ya da ölçüt-dayanaklı testler için alternatif işlemler gerekli midir? Bu testler için, norm-dayanaklı testlerden tümüyle farklı işlemler gerektiğine ilişkin kuşku vardır; norm-dayanaklı testler, bazı güçlüklere rağmen, ölçüt-dayanaklı ölçmede kullanılabilirler. Ölçüt-dayanaklı ölçme için geliştirilen bir test de bazen norm-dayanaklı ölçme yapmak için kullanılabilir ve kullanılmaktadır. (...) Testleri norm-dayanaklı ya da ölçüt-dayanaklı olarak ikiye ayırmanın yanlışlığa yol açabileceği, test geliştiriciler tarafından tartışılmaktadır" (s. 3-4).

Burada, "testlerin" norm-dayanaklı/ölçüt-dayanaklı diye ayrılmasının yanlışlığı ve ayrılması gerekenin bir ölçüte göre "değerlendirme" boyutu olması örtük bir şekilde vurgulanmasına rağmen, kullanılan terminolojinin bu karmaşayı içerdiği görülmektedir. Durum böyle olunca, ölçme aracı ile ilgili olmayan bir "güvenirlik"ten bile söz edilebilmektedir.

Hambleton ve arkadaşları (1978), ölçüt-dayanaklı "test güvenirliğini" üç kategoride ele almışlardır: (a) "geçer" (mastery) sınıflama kararları güvenirliği (bir test formunun tekrarlı ölçümleri ya da paralel test formlarıyla "geçer-kalır" sınıflama kararı verme tutarlığı), (b) ölçüt-dayanaklı "test" puanlarının güvenirliği (paralel ya da seçkisiz paralel test formlarıyla, tek tek puanların kesme puanından sapmalarının karelerine dayanan tutarlığı), (c) alan puanı kestirimlerinin tutarlığı. Berk (1980) bu sınıflamayı temel alarak literatürü gözden geçirmiş ve (a) kategorisine giren indeksleri "eşik-kayıp fonksiyonu" (threshold

loss function); (b) kategorisine girenleri "karelenmiş-hata kayıp fonksiyonu" (squared-error loss function) indeksleri şeklinde gruplamıştır. Berk'e göre, (c) kategorisine giren ve bireyin alan puanının ya da doğru cevaplar yüzdesinin (proportion correct) kararlılığını kestirmek için kullanılan indeksler daha az dikkat çekmiştir.

Yukarıdaki açıklamalardan hareketle, "ölçüt-dayanaklı testlerin güvenirliği" yerine, "sınıflama kararlarının tutarlığı" şeklinde ifadelendirme daha yerinde olacak gibi görünmektedir. Çünkü, ölçüt-dayanaklı/norm-dayanaklı/alan-dayanaklı ayırımı, ölçme araçlarının ne kadar güvenilir ölçme yaptıklarına değil, ölçme araçlarıyla elde edilen ölçümlerin kullanılma amaçlarına ve seçilen ölçüte dayanmaktadır. Bir başka deyişle, ölçme sürecinin ya da işleminin değil, ölçme sonuçlarının değerlendirilme ve yorumlanma aşamasında bu ayırım önemli olmaktadır.

Sınıflama Kararlarının Tutarlığına İlişkin Geliştirilen İndeksler

Ölçüt-dayanaklı testlerin, norm-dayanaklı testler için önemli olan puan değişkenliğini gerektirmediği savından hareketle (Popham ve Husek, 1969), ölçüt-dayanaklı testler için alternatif indeksler arayışı başlamıştır. Bu indeksler, tarihsel gelişim içerisinde aşağıda kısaca ele alınacaktır.

Cox ve Graham (1966), eğitim hedeflerinin ardışık bir yapıya sahip olduğu durumlara uygun biçimde ölçeklenmiş ölçüt-dayanaklı testler için "yeniden üretilebilirlik" (reproducibility) katsayısı önermişlerdir (aktaran, Hambleton ve Novick, 1973). Yeniden üretilebilirlik katsayısı, belirli bir sıradaki maddenin bir grup birey tarafından aşılma (başarı) derecesinin bir ölçüsüdür. Fakat, bu indeks pek kullanım alanı bulamamıştır.

Carver da (1970), ölçüt-dayanaklı testlerin "güvenirlikleri" için iki istatistik önermiştir (aktaran, Hambleton ve Novick, 1973). İlk yöntem, ölçme aracının iki uç karşılaştırma grubuna uygulanmasına dayanmaktadır ve belirlenen performans ölçütünü aşma açısından iki grup arasındaki fark ne kadar büyükse, testin o kadar güvenilir olduğu sonucuna varılmaktadır. İkinci yöntem ise, bireylerin paralel testlerde ölçütü aşma yüzdelerinin karşılaştırılmasına dayanmaktadır. Carver tarafından önerilen yöntem, sadece kesme puanını geçen bireylerin (masters) grup yüzdesinin aynı kalıp kalmadığını yansıtması ve bireysel sınıflama kararlarının tutarlığına duyarlı olmaması nedenleriyle önemli görülmemiştir (Berk, 1980).

Livingston (1972a) tarafından geliştirilen katsayı ise, ölçüt-dayanaklıdan çok norm-dayanaklı testlere yakın bir katsayı olarak görülmüştür (Harris, 1972; Hambleton ve Novick, 1973). Ancak, Livingston katsayısı, klasik yöntemlerdeki puanların ortalamadan olan sapmalarının kareleri yerine, bir performans standartından (kesme puanı) olan sapmaların karelerini kullanmaktadır:

$$K^2(X,T) = \frac{\sigma_T^2 + (\mu_T - n_i C)^2}{\sigma_X^2 + (\mu_X - n_i C)^2}$$

Eşitlikte, σ_T^2 gerçek puanların varyansı, μ_T gerçek puanların ortalaması, σ_X^2 gözlenen puanların varyansı, μ_X gözlenen puanların ortalaması, n_i madde sayısı ve C de kesme puanıdır. Gerçek puanlar bilinemeyeceği için, kestirimler kullanılarak, iki form durumunda Livingston indeksi,

$$\hat{K}^2(X,T) = \frac{\hat{\rho}_{XX'}\hat{\sigma}_X\hat{\sigma}_X + (\hat{\mu}_X - n_iC)(\hat{\mu}_{X'} - n_iC)}{\sqrt{\left[\hat{\sigma}_X^2 + (\hat{\mu}_X - n_iC)^2\right]\left[\hat{\sigma}_{X'}^2 + (\hat{\mu}_{X'} - n_iC)^2\right]}}$$

şeklinde olur. Bu eşitlikte, $\hat{\rho}_{XX'}$ iki form arasındaki korelasyonun tahmini, $\hat{\sigma}_X$ ve $\hat{\sigma}_{X'}$ iki formun standart kaymalarının tahminleri, $\hat{\mu}_X$ ve $\hat{\mu}_{X'}$ iki formun ortalamalarının tahminleridir. Eğer bir tek form varsa, Livingston indeksi,

$$\hat{K}^2(X,T) = \frac{\sigma_X^2(KR-20) + (\mu_X - n_iC)^2}{\sigma_X^2 + (\mu_X - n_iC)^2}$$

şekline dönüşür. $\hat{K}^2(X,T)$, kesme puanı dağılımın ortalamasından uzaklaştıkça artar (Subkoviak, 1976). Livingston indeksinde karar tutarlığı test puanlarının benzerliğinden etkilenir: farklı madde güçlükleri farklı dağılımlara yol açar. Hambleton ve Novick (1973), Livingston katsayısının, ölçüt-dayanaklı testler için uygun olmadığı görüşündedirler: "Sorun, bir puanın kesme noktasından ne kadar saptığı değil, daha çok, bireyin gerçek performans düzeyinin kesme puanının altında ya da üstünde olup olmadığına karar vermedir. Ölçüt-dayanaklı testler için karelenmiş-hata kaybı (squared-error loss) değil, eşik-kayıp fonksiyonu (threshold loss) daha uygundur" (s. 168). Benzer eleştirileri Harris de (1972) yapmış ve "Livingston indeksinin diğer indekslere göre büyük değerler vermesinin, bireyin durumunu daha güvenilir olarak belirlediği anlamına gelmeyeceğini" belirtmiştir. Benzer eleştiriler, Brennan-Kane indeksi için de yöneltilmiştir. Eşik-kayıp fonksiyonunu savunanların eleştirilerine karşı Brennan-Kane (1977), "özden çok, biçimle uğraşıldığı"; Livingston (1972b) ise, "güvenirlik bir tek puanın değil, bir grup puanın özelliğidir; ölçüt-dayanaklı 'testlerin güvenirliklerinin' büyük olması, gerçek puanın ölçüt puanının üstüne ya da altına düşüp düşmediğini belirlemede daha güvenilir olduğu anlamına gelmez" şeklinde yanıtlar vermişlerdir.

Harris'in (1972) önerdiği katsayı ise k=2 durumunda varyans analizine dayanmaktadır (Subkoviak, 1976):

$$\mu_C^2 = \frac{SS_B}{SS_B + SS_W}$$

Simetrik dağılımlarda μ_C^2, $C = \mu$ olduğunda maksimum değerine ulaşır; bu da indeksin kararlı olmadığını göstermektedir.

Sınıflama kararlarının tutarlığını belirlemede çok basit bir hesaplamayla, bir formun iki kez uygulanmasına ya da klasik paralel testlere dayanan karar tutarlığı (decision-making consistency) indeksi Hambleton ve Novick (1973) tarafından önerilen $\hat{P}_0$'dir. $\hat{P}_0$, dağılım sayıltısı gerektirmeyen bir indekstir. $\hat{P}_0$, elle hesaplanabilme ve kolaylıkla yorumlanabilme açısından avantajlı olmasına rağmen, özellikle tek form iki kez uygulandığında, indeks değerinin olduğundan büyük sonuçlar (overestimation) vermesi ve küçük örneklem gruplarında büyük standart hatalar ortaya çıkması (Subkoviak, 1980) gibi dezavantajlara da sahiptir. Bu indeks, eşik-kayıp fonksiyonuna dayandığından, bir performans ölçütünü (kesme puanı) her iki uygulamada da aşan bireyler (mastery) doğru-pozitif (true-positive) sınıflanmış, aşamamış bireyler (nonmastery) doğru-negatif (true-negative) sınıflanmış olurlar. Yapılabilecek hatalı sınıflamalar ise, aslında başarılı bir bireyin başarısız (yanlış-negatif/false-negative) ve aslında başarısız bir bireyin de başarılı (yanlış-pozitif/false-positive) şeklinde sınıflanmalarıdır (Hambleton ve Novick, 1973).

Tablo 1. *Basit Uyum Katsayısının Simgesel Gösterimi*

		Uygulama II		
		Başarılı	Başarısız	
Uygulama I	Başarılı	Pa= .4	Pb= .1	a+b
	Başarısız	Pc= .2	Pd= .3	c+d
		a+c	b+d	a+b+c+d= N

Yukarıdaki açıklamalardan da anlaşılacağı gibi tutarlı sınıflama indeksi (basit uyum katsayısı);

$$\hat{P}_0 = (a+d)/N = \hat{p}_a + \hat{p}_d$$

şeklinde olur. Örneğin, belirlenen bir standartı (60 puan gibi) 100 kişiden 40'ı her iki uygulamada da geçmiş; 30'u da her iki uygulamada da başarısız olmuşsa,

$\hat{P}_0$= .4+.3= .7 sonucu elde edilir. Tablo 1'e göre I. uygulamada başarısız iken II. uygulamada başarısız olanların sayısı ise 20 kişidir; aynı şekilde, I. uygulamada başarılı iken II. uygulamada başarısız olanların sayısı da 10 kişi olarak verilmektedir. Bu bakımdan Pc ve Pb oranlarının toplamı, testin bireyleri tutarlı sınıflayamamasının oranını göstermektedir. Tutarsız sınıflama olasılığı (1-$\hat{P}_0$) ise .3 olur.

$\hat{P}_0$, bir oran olduğu için 0.00 ile 1.00 arasında yer alır; indeks değeri 1.00'e yaklaştıkça sınıflama kararlarının tutarlığını, 0.00'a yaklaştıkça da tutarsızlığını gösterir. Dikkat edilirse, $\hat{P}_0$, topyekün (overall) başarılı-başarısız sınıflama tutarlığını ölçmektedir. İndeks, seçilen kesme puanının dağılım içindeki yerine, test uzunluğuna ve puan değişkenliğine (variability) duyarlıdır (Berk, 1980). Binom dağılımının özelliklerinden dolayı, kesme puanı tektepedeğerli (unimodal) dağılımın uçlarına yakınsa yüksek $\hat{P}_0$ değerleri, dağılımın ortalamasına yakınsa düşük $\hat{P}_0$ değerleri ortaya çıkma eğilimi vardır (Subkoviak, 1980). Bu eğilim ikitepedeğerli (bimodal) dağılımlarda ortaya çıkmayabilir. Diğer yandan, testin madde sayısı arttıkça ve puan varyansı büyüdükçe $\hat{P}_0$ değerinin de artması beklenmesine rağmen, 10 maddeden az ve düşük puan varyansıyla bile .75 ve daha yüksek $\hat{P}_0$ değerleri elde etmek olasıdır (Berk, 1980). $\hat{P}_0$'nin değerinin ne olacağı, testle yapılacak kararların ciddiliğine bağlı olmakla birlikte, Subkoviak (1988), bu değerin .85 ve daha yukarı olması gerektiğini ileri sürmektedir.

Swaminathan, Hambleton ve Algina (1974), $\hat{P}_0$'nun yerine, tutarlı tüm sınıflamalara testin katkısını ölçen ve şansla beklenen uyumu da dikkate alan (uyumu şanstan arındıran) Cohen'in (1960) κ'sını (Kappa) önermişlerdir. Cohen'in Kappa'sı, aslında yargıcılar arası uyumu belirlemek amacıyla, "Olası Uyum" (contingency)

katsayısına (C) alternatif olarak geliştirilmiş bir katsayıdır. Swaminathan ve arkadaşları (1974), bu katsayının geçer-kalır (başarılı-başarısız) sınıflama kararlarında 2x2'lik "olası uyum" tablosuyla kullanılabileceğini belirtmişlerdir. κ'yı hesaplamak için geleneksel basit uyum katsayısı ($\hat{P}_0$) ve şansla beklenen uyumun hesaplanması gerekir. Şansla beklenen uyum, geleneksel uyum katsayısının en alt sınırını verir (Subkoviak, 1988):

$$P_{_{ans}} = [(a+b)(a+c)+(c+d)(b+d)]/N^2$$

$P_{_{ans}}$, iki uygulamanın birbirinden bağımsız olduğu sayıltısıyla, 2x2'lik tablonun marjinallerinden hesaplanır ve şansla beklenen tutarlı sınıflamaların oranını verir. Dolayısıyla $P_{_{ans}}$? .50 olacaktır. κ katsayısı ise,

$$\kappa = (\hat{P}_0 - P_{_{ans}})/(1 - P_{_{ans}})$$

şeklindedir. $(\hat{P}_0 - P_{_{ans}})$'tan anlaşılacağı gibi, Kappa, şansla beklenenden arınık, gözlenen tutarlı sınıflamaların oranını verir. Ancak κ'nın alt ve üst sınırlarında sorun vardır: 2x2'lik olası uyum tablosunda +1.00 değeri sadece, her iki formdaki ya da ölçümlerdeki marjinaller eşit olduğunda elde edilebilir (Hambleton ve ark., 1978; Berk, 1980). κ'nın alt sınırı ise –1.00'e gider (Cohen, 1960; Hambleton ve ark., 1978). κ'nın alt sınırının ne olduğu Hambleton ve arkadaşlarına (1978) göre önemli değildir; çünkü, "ölçüt-dayanaklı ölçme" bağlamında negatif değerler tutarsızlığı, yani güvenilmez kararları gösterir" (s. 21). Belirli marjinaller ve κ değerleri için örneklem büyüklüklerini gösteren tablolar Cantor (1996) tarafından geliştirilmiştir.

κ indeksini, şans düzeltmesinde önemli olan marjinal frekanslar, test uzunluğu, puan değişkenliği (Berk, 1980), kesme puanının yeri, grup heterojenliği ve bireyleri "geçer" kategorisine ayırmada kullanılan yöntem (Hambleton ve ark., 1978) etkiler. Kısa alt testler ve homojen gruplarda ranj sınırlılığından dolayı κ'nın değeri de düşer. Diğer yandan, κ'nın kesme puanının yerine olan duyarlığı $\hat{P}_0$'nin tam tersidir: Kesme puanı, grubun puan ortalamasına yakın olduğunda yüksek κ değerleri, dağılımın uçlarında ise düşük κ değerleri elde edilir (Huynh ve Saunders, 1980; Subkoviak, 1980). Örneğin, puan grubunun ortalaması 50 ise, 50 değerine yakın belirlenen kesme puanlarında, hesaplanan κ değerleri de büyük çıkmaktadır. Belirlenen kesme puanları 50 değerinden uzaklaşıp her iki uçtaki değerlere yakın olduğunda da (örneğin, 20 ve 80 kesme puanları gibi) κ değerleri de küçük çıkmaktadır. κ'nın yanlı bir kestirici olduğu da öne sürülmektedir (Harris ve Pearlman, 1977; aktaran; Berk, 1980). Öte yandan, tek uygulamaya dayalı yöntemler, $\hat{P}_0$'ye göre daha az doğru κ kestirimleri (yaklaşık %10 hata) verirler (Berk, 1980). κ'nın uygulamada kullanışlığı ve yorumu tam olarak açık değildir; κ'nın yukarıdaki sınırlılıklarından dolayı, κ'yı etkileyen tüm faktörler, κ'yı rapor ederken yoruma katkı açısından verilmelidir (Hambleton ve ark., 1978). Subkoviak'a (1988) göre, κ ? .35 ise test kullanışlıdır.

Bir testin aynı bireylere iki kez uygulanmasıyla sınıflama tutarlığını belirleme yöntemleri ekonomik olmadığından, tek uygulamaya dayanarak $\hat{P}_0$ ve κ'yı kestiren yöntemler geliştirilmiştir.

Tek uygulamaya dayanan ilk indeks Marshall ve Haertel (1976; aktaran, Subkoviak, 1976) tarafından ileri sürülmüştür. Marshall ve Haertel'in tek uygulamaya dayanan uyum katsayısının kestirimi, testin

tüm olası iki yarımlarındaki ortalama $\hat{P}_0$ 'yi hesaplamayı gerektirir. Bu katsayı, Cronbach Alfa'nın (α) benzeri olduğu için, karıştırılmaması amacıyla β ile gösterilir. Daha sonra tüm testin sınıflama tutarlığı kestirimi için Spearman-Brown yöntemiyle adım adım β bulunur. Bu katsayı da kesme puanının yerine göre düşme ya da yükselme göstermektedir.

Subkoviak'ın (1976) geliştirdiği indeks, paralel testler (tekler-çiftler şeklinde testi iki yarıya bölerek), maddelerin eşit güçlükte olduğu ve binom ya da katışık (compound) binom dağılım sayıltılarına dayanır. Bu indekste, bir kişinin tutarlı sınıflama olasılığı,

$$P_C^{(i)} = P(X_i \,?\, C, X_i' \,?\, C) + P(X_i < C, X_i' < C)$$

şeklindedir. Bu eşitlikte C kesme puanı, X_i ve X_i' bireyin eşdeğer yarılardaki puanlarını göstermektedir. N kişilik bir grup için uyum katsayısı ise, tek tek tutarlı sınıflama olasılıklarının ortalaması olarak tanımlanabilir:

$$P_C = \frac{\sum_{i=1}^{N} P_C^{(i)}}{N}$$

Bu oran, grup için beklenen tutarlı karar olasılığını temsil eder. X_i ve X_i''nin i kişisi için bağımsız dağıldığı sayıltısıyla (bu sayıltı tartışmaya açıktır), bir kişinin tutarlı sınıflama olasılığı;

$$P_C^{(i)} = P(X_i \,?\, C).P(X_i' \,?\, C) + P(X_i < C).P(X_i' < C)$$

şeklinde yazılabilir. Binom dağılımda, maddelerin 1-0 şeklinde puanlanması ve doğru cevap oranının maddelere göre sabit kaldığı sayıltısıyla;

$$P_C^{(i)} = [P(X_i \,?\, C)]^2 + [1 - P(X_i \,?\, C)]^2$$

olur. Binom dağılımda ise,

$$P(X_i \,?\, C) = \sum_{X_i=C}^{n} \binom{n}{X_i} P_i^{X_i} (1 - P_i)^{n - X_i}$$

şeklindedir. $\hat{P}_0$ geleneksel olarak X_i / n ile kestirilmesine rağmen, regresyon yöntemi daha doğru kestirimler verir:

$$\hat{P}_i = \alpha_{21/X}\left(\frac{X_i}{n}\right) + (1 - \alpha_{21/X})\left(\frac{\mu_X}{n}\right)$$

Bu eşitlikte, $\alpha_{21/X}$ maddeleri 1-0 şeklinde puanlanan testler için KR-21 içtutarlık katsayısı, n de testteki madde sayısıdır.

Subkoviak (1976), indeksinin değişik kesme puanlarındaki durumunu incelemiş ve indeks değerinin diğer indeksler gibi kesme puanının dağılım içindeki yerinden etkilendiğini göstermiştir. $\hat{P}_C$, $\hat{P}_0$ gibi bağımsız uygulamalara dayanmadığından, $\hat{P}_0$ 'nin değeri $\hat{P}_C$ 'den daha büyük olmaya eğilimlidir. Subkoviak (1976) ayrıca, sayıltıların karşılanmamasının $\hat{P}_C$ değerlerini etkileyeceğini ileri sürmektedir. Hambleton, ve arkadaşları (1978), Subkoviak indeksinin iki küçük sorununun, şans uyumunu içerdiği için kestirimin "şişeceği" ile ölçüt-dayanaklı testlerde tüm maddelerin eşit güçlükte varsayılmasının uygun olmaması olduğunu belirtmektedirler.

Huynh'un (1976) tek uygulamayla $\hat{P}_0$ ve κ 'yı kestirme yöntemi, beta-binom (ya da negatif hipergeometrik) dağılıma dayanmaktadır:

$$f(x) = \binom{n}{x} B(\alpha + x, n + \beta - x) / B(\alpha, \beta)$$

Bu eşitlikte, $\alpha = (-1 + 1/\alpha_{21})\mu$ ve $\beta = -\alpha + n/\alpha_{21} - n$ 'dir. Huynh ayrıca, iki

eşdeğer yarıdan (X ve Y) elde edilen puanların (x ve y) yerel-bağımsız olduğunu varsayarak, bivariate -beta binom dağılımın

$$f(x,y)=\frac{\binom{n}{x}\binom{n}{y}}{B(\alpha,\beta)}B(\alpha+x+y,2n+\beta-x-y)$$

şekline dönüşeceğini öngörür.

P_{00} iki formda da "kalır" kategorisindeki oran; P_{11}, iki formda da "geçer" kategorisindeki oran; P_0 ve P_1 marjinallerin oranlarını göstermek için kullanılırsa, geleneksel $\hat{P}_0 = P_{00} + P_{11}$; $\hat{P}_{cans} = P_0^2 + P_1^2$ olur. Bu durumda, $P_{11} = \sum_{x,y=c}^{n} f(x,y)$ ve $P_1 = \sum_{x=c}^{n} f(x)$ olarak, yukarıda verilen beta-binom fonksiyonları yardımıyla hesaplanabilir. Buradan hareketle, kesme puanları n'e yakın olduğunda,

$$\kappa = (P_{11} - P_1^2)/(P_1 - P_1^2);$$

küçük kesme puanlarında ise,

$\kappa = (P_{00} - P_0^2)/(P_0 - P_0^2)$ ile κ bulunur.

Huynh (1976), n büyük olduğunda, yukarıdaki hesaplamaları azaltan arcsinüs dönüştürmesi önermiştir ($x' = \sin^{-1}\sqrt{x/n}$). İşlemin adımları aşağıdaki şekildedir:

a) $\overline{X}$'yı μ ve S'yi σ ile değiştirerek, $\mu_{x'} = \sin^{-1}\sqrt{\mu/k}$ ve

$\sigma_{x'} = [(\alpha_{21}+1)/(\alpha+k)]^{1/2}$ hesaplanır. Eşitlikte, $\alpha = (-1+1/\alpha_{21})\mu$ ve k madde sayısıdır.

b) $\rho = \alpha_{21}[(k-1)/(k-\alpha_{21})]^{1/2}$ bulunur.

c) $x' \; ? \; c'$ koşulu için (c, kesme puanı), $c' = \sin^{-1}\sqrt{(c-.5)/k}$ bulunduktan sonra,

d) $z = (c' - \mu_{x'})/\sigma_{x'}$ hesaplanır.

e) Huynh'daki (1976) tablodan z ve ρ'yu girerek P_{00} değeri, Gupta'nın (1963) tablolarından da P_0 bulunur.

f) Bulunan değerler, $\kappa = (P_{00} - P_0^2)/(P_0 - P_0^2)$ eşitliğinde yerine konarak κ kestirilir.

Huynh'un κ'yı hesaplama yöntemi, yoğun işlemler gerektirdiği için eleştirilmiştir (Hambleton ve ark., 1978).

Daha önce de değinildiği gibi, "karelenmiş-hata kaybı"na dayanan (Berk, 1980) ve Livingston indeksine benzer indeks, Brennan ve Kane (1977) tarafından önerilmiştir. Berk'e (1980) göre, iki indeks de (a) dağılım sayıltısı gerektirmemesi, (b) şanstan arınık uyumu dikkate almaması, (c) kesme puanının yerinin dağılımın ortalamasından uzaklaşmasına bağlı olarak indeks değerlerinin artması, (d) indeks değerlerinin ölçmenin standart hatasından bağımsız değişebilmesi, (e) indeks değerlerinin, test uzunluğu arttıkça yükselmesi gibi yönlerden benzer olmasına rağmen, test formlarına ilişkin sayıltıları ve bu sayıltılara dayanan hata varyansı ya da karelenmiş hata kaybı tanımları açısından farklılık gösterirler. Brennan ve Kane'in indeksi, aslında Cronbach, Gleser, Nanda ve Rajaratnam'ın (1972) genellenebilirlik katsayısından çeşitli sayıltılarla ölçüt-dayanaklı testler için türetilmiş bir indekstir.

Çeşitli sayıltılarla türetilen (1-0 biçiminde ikili puanlanan maddeli testler için) indeks;

$$\hat{M}(C) = 1 - \frac{1}{n_I - 1} \frac{X_{PI}(1 - X_{PI}) - S^2(X_{PI})}{(X_{PI} - C)^2 + S^2(X_{PI})}$$

şeklindedir (çoklu puanlanan maddeli testler için olan eşitlik burada verilmemiştir). Eşitlikte, X_{PI}, n maddelik örneklemde ortalama p kişisinin gözlenen puanını ve C de kesme puanını göstermektedir. Brennan ve Kane (1977) kendi indekslerine, geleneksel güvenirlik ve ölçüt-dayanaklı sınıflama tutarlığından farklı oluşu nedeniyle "güvenilebilirlik" (dependability) adını vermişlerdir. Kesme puanı, örneklemin puan ortalamasına eşit olduğunda Livingston indeksi ($\hat{K}(X,T)$) KR-20'ye, Brennan ve Kane indeksi de KR-21'e eşdeğer olur. Ayrıca, $\hat{K}(X,T)$, $\hat{M}(C)$'den daha büyük kestirimler verir (Crocker ve Algina, 1986). Livingston katsayısı gibi, Brennan ve Kane indeksi de "karelenmiş hata kayıp" fonksiyonuna dayanmaktadır. Brennan ve Kane, kendilerini eleştiren "eşik-kayıp" fonksiyoncularına (Hambleton ve Novick, 1973; Huynh, 1976; Subkoviak, 1976) karşı, "eşik-kayıp fonksiyonu, tüm yanlış sınıflamaların eşit ciddilikte olduğunu varsayar; oysa bir ölçüt-dayanaklı test için yanlış sınıflamaların eşit ciddilikte olduğu açık değildir. Geçerler ve kalırlar arasında kesin bir ayırım yapmak zordur ve kesme puanları kaçınılmaz bir şekilde keyfidir. Bir alandaki uzmanlık (geçer) için %90 kesme puanı tanımlanmışsa, %89 evren puanına sahip birinin %20 evren puanına sahip biri gibi yanlış sınıflandığı sayıltısını iddia etmek genellikle güçtür" (s. 286) şeklinde karşılık vermektedirler.

Peng ve Subkoviak (1980) ise, Huynh'un işlemine basit bir yaklaştırma önermişlerdir: Eş normal dağılımlara ve "birleşik iki yönlü dağılıma" (joint bivariate distribution) sahip testler birbirinin yerine kullanılabileceğinden, X'in $n_i C$'den daha az olma olasılığı ($P_{0.} = \Pr(X < n_i C)$ şu şekilde hesaplanabilir:

$$P_{0.} = \Pr\ z < \frac{n_i C - \mu_X}{\sigma_X} ?$$

P_{00} ise, her iki formdaki puanların $n_i C$'den daha az olma olasılığını verir. İki değişken arasındaki korelasyon bilindiğinde, bu olasılıklar Gupta'nın (1963) hazırladığı dağılım tablolarından bulunabilir. Huynh (1976) bir tek uygulamayla elde edilen KR-21 ile bu olasılıkların bulunabildiği tablolar vermiştir.

Subkoviak (1988), standart normal dağılımdan yararlanarak ($|z| = \frac{(C - .5 - \bar{X})}{S}$) testin ortalaması, standart kayması ve KR-20 ya da KR-21 gibi bir güvenirlik katsayısı bilindiğinde $\hat{P}_0$ ve κ'nın yaklaşık değerlerinin bulunabileceği tablolar geliştirmiştir.

Breyer ve Lewis'in (1994) "basitleştirilmiş yöntem" diye ifade ettikleri ve Alfa gibi bir içtutarlık katsayısı gerektirmeyen indekslerinin hesaplanması aşama aşama aşağıdaki şekildedir:

1. Tüm test, her birinin ayrı ayrı kesme puanları bulunan, iki yarının kesme puanları toplamı tüm testin kesme puanını verecek şekilde iki yarıya bölünür.

2. İki yarı test için de başarısız olanların ortalama oranı,

$$P_{k,yar?} = \frac{2X_{11} + X_{12} + X_{21}}{2N}$$

formülüyle hesaplanır. Eşitlikte X_{11} iki yarı testteki başarısız bireylerin frekansını; X_{12} 1.

yarı testte başarısız, 2. yarı testte başarılı olan bireylerin frekansını; X_{21} 1. yarı testte başarılı, 2. yarı testte başarısız olan bireylerin frekansını göstermektedir.

3. $P_{k,yar?}$'a karşılık gelen z-puanı ($z_{yar?}$) ilgili tablolardan bulunur.

4. Her iki yarı testte de başarısız olan bireylerin oranı;

$P_{kk,yar?} = \frac{X_{11}}{N}$ formülüyle hesaplanır.

5. $z_{yar?}$ ve $P_{kk,yar?}$ değerleri kullanılarak, iki yarı test arasındaki korelasyon, tetrakorik korelasyonun özel bir hali (Huynh, 1976; Gupta, 1963) ile bulunur ($r_{yar?}$).

6. $r_{yar?}$ ve Spearman-Brown formülü kullanılarak, tüm testin korelasyonu ($r_{tüm}$) bulunur.

7. İki yarı testin toplamının standart kayması;

$S_{tüm} = \sqrt{1+1+2r_{yar?}}$ formülüyle hesaplanır.

8. $S_{tüm}$ ve $z_{yar?}$ kullanılarak, tüm testin standart puanı bulunur:

$z_{tüm} = \frac{2z_{yar?}}{S_{tüm}}$

9. Tablolardan $z_{tüm}$'e karşılık gelen $P_{k,tüm}$ bulunur.

10. $z_{tüm}$ ve $r_{tüm}$ yardımıyla $P_{kk,tüm}$ hesaplanır.

11. $P_{k,tüm}$ ve $P_{kk,tüm}$ ile de tüm testin tutarlık indeksi,

$P_{CC,tüm} = 1 - 2(P_{k,tüm} - P_{kk,tüm})$ bulunur.

"Basit" geçer-kalır indeksinin öyküsünün hiç de basit olmadığı yukarıdaki açıklamalardan kolaylıkla görülebilmektedir. Breyer ve Lewis (1994), bu işlemler için "logaritmik olasılık (log-odds)" tabloları da hazırlamışlardır. Breyer ve Lewis, indekslerinin, maddeleri sadece 1-0 şeklinde değil, her türlü puanlanan testlere ve ayrıca içtutarlık katsayısı gerektirmediği için benzeşik (homojen) olmayan testlere de uygulanabileceğini ve bu nedenle de kullanışlı bir indeks olduğunu belirtmektedirler.

Livingston ve Lewis'in (1995) indeksi ise, yine iki yarıya bölme yöntemine dayanmakla birlikte, testin maksimum ve minimum olası puanları ve puan dağılımlarına ilişkin bilgiden hareketle, etkili test uzunluğunun gerçek puan dağılımının, her bir gerçek puan düzeyinde bireylerin testin diğer formundaki sınıflamaların koşullu dağılımının ve gerçek puanlara dayalı sınıflamaların "birleşik (joint) dağılımlarının" kestirimlerini gerektirmektedir. Maddelerinin 1-0 şeklinde puanlandığı bir test için orantısal puan,

$$P = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

bireyin gerçek orantısal puanı ise,

$$T_P = \frac{E(X) - X_{\min}}{X_{\max} - X_{\min}}$$

ile hesaplanır.

$X_{\min}$'den $X_{\max}$'a puanlar 0 ile n arasındaki yeni bir ölçeğe,

$$X' = np = n\frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

ile dönüştürülür.

Livingston ve Lewis, işlemlerinin adımlarını şöyle özetlemektedir:

"1. Geniş bir gruba uygulanan bir testi, madde kapsamı ve formatı benzer olacak şekilde iki yarıya bölün.

2. Her birey için iki yarıdaki puanları bulun; her iki yarıdaki puanların dağılımını ve iki yarı test arasındaki korelasyonu hesaplayın.

3. Her yarı test için, tüm testin gerçek kesme puanlarıyla oldukça aynı yüzdelik sırada olacak şekilde, kesme puanları seçin ve her kesme puanındaki 2x2 "olası uyum" (contingency) tablosundaki sınıflamaları bulun.

4. İki yarı test puanları arasındaki korelasyonu güvenirlik katsayısı gibi kullanarak, yarı testlerden birindeki puanlara kestirim yöntemini uygulayın; bulduğunuz sonuçları Adım 3'teki 2x2'lik olası uyum tablosunun sonuçlarıyla karşılaştırın."

Livingston ve Lewis, indekslerinin, testin puan ranjından ve kesme puanlarının yüzdelik sırasından etkilendiğini aynı çalışmada görgül olarak sınadıkları 7 test üzerinde göstermişlerdir.

Literatüre bakıldığında, yukarıdaki indekslerden en çok kullanılan ve üzerinde çalışılanlarının $\hat{P}_0$ ve κ katsayıları olduğu görülmektedir. Kappa ve Ağırlıklı Kappa'nın (Cohen, 1968) "yokluk hipotezi" durumunda test edilmesi için çalışmalar da yapılmıştır (Rae, 1997).

Sınıflama tutarlığı indeksleri, Cohen'in (1960) yargıcılar arası uyumun bir göstergesi olarak önerdiği κ 'ya dayandığından, "uyum (agreement) katsayıları" olarak da anılmalarına rağmen; sınıflama tutarlığı indeksleriyle uyum indeksleri arasında bir ayırım yapmak yararlı olacaktır. İki yargıcının aynı bireyleri derecelendirdiği ya da aynı bireylerin herhangi bir işlemden önce ve sonra aynı yöntemle derecelendirildiği durumlarda, yapılan iki derecelendirme arasında (genellikle sıralama türü veriler üzerinde) uyumun derecesi bilinmek istenebilir. Bu gibi durumlarda uyumun derecesini veren istatistikler ise, geleneksel olarak Spearman'ın rho'su, Kendall'ın tau'su ve Goodman ve Kruskal'ın gamma'sıdır (Cicchetti, 1972). Cicchetti'nin (1972; 1977) C istatistiği ve Cohen'in (1960) κ 'sı da bu uyum katsayıları içindedir.

Yukarıda ele alınan sınıflama tutarlığı indekslerini, isminden de anlaşılacağı gibi, uyum indekslerinden ayırmak gerekecektir: Uyum katsayıları daha çok sıralama türü veriler için uygun iken, sınıflama tutarlığı indeksleri, verilerin niteliği ne olursa olsun, testlerin bireyleri sınıflama tutarlığını (binom dağılıma uygun) yansıtmaktadır.

Ölçüt-Dayanaklı Yaklaşım İle İlgili Genel Sorunlar

Ölçme-Değerlendirme ve Güvenirlik-Geçerlik Açısından Ölçüt-Dayanaklı ve Norm-Dayanaklı Ayırımı

Literatürde ölçme, çok çeşitli biçimlerde tanımlanmasına rağmen, genel olarak "nesneleri ve özellikleri sayılarla ayırma işlemi" (Nunnally, 1970) şeklinde tanımlanabilir. Biraz daha açılacak olursa, bu özellikleri belirli kurallara uyarak niceleme işlemi de denilebilir. Örneğin, Can'ın boyunun, metre denilen standart bir ölçme aracıyla belirlenip, cm. cinsinden 110 cm. olarak ifade edilmesi bir ölçme işlemidir. Değerlendirme ise, bu ölçme sonucunun bir ölçüte vurularak, ölçülen nitelik hakkında bir yargıya varma sürecidir (Turgut, 1992). Bu anlamda, bir ölçme sonucu tek başına bir şey ifade etmez; onun anlamlandırılması ancak bir değerlendirme süreciyle olur. Örneğimizdeki

Can'ın "uzun mu, kısa mı" olduğu konusunda bir yargıda bulunulabilmesi için bir ölçüt gereklidir. Eğer Can yeni doğmuş bir bebekse, "anormal" derecede uzun; yok eğer bir yetişkinse, çok kısa olarak değerlendirilecektir. Aynı şekilde, "Can'ın matematik puanı 60" demek de bir şey ifade etmez: Eğer değerlendirme 60 üzerinden ise farklı, 100 üzerinden ise daha farklı bir yargıya varılır; yine, aynı 60 puan, 50 kesme puanına göre farklı, 65 kesme puanına göre farklı değerlendirilir. Öte yandan, Can'ın 60 puanı içinde bulunduğu grubun ortalama ve standart sapmasına göre standart puan cinsinden ele alınırsa, Can'ın grup içindeki yeri ile ilgileniliyor demektir ve bu da grubun normlarına göre yorumlamayı beraberinde getirir.

Bu açıdan bakıldığında, ölçüt-dayanaklı ve norm-dayanaklı ayırımı, ölçme sonuçlarının bir ölçüte vurulma ve buna göre bir karar verme aşamasında ortaya çıkan değerlendirme ve yorumlama ayırımına dayanmaktadır. Bir önceki bölümde sunulan indeksler incelendiğinde, bu indekslerin, test (ölçme) sonuçlarının ne kadar hatasız-güvenilir ölçme yaptıklarıyla değil, test sonuçlarını bir ölçüte vurarak verilecek kararların tutarlı olup olmadığı ile ilgili oldukları rahatlıkla görülebilir. Berk'e (1980) göre de, ölçüt-dayanaklı testler için salık verilen indeksleri tanımlamak için "güvenirlik katsayısı" yerine "uyum indeksi" terimini kullanmak daha uygundur. Bu bakımdan, bu indeksler için yeni bir kavramlaştırmaya gitmek gerekmektedir. Hambleton ve Novick (1973) ile Hambleton ve arkadaşlarında (1978), sözünü ettiğimiz ayırımın yapaylığına ilişkin izler görünmekle birlikte, bu ayırımı açıkça vurgulayan bir yazıya literatürde rastlanmamıştır. Hambleton ve arkadaşlarında (1978) ve Popham ve Husek'de (1969) test yapımı açısından ölçüt-dayanaklı ile norm-dayanaklı testler arasında pek de farklılık bulunmadığı belirtilmektedir. Ölçüt-dayanaklı ve norm-dayanaklı (aslında *değerlendirmeler* olmasına rağmen, "testler", "ölçekler" gibi yanlış bir kavramlaştırma, görüldüğü gibi devam etmektedir) ayırımında önemli farklılığın ölçütte yattığı görülmektedir: Bir ölçme sonucu (isterse aynı testten elde edilsin), grubun normları ölçüt alınarak *değerlendiriliyorsa* norm-dayanaklı; bir kesme puanı gibi gruptan bağımsız bir ölçüt alınarak *değerlendiriliyorsa* ölçüt-dayanaklı (kesme puanı-dayanaklı demek daha uygun); testten alınabilecek maksimum puan ölçüt alınarak *değerlendiriliyorsa* alan-dayanaklı olarak adlandırılmaktadır. Bu bakımdan, "ölçüt-dayanaklı ya da norm-dayanaklı *testlerin güvenirliğ*inden çok, *ölçme sonuçlarının değerlendirilmesi* şeklinde bir ifadelendirme daha doğru görünmektedir.

Klasik güvenirlik tanımının, bir ölçme aracının hatadan arındırılmış ve tutarlı (işlemsel olarak tekrarlı ölçmeler yapılmışsa, aynı sonucun alınması) ölçüm yapıp yapmadığı ile ilgili olduğu bilinmektedir (Guilford, 1954; Gulliksen, 1967; Lord ve Novick, 1968; Crocker ve Algina, 1986). Geçerlik ise, test puanlarından yapılacak vardamaların (inferences) kullanışlığını, yani testin geliştirilme amacına uygunluk derecesini (APA, AERA ve NCME, 1985) ifade eder. Testler bu bakımdan genelde bireyler hakkında çıkarsamalarda bulunmak, kararlar vermek amacıyla kullanılırlar: Testin uygulandığı birey geçecek mi-kalacak mı, işe ya da okula alınacak mı-alınmayacak mı,... gibi. Dolayısıyla, karar aşamasında ya da değerlendirme aşamasında, ölçme sonuçları bir ölçüte göre çoğunlukla sınıflama düzeyine indirgenmiş olmaktadır. Bu bakımdan, bir önceki kısımda ele alınan indekslerin, güvenirlik ya da uyumdan çok, "geçerlik" indeksleri olduğunu ileri sürmek pek yanlış olmasa gerektir. Çünkü söz konusu indeksler,

testten elde edilen puanların hatasız ve tutarlı olup olmadığıyla değil, o puanların bir ölçüte dayanarak verileceği kararların ne kadar tutarlı olduğu (bir başka deyişle, amaca ne kadar ulaşıldığı) ile ilgilidirler. Bu kararlar da, ölçme sonuçlarından yapılacak çıkarsamalarla, yani ölçme sonuçlarının geçerliği ile ilgilidir. Tek cümleyle özetlemek gerekirse, ölçüt-dayanaklı test yoktur, ölçüt-dayanaklı değerlendirme vardır; bir test (aslında test sonuçları) amaca bağlı olarak ölçüt-dayanaklı olarak da, norm-dayanaklı olarak da kullanılabilir.

Ölçüt-dayanaklı testlerin geçerliklerine ilişkin hiçbir çalışma yapılmaması ve bu konuda hiçbir yazıya rastlanmaması da yukarıdaki yargıyı destekliyor görünmektedir. Popham ve Husek (1969), norm-dayanaklı testlerden farklı olduğunu ileri sürdükleri ölçüt-dayanaklı testlerin geçerlikleri konusunda klasik kapsam geçerliği (content validity) ile yapı geçerliğini (construct validity); Hambleton ve Novick de (1973), klasik kapsam geçerliği ile ölçüt-bağıntılı geçerliği (criterion-related validity) önermektedirler. Livingston ve Lewis (1995) ise, biraz farklı bir şekilde, alternatif formlar temelinde yapılacak sınıflamalar ile gerçekten uygulanan forma dayanarak yapılan sınıflamalar arasındaki uyum olarak (güvenirlik yerine) "karar tutarlığı" nı (decision consistency); test edilenlerin gerçek puanları temelinde yapılacak sınıflamalar ile gerçekten uygulanacak forma dayanarak yapılan sınıflamalar arasındaki uyum olarak da (geçerlik yerine) "karar doğruluğu"nu (decision accuracy) önermektedirler. Subkoviak (1988) ise, Kappa ve uyum katsayılarını "iki test uygulamasında 'geçer-kalır' sınıflamaların tutarlığının bir ölçüsü" olarak ele almasına ve bu katsayıların geleneksel güvenirlik katsayılarından farklı yorumlar gerektirmesini belirtmesine rağmen, iki indeksi "geçme (mastery) testlerinin güvenirlik indeksleri" şeklinde tanımlamaktadır. Brennan ve Kane (1977), "geçme testi, tek kesme puanına sahip alan-dayanaklı test olarak tanımlanabilir" demekte ve kendi indeksleri için "güvenilebilirlik (dependability)" terimini kullanmayı yeğlemektedirler. Huynh (1976), ilk yazılarından birinde, hem ölçüt-dayanaklı hem alan-dayanaklı testler, hem de kararların tutarlığı ile güvenirlik terimlerini birbiri yerine kullanmaktadır. Huynh ve Saunders (1980), "geçmeyi ölçmede, güvenirliği genellikle, tekrarlı ölçmelerdeki 'geçer-kalır' kararların tutarlığı olarak gördüğünü" belirtmektedir. Görüldüğü gibi, çoğu makalede "sınıflama kararlarının tutarlığı" ile "güvenirlik" eşdeğer şekilde ele alınmakta, geçerliğe ilişkin de, klasik yolların (kaçınılmaz biçimde) önerilmesinden başka bir şey yapılmamaktadır. Bu durum, tamamen bir kavram karmaşasının sonucudur.

Kesme Puanı ile İlgili Sorunlar

Kesme puanı ile ilgili sorunlar, kesme puanının doğasından ve kesme puanının indekslerde kullanımından kaynaklanan sorunlar olarak iki grupta toplanabilir.

Kesme puanının doğasından kaynaklanan sorunlar

Kesme puanı hangi yöntemle (Nedelsky, 1954; Angoff, 1971; Ebel, 1972; Glass, 1978; Mills, 1983) ve hangi kesinlikte belirlenirse belirlensin, bir keyfilik (arbitrary) taşımaktadır (Hambleton ve Novick, 1973; Huynh, 1976). Öte yandan, politik-kurumsal kararlar ve (eğer test, işe ya da okula giriş için kullanılıyorsa) başvuran aday sayısı ile alınacak aday sayısı arasındaki fark da (Vos, 1997) aynı ölçme aracına ilişkin kesme puanlarının farklı olmasına yol açabilmektedir. Örneğin, aynı ölçme aracıyla test edilmelerine karşın, farklı bölgelerdeki Anadolu Liseleri farklı kesme puanlarıyla öğrenci almaktadırlar. Bu

bağlamda, önceki kısımda ele alınan indekslerin "keyfi" ve oynak bir kesme puanına dayanması ve bu indekslerin "güvenirlik" olarak adlandırılması başlı başına bir sorundur.

Dwyer'ın (1996), kesme puanlarının yapısı ve buna bağlı sorunlara ilişkin görüşleri şu şekilde özetlenebilir:

a) Eğitim ve psikolojide, açık veya örtük bir şekilde kullanılan kesme puanlarını oluşturmadaki yöntemlerin tümü bir yargıya bağımlıdır. Kesme puanlarını oluştururken görüş ve deneyimlerine başvurulan yargıcıların tümünün yargıcılık niteliğinin aynı olmaması ve bazılarının bu konudaki eğitimlerinin yetersiz olması ile küçük yargıcı örneklemlerinin kullanılması, kesme puanlarının değeri konusunda sorunlar çıkarmaktadır.

b) Testlerin çoğundaki puanlar, bireysel farkları, bir anlamda bireyleri birbirlerine göre sıralayarak yansıtırlar. Bu nedenle kesme puanının hemen üstündeki bir birey, kesme puanının hemen altındaki bir bireyden çok farklı değildir. Kesme puanlarına ilişkin sorunlar, bireylerin hatalı sınıflanmalarına yol açmaktadır. Bu hatalı sınıflamanın birey ve toplum açısından maliyeti, onarılamayacak şekilde ciddi olabilir.

c) Sürekli boyutta yer alan bilgi, beceri ve yetenek, kesme puanı kullanıldığında sınıflama düzeyine indirgenmekte, bu da bilgi kaybına neden olmaktadır. Örneğin, 100 üzerinden 60 kesme puanı kabul edilmişse, 59 puan alan birey ile 10 puan alan birey "kalır", 60 ile 90 puan alan bireyler de "geçer" kategorisine; oysa 59 ile 60 puan alan iki birey farklı kategoriler içine sokulmaktadır. Bu bakımdan, bireyler arasındaki derece farkları dikkate alınmamaktadır.

Kesme puanını ölçüt alarak, bireylerin tutarlı sınıflanıp sınıflanmadığına ilişkin geliştirilen indeksler, yukarıdaki durum göz önüne alınarak değerlendirildiğinde ortaya karamsar bir tablo çıkmaktadır. Testin kendi özellikleri yerine, "keyfi" bir kesme puanına dayanarak bir testin "güvenirliğinin" belirlendiğini iddia etmek ise güç olsa gerektir. Her şeyden önce, bir ölçme aracının orijinal güvenirliği yeterli olmalıdır ki, bu ölçme aracından elde edilen sonuçlara dayanarak yapılan sınıflamanın da doğruluğu sınanabilsin. Bireyler hakkında çok ciddi kararlar (geçer-kalır, kabul-red, kendilik değeri düşük-yüksek gibi) vermede bir standart ölçüt olarak alınan kesme puanının "oynaklığı", "keyfiliği" bir yana, öncelikle ölçme aracının klasik anlamda güvenilir olduğunun gösterilmesi gerekir. Bu bakımdan, bu "keyfi" kesme noktalarına göre bireylerin ne kadar tutarlı sınıflandığını belirlemek için geliştirilen indeksleri "güvenirlik indeksleri" şeklinde tanımlamak doğru bir kavramlaştırma

İndekslerin kesme puanına dayanmasından kaynaklanan sorunlar

Ele alınan indekslerin tümü, kesme puanının puan dağılımındaki pozisyonuna duyarlıdır (Berk, 1980). Kappa katsayısı dışında, tümü, kesme puanı grubun ortalamasına ya da ortancasına yaklaştıkça düşük, dağılımın uçlarına gidildikçe de yüksek değerler almaktadır. κ da ise bu durum tam tersidir.

Tablo 2 ve 3'ün incelenmesinden de görülebileceği gibi, binom dağılımının özelliğinden dolayı, dağılımın ortalarında binom dağılımının standart hatası en büyük değeri aldığından indeks değerleri de, kesme puanının dağılımdaki yerine göre değişmektedir.

Kesme puanının pozisyonundan başka, testin uzunluğu (Huynh, 1978; Crocker ve Algina, 1986), iki formun puan dağılımlarının

benzerliği ve puan değişkenliği de (Huynh, 1976) indeks değerlerini etkilemesine rağmen, bu etkiler klasik güvenirlikteki etkilerine benzerdir.

Ölçüt-Dayanaklı Yaklaşımda Geçen İlk 30 Yılın (1963-1993) Değerlendirilmesi

Ölçüt-dayanaklı değerlendirme alanındaki ilk 30 yılın ele alındığı bir süreli yayında (Educational Measurement: Issues and Practice, 1994), bu alana önemli katkılar yapmış araştırmacıların özeleştirileri yer almaktadır.

- Glaser (1994a), ölçüt-dayanaklı testler fikrini o zamanlar gündemde olan davranışçı yaklaşımın etkisine bağlamakta ve şimdilerde ise bilişsel açıklamaların ağır basmaya başladığını belirtmektedir. Davranışçı akımın etkisiyle bilişsel yanın ihmal edildiğini ileri süren Glaser (1994b), ölçüt-dayanaklı "performans" kuramının temelinde yatan bilişsel yapıları; a) yapılandırılmış-ilkeli bilgi (structured; principled knowledge), b) işlem yolu belirlenmiş bilgi (proceduralized knowledge), c) becerili bellek ve otomatiklik (skilled memory and automaticity), d) etkili problem temsili (effective problem representation) ve e) kendi kendini düzenleyici

Tablo 2. *Kesme Puanındaki Değişimin* $\hat{P}_0$, $\hat{K}^2(X,T)$ *ve* $\hat{M}(C)$ *Üzerindeki Etkisi*

	Kesme Puanı				
İndeks	.1	.3	.5	.7	.9
$\hat{M}(C)$	.94	.88	.76	.76	.88
$\hat{K}^2$ (X,T)	.96	.92	.83	.83	.90
$\hat{P}_0$	1.00	.90	.70	.70	.90

Crocker ve Algina (1986), s: 207'den kısaltılarak alınmıştır. Tabloda puan dağılımında kesme puanlarının bulundukları yüzdeliklere göre, söz konusu üç indeksin aldığı değerler gösterilmektedir. Örneğin, kesme puanı, puan dağılımının .5 ile .7 dilimi arasında yer aldığında üç indeks de düşük, .1 ile .9 dilimlerine çekildiğinde ise üç indeks de büyük değerler almaktadır.

Tablo 3. *Kesme Puanının Bir Fonksiyonu Olarak* κ *'nın Değişimi*

	Kesme Puanı					
Veriler	9	11	13	15	17	19
I	.277	.339	.361	.331	.243	.107
II	.577	.616	.641	.650	.637	.572
III	.738	.745	.741	.727	.693	.606

Huynh (1976), s: 260 ve Huynh (1978), s: 323'den kısaltılarak alınmıştır. Tabloda, üç ayrı veri grubu, aynı kesme puanlarına göre birarada verilmiştir. 9 ile 19 arasındaki çeşitli kesme puanlarında κ 'nın değişimine bakıldığında, I. veri grubunun ortalaması olan 13 kesme puanı olarak alındığında κ 'nın en büyük değerine ulaştığı, kesme puanı puan dağılımının uçları olan 9 ile 19'a doğru çekildiğinde κ 'nın düştüğü görülmektedir.

beceriler (self-regulatory skills) şeklinde özetleyerek yeni bir tartışma başlatmaya çalışmaktadır. Ölçüt-dayanaklı ölçmeyi, "performans-dayanaklı ölçme" olarak değiştirmeyi uygun gören Glaser (1994b), gelecekte, geleneksel psikometrik kavramların değişeceğini, bu değişiklik de biliş, öğrenme ve yeterlik kavramlarının önemli etkileri olacağını ve belki de, ölçüt-dayanaklı test kavramının özünün bu çerçevede değişeceğini, ölçüt-dayanaklı ölçme teriminin yeni terimlerle yer değiştireceğini ileri sürmektedir.

- Ölçüt-dayanaklı ölçmeyi, "puana-dayanan çıkarsamalar (score-based inferences)" terimiyle değiştiren Popham (1994), bu alanın psikometriye önemli katkılar sağlamakla birlikte, bugün artık yüksek düzeyde düşünme becerileri ve daha geniş bilgi alanlarını dikkate alan testlerin gerektiği bir durumda, telefon rehberi gibi belirtkelerin gerekli olmadığını belirtmekte ve artık, bilişsel yeteneğe ilişkin tutarlı çıkarsamalar yapmaya olanak sağlayan değişik türde test maddeleri hazırlanması gerektiğini ileri sürmektedir.

Linn (1994) ise, Glaser'ın ilk önerilerinin şimdiki performansa-dayalı değerlendirme hareketiyle tutarlı olduğunu belirtmektedir. Linn'e göre, maalesef ölçüt-dayanaklı ölçme terimi temel anlamından uzaklaştırılmış, yanlış kavramlaştırmaya yol açılmıştır. Bunların örnekleri, a) norm-dayanaklı ile ölçüt-dayanaklı yorumların tek bir ölçü için birlikte varolamayacağı, b) ölçüt-dayanaklı ölçmenin muhakkak kesme puanı içerdiği yorumu, c) ölçüt-dayanaklı ve alan-dayanaklı ölçmeyi eşgörme, d) ölçüt-dayanaklı ölçmeyi, çoğu ikili, dar davranışçı kavramlaştırma ile tipik hiyerarşik beceri ve davranışlarla sınırlama şeklinde sıralanabilir. Linn, "A öğrencisinin problemleri B öğrencisinden daha hızlı çözüp çözmediği ya da A öğrencisinin ülkedeki öğrencilerin %90'ından daha fazla soruyu doğru cevaplayıp cevaplamadığı" şeklinde bağıl yorumların ölçüt-dayanaklı yorumlara eklenebileceği önerisini getirmektedir. Dikkat edilirse, bu öneri, gruba bağlı değerlendirmelerin (özellikle sıralama-derecelemeye dayalı bilgilerin) gözönüne alınması gerektiğini vurgulamakta, bir anlamda da testler bağlamında "kesin ölçüt-dayanaklı/norm-dayanaklı ayırımı gereksizliğinin" dışavurumuna işaret etmektedir. Aynı makalesinde Linn, ölçüt-dayanaklı testler için kesme puanlarının aslında temel bir kavram olmadığını, ölçütün daha çok bir yapı (construct) olduğunu, bugüne kadar ölçüt-dayanaklı ölçmeye damgasını vuran davranışçılık yerine bilişsel psikolojinin zihinsel şemalar, aktif yapı temsilleri gibi konulara yönelinmesini, "özgül değerlendirme" kavramının daha uygun olacağını belirtmektedir.

- Millman'a (1994) göre, yerine getirilemeyen savlardan biri, ölçüt-dayanaklı testlerin bir öğrencinin "ne yapabileceği ve yapamayacağına ilişkin geçerli çıkarsamalara olanak sağlayacağı" idi ki, bu da yanlış bir sayıltıya dayanıyordu ve bu bakımdan başarısız olunmuştu.

- Hambleton'a göre (1994) ise, ölçüt-dayanaklı ölçme, 1960'lı yılların Amerikan eğitimindeki "gruba-dayalı" yaklaşımdan "bireye-dayalı" yaklaşıma sıçramanın bir ürünüdür; bu hareketle birlikte, davranışsal hedefler önemli duruma gelmiş ve 1980'lerin ortalarına kadar bu konuda çok sayıda makale yayınlanmasına rağmen, son yıllarda bu alandaki makale sayısında büyük düşüşler gözlenmiştir. Ona göre, artık üst düzeyde düşünme ve muhakeme becerileri üzerine yoğunlaşılmakla birlikte, bu değerlendirme biçimleri de ölçüt-dayanaklı karakterdedir. Hambleton'a göre de Glaser yanlış anlaşılmış, ölçüt-dayanaklı ölçme çok dar eğitim hedefleri

ile çoktan seçmeli test maddelerine sıkışıp kalmıştı. Bugün aslında ölçüt-dayanaklı kavramı hala canlıdır, eskisine göre daha geniş bir alanı kapsamaktadır ve Glaser'ın orijinal amaçlarıyla daha tutarlıdır.

Yukarıda kısaca özetlenen yazılara eklenecek pek fazla bir şey kalmıyor görünmekle birlikte, dikkat edilirse hala kavram karmaşası devam etmektedir. Bilişsel öğelere ağırlık verilmesi ise, bu kavram karmaşasını çözmeye yetmemektedir.

Sonuç

Psikolojik ölçme araçlarından elde edilen sonuçlara dayanarak bireyler hakkında çok ciddi kararlar verilmektedir. Örneğin, eğitimde geçer-kalır; işyerlerine eleman seçiminde kabul-red; klinik ortamlarda çeşitli tanı kararları gibi. Bu kararların çoğu da, ölçme aracı hangi düzeyde ölçme yaparsa yapsın, sınıflama düzeyindedir. Bu noktada, ölçme araçlarının bu sınıflamayı ne kadar doğrulukla yaptığının belirlenmesi, verilecek kararların ciddiliği ile yakından ilişkilidir. Ancak, bundan önce de, söz konusu ölçme araçlarının klasik anlamda –bilinen yollarla güvenirlik ve geçerliklerinin sağlanması gerekmektedir. Öncelikle bu işlemler yapılmalı, daha sonra sınıflama tutarlığı saptanmalıdır. Bireylerin ne kadar tutarlı sınıflandığını belirlemek için geliştirilen indekslerin oturtulacağı yer de belirlenmek durumundadır ki, bu da yeni bir kavramsallaştırmayı gerekli kılmaktadır (Bu konuda yapılan denemelik bir sınıflama Erkuş, 2001'de verilmiştir).

Yukarıda ele alınan indeksler, "keyfi" bir kesme puanına dayanmalarına ve bu nedenle zaaflı yanları olmasına rağmen, gereklidirler. Ancak, öncelikle bu indeksler bağlamındaki yanlış kavramsallaştırmalardan kurtulunması gerekmektedir.

1) Ölçüt-dayanaklı ya da norm-dayanaklı "testler", "ölçme", vb. kavramlaştırmalar doğru görünmemektedir. Çünkü söz konusu ayırım test veya ölçme değil, değerlendirme aşamasındaki ayırımdır.

2) Ölçüt-dayanaklı yaklaşım bağlamında geliştirilmiş olan indeksleri "güvenirlik" indeksleri veya "uyum" katsayıları olarak adlandırmak doğru görünmemektedir. Bu tür indeksleri "sınıflama kararlarının tutarlığı" veya "sınıflama geçerliği" indeksleri şeklinde adlandırmak daha uygun olabilir.

3) Değerlendirme türlerinin sınıflanması da yeniden yapılmalıdır: Norm-dayanaklı, kesme puanına-dayalı ve alan-dayanaklı şeklinde bir sınıflama uygun görünmektedir.

Geliştirilmiş bulunan sınıflama tutarlığı indekslerinin tümü kesme puanının dağılım içindeki pozisyonundan etkilenmekte ve radikal değişiklikler göstermektedirler. Ayrıca kesme puanına bağlı olunması yanlış sınıflama olasılıklarını artırmaktadır. Bu durum, testten ayrı ve keyfi kesme puanı gibi bir ölçüt yerine, testin kendi içinde bir ölçüte dayanan ve yanlış sınıflama olasılıklarını en aza indiren yeni indekslerin geliştirilmesi gerektiğini öne çıkarmaktadır. Bu konuda, yeni bir yöntem ve buna dayanarak geliştirilen bir indeks (Erkuş, 2000) önerilmiş bulunmaktadır.

Kaynaklar

Angoff, W. H. (1971). *Norms, scales, and equivalent scores.* In R. L. Thorndike (Ed) Washington: American Council on Education.

APA, AERA & NCME (1985). *Standards for educational and psychological testing.* Washington: APA.

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

Brennan, R. L. & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14 (3),* 277-289.

Breyer, F. J. & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method.* New Jersey: ETS, Research Report.

Cantor, A. B. (1996). Sample-size calculations for Cohen's Kappa. *Psychological Methods, 1(2),* 150-153.

Cicchetti, D. V. (1972). A new measure of agreement between rank-ordered variables. *Proceeding, 80th Annual Convention, APA,* 17-18.

Cicchetti, D. V. (1977). Comparison of the null distributions of weighted Kappa and the C ordinal statistic. Applied Psychological Measurement, 1(2), 195-201.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, XX(1),* 37-47.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provisions for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-220.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: CBS College Pub. Co.

Cronbach, L. J., Linn, R. L., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: John Wiley.

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment, 8(4),* 360-362.

Ebel, R. L. (1972). Essentials of educational measurement (2nd ed). Englewood Cliffs, New Jersey: Prentice-Hall.

Erkuş, A. (2000 basımda). Yeni bir indeks önerisi: Çift tutarlık indeksi. *Türk Psikoloji Dergisi.*

Erkuş, A. (2001 basımda). Psikometri üzerine yazılar I. Ankara: Türk Psikologlar Derneği Yayınları.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Glaser, R. (1994a). Criterion-referenced tests: Part I. Origins. *Educational Measurement: Issues and Practice, 1,* 9-11.

Glaser, R. (1994b). Criterion-referenced tests: Part II. Unfinished business. *Educational Measurement: Issues and Practice, 1,* 27-30.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-262.

Gleser, G. C., Cronbach, L. J. & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika, 30,* 395-418.

Guilford, J. P. (1954). *Psychometrics methods (2. Ed).* New York: McGraw-Hill Book Co.

Gulliksen, H. (1967). *Theory of mental tests (6. Ed).* New York: John Wiley and Sons Inc.

Gupta, S. S. (1963). Probability integrals of multivariate normal and multivariate. *Annals of Mathematical Statistics, 34,* 792-828.

Hambleton, R. K. & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10 (3),* 159-170.

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.

Hambleton, R. K. (1994). The rise and fall of criterion-referenced measurement. *Educational Measurement: Issues and Practice, 1,* 21-26.

Harris, C. W. (1972). An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement, 9,* 27-29.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika, 41 (1),* 65-78.

Huynh, H. (1978). Reliability of multiple classifications. *Psychometrika, 43 (3),* 317-325.

Huynh, H. & Saunders, J. C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement, 17(4),* 351-358.

Linn, R. L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice, 1,* 12-14.

Livingston, S. A. (1972a). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9,* 13-26

Livingston, S. A. (1972b). A reply to Harris' "An interpretation of reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement, 9,* 31.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179-198.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* New York: Addison-Wesley Pub. Co.

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed), *A guide to criterion-referenced test construction,* 29-48. Baltimore: Johns Hopkins Unv. Press.

Millman, J. (1994). Criterion-referenced testing 30 years: Promise broken, promise kept. *Educational Measurement: Issues and Practice, 1,* 19-39.

Mills, C. N. (1983). A comparison of three methods of establishing cutoff scores on criterion-referenced tests. *Journal of Educational Measurement, 20,* 283-292.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Nunnally, J. C. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill Book Co.

Ornstein, A. C. & Gilman, D. A. (1991). The striking constrats between norm-referenced and criterion-referenced tests. *Contemporary Education, 62(4),* 287-293.

Peng, C-Y. J. & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368.

Popham, W. J. & Husek, T. R. (1969). Imlications of criterion-referenced measurement. *Journal of Educational Measurement, 6(1),* 1-9.

Popham, W. J. (1975). *Educational evaluation.* Englewood Cliffs, New Jersey: Prentice-Hall.

Popham, W. J. (1994). The instructional consequences of criterion-referenced clarity. *Educational Measurement: Issues and Practice, 1,* 15-18.

Rae, G. (1997). Sampling behaviour of Kappa and weighted Kappa in the null case. British *Journal of Mathematical and Statistical Psychology, 50,* 1-7.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement, 13,* 265-276.

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. A. Berk (Ed), *criterion-referenced measurement,* 129-185. Baltimore: John Hopkins Unv. Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25(1),* 47-55.

Swaminathan, H., Hambleton, R. K. & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263-267.

Turgut, M. F. (1992). *Eğitimde ölçme ve değerlendirme metotları (8. baskı).* Ankara: Saydam Matbaacılık.

Turgut, M. F. & Baykul, Y. (1992-1). *Ölçekleme teknikleri.* Ankara: ÖSYM Yayınları.

Vos, H. J. (1997). Simultaneous optimization of quota-restricted selection decisions with mastery scores. *British Journal of Mathematical and Statistical Psychology, 50,* 105-125.