

PAPER DETAILS

TITLE: Item and test analysis of multiple-choice exams which applied to the First term students of pamukkale university medical faculty In 2006-2007 academic year

AUTHORS: Kenan TOPAL,Hülya AYBEK,Cüneyt Orhan KARA,Akile BÜKE,Zafer AYBEK

PAGES: 120-126

ORIGINAL PDF URL: <https://dergipark.org.tr/tr/download/article-file/422270>

PAÜ TIP FAKÜLTESİ DÖNEM I ÖĞRENCİLERİNE 2006-2007 EĞİTİM VE ÖĞRETİM YILINDA UYGULANAN ÇOKTAN SEÇMELİ SINAVLARIN MADDE VE TEST ANALİZLERİ

ITEM AND TEST ANALYSIS OF MULTIPLE-CHOICE EXAMS WHICH APPLIED TO THE FIRST TERM STUDENTS OF PAMUKKALE UNIVERSITY MEDICAL FACULTY IN 2006-2007 ACADEMIC YEAR

Kenan TOPAL*, Hülya AYBEK**, Cüneyt Orhan KARA***, Akile BÜKE****, Zafer AYBEK*****

*Pamukkale Üniversitesi, Tıp Fakültesi, Tıp Eğitim AD.

** Pamukkale Üniversitesi, Tıp Fakültesi, Biyokimya AD.

*** Pamukkale Üniversitesi, Tıp Fakültesi, Kulak Burun Boğaz AD.

**** Pamukkale Üniversitesi, Tıp Fakültesi, Çocuk Cerrahisi AD.

***** Pamukkale Üniversitesi, Tıp Fakültesi, Üroloji AD.

Özet

Çoktan Seçmeli Sınavlar bilişsel düzeydeki öğrenme hedeflerini değerlendirmek amacıyla sıkça kullanılır. 2006-2007 Eğitim ve Öğretim yılında Dönem I öğrencilerine uygulanan Çoktan Seçmeli Sınavların Madde ve Test Analizleri yapılmıştır. Sonuçların öğretim üyeleri ile paylaşarak daha kaliteli test maddelerinin hazırlanması ve bir soru bankasının yapılandırılması amaçlanmıştır.

Bir yılda uygulanan toplam 23 Çoktan Seçmeli Sınav ve 1368 madde incelenmiştir. Maddelerin Zorluk ve Ayıricılık İndeksleri Kelley tarafından önerilen sistemle hesaplanmıştır. Her bir testin Güçlük, Çarpıklık değerleri hesaplanmış ve Kuder-Richardson katsayısı bulunmuştur.

Zorluk İndeksi açısından önerilen ve kabul edilebilir sınırın içinde 700 adet madde (%51.1) bulunmuştur. Ayıricılık İndeksine göre ise 699 madde çok iyi (%51.0) ve 284 madde iyi (%20.7) olarak saptanmıştır. Zorluk ve Ayıricılık İndekslerine göre toplam 651 madde (%47.5) kaliteli olarak belirlenmiştir. Testlerin Güçlük değerinin ortalaması (0.644) 0.5'e yakın bulunmuştur. Çarpıklık değeri bütün testler için negatiftir. Testler genel olarak güvenilirlik açısından değerlendirildiğinde Dizi genişliği/Standart sapma oranı 19 test için 4-6 arasında, testlerin tamamında Kuder-Richardson 21 indeksinin 0.5'in altında olduğu görülmüştür.

Testlerin bir bütün olarak analizi güvenilirliğin iyi olduğunu göstermiştir. Ancak KR-21 indeksinin düşük olması (<0.5) testlerin kendi içinde kararlı olmadığına bir göstergesidir. Sonuçların ilgili öğretim üyeleri ile paylaşarak daha kaliteli sınav sorularının hazırlanmasına ve yapılandırılacak bir soru bankası ile çalışmaların sürdürülmesine karar verilmiştir. (Pam Tıp Derg 2008;1(3):120-126).

Anahtar kelimeler: Çoktan seçmeli sınavlar, madde ve test analizi, soru bankası

Abstract

The Multiple Choice Questions are often used to evaluate the educational objectives for cognitive learning. Item and test analysis of the Multiple Choice Tests which applied to first term students was performed in 2006-2007 academic year. Sharing the results with academicians to prepare better quality test items and construct a test question bank was aimed.

A total of 23 Multiple Choice Tests and 1368 items which were applied in one year were analysed. The difficulty and discrimination index of items were measured with the system proposed by Kelley. The facility and discrimination index and Kuder-Richardson coefficient were calculated for each test.

There were 700 items (51.1%) in the suggested and acceptable level according to difficulty index of the questions. Analyzing the discrimination index revealed that 699 items (51.0%) were very good and 284 (20.7%) items were good. Total 651 items (47.5%) were determined of good quality according to difficulty and discrimination index. The average facility index (0.644) of tests was found near 0.5. The discrimination indexes of all tests were negative. When the reliability of tests was appraised, it was seen that the Range/Standard deviation ratio was between 4 and 6 for 19 tests and Kuder-Richardson 21 index is under 0.5 for all tests.

The entire analyses of tests showed that the reliability was good. However the low index of Kuder-Richardson 21 indicates the poor inner consistency of the tests. Sharing the results with academicians for preparing better quality Multiple Choice Questions and to continuing the studies for constructing of a test question bank is concluded. (Pam Med J 2008;1(3):120-126).

Key words: Multiple choice tests, item and test analyse, test question bank

Giriş

Öğrencilerin uygulanan programın hedeflerine ulaşip ulaşmadıklarını belirlemek için kullanılan

Ölçme ve Değerlendirme eğitim sisteminin ana öğelerinden birisidir. Objektif test türlerinden birisi olan Çoktan Seçmeli Sınavlar bilişsel

Kenan TOPAL

Pamukkale Üniversitesi Tıp Fakültesi, Tıp Eğitimi Anabilim Dalı, Denizli
e-posta: ktopal@pau.edu.tr

düzeydeki öğrenme hedeflerini değerlendirmek amacıyla Tıp Eğitiminde sıkça kullanılır [1-3]. Bu amaçla uygulanan sınavların adil olması, geçerli ve güvenilir sonuçlar vermesi, kullanışlı olması ve öğrenmenin düzeyini tam ve doğru olarak saptaması beklenir [4, 5].

Bu çalışmada Pamukkale Üniversitesi Tıp Fakültesinde 2006-2007 öğretim yılında Dönem I öğrencilerine uygulanan Çoktan Seçmeli Sınavların madde analizleri ve test analizleri yapılarak sonuçların ilgili öğretim üyeleri ile paylaşılması amaçlanmıştır. Böylece daha kaliteli test maddelerinin hazırlanması için tekrar kullanılmayacak olanların ayıklanması, madde dosyasına ayrılacak olanların belirlenerek bir soru bankasının yapılandırılması planlanmıştır.

Gereç ve Yöntem

2006-2007 öğretim yılında Dönem I öğrencilerine uygulanan 16 adet Modül Sonu Değerlendirme (MSD), 5 adet Blok Sonu Değerlendirme (BSD) ve 2 adet Öğrenci Başarı Değerlendirmesi (ÖBD) sınavı olmak üzere toplam 23 test ve 1368 madde incelenmiştir. Madde analizi; Zorluk İndeksi (P) ve Ayırıcılık İndeksi (D) hesaplanarak yapılır. Zorluk İndeksi bir sorunun doğru yanıtlanma oranını ölçer. Temel olarak Zorluk indeksi %0-%100 arasındaki aralık içerisinde değerlendirilir, iyi sorunun P değeri %50-%60 arasında ve kabul edilebilir sorunun P değeri %30-%70 arasında olmalıdır. $P < \%30$ olan sorular “çok zor”, $P = \%30-49$ arası ise “kabul edilebilir alt sınır”, $P = \%50-59$ ise “önerilen”, $P = \%60-69$ ise “kabul edilebilir üst sınır” ve $P > \%70$ ise “çok kolay” olarak değerlendirilir. İndeks değeri yükseldikçe soru kolaylaşır. Kabul edilebilir sınırın dışındaki bir soru ($P < \%30$, $P > \%70$) gerekirse sınavdan çıkarılmalıdır. Bir test için en uygun olanı çoğunlukla orta zorlukta ($P = \%50-59$), düşük performanstaki öğrenciler için birkaç tane kolay ($P = \%60-69$), yüksek performanstaki öğrencileri hırslandırmak için de birkaç tane de zor madde ($P = \%30-49$) konulmasıdır. Tüm ölçümün madde yoğunluğunun %50-60 arasında olması en idealdir. Ayırıcılık indeksi her sorunun alt ve üst gruptaki öğrencileri ayırabilme gücünü ölçer. Yani bilenle bilmeyeni ayırt etme gücünü (D) ölçer. Ayırıcılık indeksinin aralığı (-1.00 ile +1.00) arasındadır. Temel olarak D değeri, +0.35 ve üzeri ‘mükemmel soru’, +0.25 ve +0.34 arası ‘iyi soru’, +0.15 ve +0.24 arası ‘gözden geçirilmesi gereken soru’, +0.15 altında olan soru ise ‘kötü soru’ olarak kabul edilir ve gerekirse sınavdan çıkarılmalıdır. İndeksin değeri yükseldikçe sorunun ayırt ediciliği de artar Bu indeks maddeyi doğru yapan ve yüksek not alan öğrenci sayısı ile maddeyi doğru yapan ve düşük not alan öğrenci sayısı karşılaştırılarak bulunur. Değerlendirme sonucunun (-1.00) olması tüm

düşük not alan öğrencilerin o maddeyi doğru yaptığını ve tüm yüksek not alan öğrencilerin o maddeyi yanlış yaptığını gösterir. Değerin +1.00 olması maddenin olması gereken gibi ifade edildiğini gösterir. Değerin 0.00 olması ise maddenin konuyu bilen bilmeyen ayırımını yapmadığını gösterir ki istenmeyen bir durumdur. En mantıklı olanı değer +0.25-0.50 arasında olmasıdır. Maddelerin Zorluk ve Ayırıcılık İndeksleri Kelley tarafından önerilen sisteme dayanılarak geliştirilen bağlantı ve formül sistemlerini kullanan bilgisayar programlarıyla hesaplanmıştır [5-11].

Daha sonra testlerdeki puanların dağılımına göre test analizleri yapılmıştır. Genel olarak, standart sapmanın küçük olması; ortalamadan sapmaların ve riskin az olduğunun, büyük olması ise; ortalamadan sapmaların, riskin çok olduğunun ve oynaklığın göstergesidir [12]. Ancak bir testten elde edilen puanların standart sapması (S) büyüdükçe o testin güvenilirliği artar. Dolayısıyla farklı başarı düzeyindeki öğrencileri birbirinden ayırmaya yönelik bir testin standart sapması da büyük olmalıdır. Dağılımın standart sapması ile dizi genişliği ($R = \text{en büyük puanla en düşük puan arasındaki fark}$) birlikte de düşünülebilir. Dizi genişliğinin standart sapmaya oranı (R/S) güvenilirlik ile ilgili bir fikir verir. R/S oranının 4-6 arasında bir değer olması istenir [3]. Bundan başka her bir testin Güçlük ve Çarpıklık değerleri hesaplanır. Bir testin güçlük derecesi o testin maksadına bağlı olarak değerlendirilir. Testin güçlüğü aritmetik ortalamanın (X) alınabilecek en yüksek puana (AMP) oranlanmasıyla bulunur (X/AMP). Oran 1’e yaklaştıkça güçlük azalır. Biçimlendirme-yetiştirmeye yönelik testlerde her öğrencinin sahip olması gereken davranışları ölçmek ya da öğrencinin öğrendiği konuyu pekiştirmek amaçlandığından güçlük değerini hesaplamak gereksizdir. Ancak farklı başarı düzeylerini ayırt etmek ve öğrenme derecelerine göre sıralama yapmayı amaçlayan başarı testlerinin ortalama güçlüğü 0.50 civarında olması istenilir. Eğer testin güçlüğü 0.50’den küçükse ya test öğrencilere güç gelmiştir ya da öğretim yetersizdir; istenilen düzeye çıkarılamamış zayıf öğrenciler çoğunluktadır. Eğer ortalama güçlük 0.50’den büyükse öğrencilerin çoğunun öğrenilecek konuları öğrendiği ya da testin öğrencilere kolay geldiği şeklinde yorumlanabilir [3]. Bir dağılımın ortalaması, ortancası ve standart sapmasını kullanan bir formülle o dağılımın çarpıklığı hakkında bir fikir edinilebilir [13]. Çarpıklık, $([3(X - \text{Medyan})] / S)$ formülüyle hesaplanır ve bu çarpıklık ölçüsünden testin zorluğuna ya da kolaylığına ilişkin yoruma gidilir (Tablo 1). Çarpıklık değerinin grafik olarak yorumlanması Şekil 1’de verilmiştir.

Kuder Richardson 20 ve 21 formülleri (KR-20, KR-21) test kalitesi veya testin güvenilirlik değerinin, kısaca testin iç tutarlılığının bir göstergesidir. Eğer sınav tekrar yapılsa birincinin etkisinde kalmaksızın öğrencilerin ne derecede yine aynı puanı alabileceklerini yansıtır. KR-20 formülü her maddenin birbiri ile paralel olduğu ve aynı ortalama ve varyansa sahip olduğu yani testin ölçtüğü şeyin homojen olduğu varsayımından hareketle geliştirilmiştir. Formülün uygulanmasıyla 0.0 ile 1.0 arasında değişen değerler elde edilir. Değer 1.0'e yaklaştıkça güvenilirliğin yüksek olduğu kabul edilir. Maksimum olası KR-20 değeri 1.0'dir. Grup karşılaştırmasında kullanılmak üzere hazırlanan testlerin güvenilirlikleri 0.6-0.8 arasında olabilir. Bireyler hakkında karar vermede kullanılacak testlerin güvenilirliklerinin 0.8'in, karar çok ciddi sonuçlara yol açabilecekse 0.9'ın üzerinde olması beklenir. Eğer KR-20 değeri 0.8'in altında ise o teste dayanarak öğrenci hakkında önemli bir bireysel karara varılmamalıdır. KR-20 İndeksi Tablo 2'de verilen oranlara göre yorumlanır [11, 14].

Eğer bir testteki maddelerin güçlük dereceleri birbirinden önemli ölçüde farklı değilse o testin güvenilirliğinin tahmini için KR-20 formülünün özel hali olan KR-21 kullanılır. Formülün bu hali madde güçlük indekslerinin bilinmediği ya da hesaplanmadığı durumlarda tercih edilir. KR-21 formülü test puanları ortalaması, standart sapması ve madde sayısına bağlı olarak hesaplandığından madde analizi yapılmamış testlere de uygulanabilmektedir. Oysa KR-20 formülü madde analizi yapılmamış testlere uygulanamaz. KR-21 madde güçlükleri indeksinin eşit olduğu yani, güçlük derecesinin %50 olduğu varsayımına dayanır. Bu varsayım pratikte nadiren gerçekleşir ve bu nedenle testi oluşturan maddelerin güçlükleri arasındaki farklılıklara duyarlı değildir. Dolayısıyla testin maddeleri eşit güçlükte olamayacağı için KR-21 eşitliği ile elde edilen güvenilirlik katsayıları KR-20 ile elde edilen güvenilirlik katsayılarına göre genellikle küçük çıkar (KR-20 > KR-21). Bu sebeple, KR-21 ile hesaplanan değer testin güvenilirlik katsayısının alt sınırı olarak kabul edilir. Genel olarak bir testin KR-20 veya KR-21 güvenilirliği yüksek bulunmuşsa, bu testteki maddelerin aynı yeterliği ölçtüğü ve test puanlarının tesadüfi hatalardan arınık olduğu sonucuna ulaşılabilir [3, 4, 15].

Bulgular

Testlerin Madde Analizleri: Yapılan 23 testte bulunan toplam 1368 maddenin Zorluk İndeksi (P) ve Ayırıcılık İndeksi (D) bulunmuştur. Bu maddeler Zorluk İndeksi açısından değerlendirildiğinde 104'ü çok zor (% 7.6) ve 564'ü çok kolay (%41.2), 289'u kabul edilebilir

üst sınır (% 21.1), 218'i kabul edilebilir alt sınır (% 15.9) ve 193'ü önerilen (%14.1) olarak belirlenmiştir. Zorluk İndeksi açısından önerilen ve kabul edilebilir sınırın içinde ($P > \%30 - P < \%70$) 700 adet madde (% 51.1) bulunmuştur. Ayırıcılık İndeksi açısından değerlendirildiğinde ise 699 madde çok iyi (%51.0), 284 madde iyi (% 20.7), 176 madde gözden geçirilmesi gereken, sınırdan (%12.8) ve 209 madde zayıf (%15.2) olarak saptanmıştır. Zorluk ve Ayırıcılık İndekslerine göre ($P > \%30 - P < \%70$ ve $D > +0.15$) 651 madde (% 47.5) kaliteli olarak belirlenmiştir (Tablo 3).

Testlerin Bir Bütün Olarak Analizi: Puanların dağılımına dayanarak, uygulanan 23 testin analizleri sonucunda elde edilen Aritmetik Ortalama (mean), Ortanca (median), Tepe Değer (mod), Standart Sapma (S), Dizi Genişliği (R), Testin Dizi Genişliği (TR) ve Beklenen Dizi Genişliği (BR) değerleri Tablo 4'de verilmiştir.

Güçlük (X/AMP), çarpıklık, dizi genişliliğinin standart sapmaya oranı (R/S) ve Kuder-Richardson (KR-21) değerleri Tablo 5'de verilmiştir. Buna göre bütün testlerin Güçlük değerinin (X/AMP) ortalaması 0.50'nin üzerinde (0.644) bulundu. Çarpıklık değeri bütün testler için negatifti ve ortalaması -0.573 bulundu. Dizi genişliliğinin standart sapmaya oranı 19 test için 4-6 arasında idi, 2 testte 6'nın üzerinde 2 testte ise 4'ün altında bulundu. Testlerin tamamında KR-21 indeksi 0.5'in altında bulundu.

Tartışma

Çoktan seçmeli sınavlar Tıp Eğitiminde öğrencinin değerlendirilmesi için çok yaygın kullanılır. Testi alanların, yanıtlarını belirli sayıdaki seçenek arasından birini seçerek vermesini gerektiren sorulardan oluşan ölçme araçlarıdır. Eğer iyi ve uygun hazırlanırsa çoktan seçmeli sorularla bilişsel alanın hemen her düzeyindeki davranışları ölçmek mümkün olabilir. Öğrencilere objektif, güvenilir ve geçerli çoktan seçmeli sınavlar yapılması için sınav sorularının madde analizini yapmak gereklidir. Asıl olarak istenen niteliklere sahip bir test hazırlamak amacıyla soruların saklanması, yeniden gözden geçirilmesine veya sorunun atılıp atılmayacağına karar verilmesini sağlar. Böylece iyi çalışmış maddeler ortaya çıkartılır, iyi çalışmayan maddeler düzeltilerek yeniden kullanılır ya da tamamen çıkartılır. Testler bir maddenin doğru yanıtlanma oranını ölçen Zorluk İndeksi açısından değerlendirildiğinde çok zor ve çok kolay maddeler dışarıda bırakıldığında kabul edilebilir sınırın içinde ($P > \%30 - P < \%70$) 700 adet madde (% 51.1) bulunmuştur. Ayırıcılık İndeksi açısından incelendiğinde ise 875 madde çok iyi ya da iyi (% 63.9), 176 madde ise gözden geçirilmesi gereken (%12.8) olarak saptanmıştır. Bu da bize sorularımızın bilenle bilmeyeni ayırt

etme açısından iyi durumda olduğunu göstermektedir. Hem Zorluk hem de Ayırıcılık İndekslerine göre ($P > \% 30 - P < \% 70$ ve $D > +0.15$) toplam 651 adet (% 47.5) kaliteli madde belirlenmiş ve bunların madde dosyasına ayrılması önerilmiştir.

Eskiocak ve arkadaşlarının yaptığı bir çalışmada Trakya Üniversitesi Tıp Fakültesinde son beş yılda yapılan Biyokimya testlerinde sorulan 1290 sorudan 634 (% 49.1) tanesi zorluk indeksi açısından kabul edilebilir sınırlar içerisinde bulunmuştur [5]. Yine aynı çalışmada Ayırıcılık İndeksi açısından 814 sorunun çok iyi ve iyi düzeyde (%62.7) olduğu bulunmuştur. Zorluk ve Ayırıcılık İndekslerine göre 589 madde (% 45.6) kaliteli olarak belirlenmiştir. Bizim yaptığımız analizlerde de benzer bulgulara ulaştığımız görülmektedir.

Dizi genişliği (Ranj) en büyük ve en küçük puanlar arasındaki farka eşittir. Dizi genişliği uygulandığı grup ve yöneldiği maksat için testin uygunluğunu belirlemeye yardımcıdır. Testin dizi genişliği (TR) beklenen dizi genişliğine (BR) yani testten elde edilmesi gereken en yüksek puanın yarısına yakın olmalıdır. Bizim yaptığımız analizlerde TR'nin BR'e göre daha düşük olduğu bulunmuştur, bu bize testlerin güçlük düzeyi yönünden kolay sorulardan oluştuğunu veya öğrencilerin çoğu konuları öğrendiğini gösterir.

Analizler sonucunda 23 testten 19'unda dizi genişliğinin standart sapmaya oranı (R/S) 4 ile 6 arasında bulundu, bu da bize genel olarak testlerin güvenilirliğin iyi olduğu hakkında bir fikir verir. Testlerin Güçlük değerinin (X/AMP) ortalaması 0.50 civarında (0.644) bulunması başarı testlerinde olması istenilen bir durumu ifade eder. Diğer yandan testlerdeki puan dağılımının çarpıklığı ortalama -0.573 olarak bulunmuştur, bu negatif çarpıklık bize testlerin kolay olduğunu ya da özellikle, biçimlendirme-yetiştirmeye yönelik testlerde olduğu gibi

öğrenciler testin ölçtüğü içeriği tam olarak öğrenmiş olabileceğini gösterebilir. Yapılan analizler sonucunda testlerin tamamında KR-21 indeksini 0.5'in altında bulunmuştur. Bunun anlamı bu testlere dayanarak öğrenci hakkında önemli bir bireysel karara varılmaması gerektiği şeklinde ifade edilir. KR-21'in düşük olma nedenleri ise: çok kolay soruların fazla olması, çok zor soruların fazla olması, ayırıcılık özelliği bulunmayan kötü yazılmış sorular ve içeriği bütünüyle test etme ön koşuluna uyulmaması olarak sayılabilir. Bir testin ya da o testi oluşturan maddelerin sahip olması istenen özellikler testin maksadına bağlıdır. Eğer bir testin niçin verildiği bilinmiyorsa, onu uygulamanın ve uygulama sonuçlarını analiz etmenin anlamı yoktur. Dolayısıyla bir testin ne derecede amaca yönelik olduğunu anlamak, testin mantıksal ve istatistiksel açıdan çözümlenmesini gerektirir [16-18].

Sonuç

PAÜ Tıp Fakültesi Dönem 1 öğrencilerine 2006-2007 eğitim öğretim yılında uygulanan testlerin analizi ile elde edilen veriler ilgili öğretim üyelerine geri bildirimde bulunularak daha kaliteli sınav sorusu hazırlanmasına katkı sağlamak hedeflenmiştir. Yapılan analiz ve sonuçların öğretim üyeleriyle paylaşılması soruların niteliğinde olumlu değişikliklere yol açacak, istenilen nitelikleri taşıyan bir testin geliştirilmesinde uygun soruların seçilmesini daha kolaylaştıracak, sorulamayacak kadar zor ve kolay soruların azaltılmasını sağlayacaktır. Amacımız soruların ve sınavların isteneni ölçen, hakkaniyetli, dengeli ve özgün olmalarının sağlanmasıdır. Ayrıca uzun dönemde madde ve seçenek analizi çalışmalarının sürdürülerek yapılandırılacak bir soru bankasının temel öğeleri olan madde dosyalarını hazırlanması planlanmıştır.

Tablo 1: Testlerin Çarpıklık değerine göre yorumlanması.

Çarpıklık	Grup için testin güçlüğü
Negatif (-1.0 ile 0.0)	Kolay
Pozitif (0.0 ile +1.0)	Zor
0.10'dan küçük	Hafif zor
0.10 - 0.25	Orta güçlükte
>0.25	Çok zor

Kaynak: Tekin H. Eğitimde Ölçme Ve Değerlendirme, Yargı Yayınevi: 17. Baskı, Ankara; 2004.

Tablo 2. KR-20 İndeksinin yorumlanması.

KR-20 İndeksi	Yorumu
> 0.9	Mükemmel Güvenilirlik
> 0.8	İyi Güvenilirlik
0.8-0.5	Sınırdaki Test (düzeltilmeli)
< 0.5	Test sonuçlarına güvenilemez

Kaynak: Kocabaşoğlu YE. Çoktan Seçmeli Soru Analizi ve Çoktan Seçmeli Soru Sınavları, Akdeniz Ü. Tıp Fak. Ölçme Değerlendirme Kursu: Antalya; 2002.

Tablo 3. Testlerin zorluk ve ayıricılık indeksine göre yapılan madde analizi sonuçları.

ZORLUK İNDEKSİ (P)	AYIRICILIK İNDEKSİ (D)								Toplam	
	Çok İyi		İyi		Sınırdan		Zayıf			
	D >+0.35		D =+0.25- +0.34		D =+0.15- +0.24		D <+0.15		N	%
	N	%	N	%	N	%	N	%		
Çok Zor P < %30	13	0.95	18	1.30	19	1.38	54	3.94	104	7.60
Kabul Alt Sınır P = %30-49	118	8.60	50	3.65	30	2.1	20	1.46	218	15.9
Önerilen P = %50-59	138	10.0	21	1.53	19	1.38	15	1.09	193	14.1
Kabul Üst Sınır P = %60-69	219	16.0	36	2.63	20	1.46	14	1.02	289	21.1
Çok Kolay P > %70	211	15.4	159	11.3	88	6.4	106	7.74	564	41.2
Toplam	699	51.0	284	20.7	176	12.8	209	15.2	1368	100.0

Not: gri boyalı alan Zorluk ve Ayıricılık İndekslerine göre belirlenen 651 kaliteli maddeyi göstermektedir.

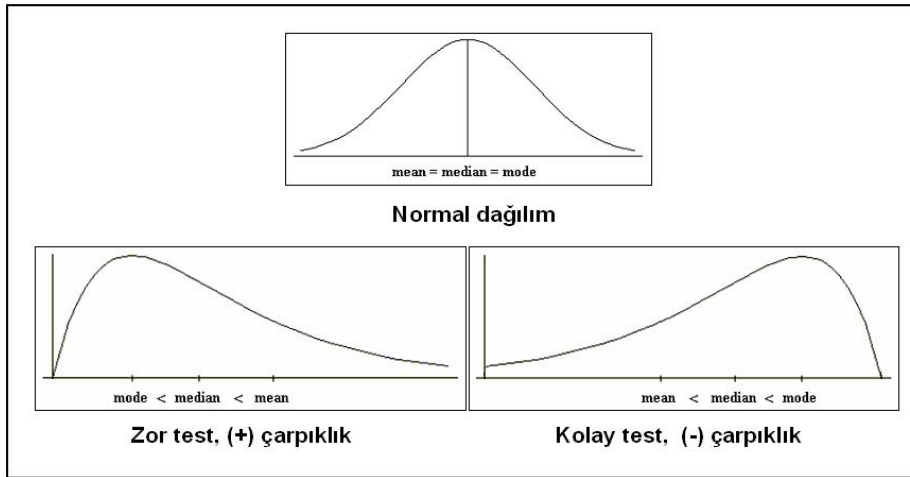
Tablo 4. Uygulanan 23 testin analizleri sonucunda elde edilen Aritmetik Ortalama, Ortanca, Tepe Değer, Standart Sapma, Dizi Genişliği, Testin Dizi Genişliği ve Beklenen Dizi Genişliği değerleri.

Test	Aritmetik Ortalama Mean	Ortanca Medyan	Tepe Değer Mod	Standart Sapma S	Dizi Genişliği R	Testin Dizi Genişliği TR	Beklenen Dizi Genişliği BR
MSD 1	67.9	70.0	68.7	13.0	90.0	45.0	50.0
MSD 2	66.8	70.2	72.3	14.9	87.2	43.6	49.9
MSD 3	68.5	72.0	74.0	14.5	96.0	48.0	50.0
MSD 4	71.3	75.0	82.0	15.4	88.0	44.0	50.0
MSD 5	65.1	68.7	64.5	17.7	87.5	43.7	49.9
MSD 6	72.4	73.9	72.9	16.7	93.7	46.8	49.9
MSD 7	65.6	66.6	66.6	17.1	92.8	46.4	49.9
MSD 8	59.9	62.5	62.5	17.0	82.5	41.2	50.0
MSD 9	57.8	60.4	62.5	19.4	87.5	43.7	49.9
MSD 10	53.3	60.0	68.0	21.4	86.0	43.0	50.0
MSD 11	60.2	66.0	68.0	21.5	86.0	43.0	50.0
MSD 12	55.3	60.0	68.0	20.5	88.0	44.0	50.0
MSD 13	73.7	80.0	84.0	22.4	92.0	46.0	50.0
MSD 14	67.7	72.0	74.0	20.2	90.0	45.0	50.0
MSD 15	62.7	68.0	78.0	21.9	86.0	43.0	50.0
MSD 16	68.0	72.2	84.4	21.8	93.3	46.6	49.9
BSD 1	60.8	62.0	66.6	13.7	81.3	40.6	49.9
BSD 2	67.6	70.0	71.0	15.2	89.0	44.5	50.0
BSD 3-1	66.0	68.0	68.0	17.9	96.0	48.0	50.0
BSD 3-2	58.9	61.8	62.0	20.9	85.0	42.5	50.0
BSD 4	68.6	74.0	74.0	22.2	91.0	45.5	50.0
ÖBD 1	61.3	65.7	65.2	19.2	81.6	40.8	49.9
ÖBD 2	61.4	65.9	68.0	20.1	85.4	42.7	49.9

Kısaltmalar: MSD; Modül Sonu Değerlendirme, BSD; Blok Sonu Değerlendirme, ÖBD; Öğrenci Başarı Değerlendirmesi

Tablo 5: Testlerin Güçlük (X/AMP), Çarpıklık, Dizi genişliğinin standart sapmaya oranı (R/S) ve Kuder–Richardson (KR-21) değerleri.

Test	Güçlük (X/AMP)	Çarpıklık [3 (X-Medyan)] / S	Dizi genişliğinin standart sapmaya oranı (R/S)	Kuder –Richardson (KR-21)
MSD 1	0.679	-0.477	6.91	0.14
MSD 2	0.669	-0.670	5.81	0.12
MSD 3	0.685	-0.712	6.58	0.12
MSD 4	0.713	-0.717	5.68	0.12
MSD 5	0.652	-0.622	5.06	0.07
MSD 6	0.725	-0.265	5.59	0.13
MSD 7	0.656	-0.178	5.40	0.12
MSD 8	0.599	-0.455	4.85	0.10
MSD 9	0.578	-0.400	4.50	0.03
MSD 10	0.534	-0.925	4.00	0.00
MSD 11	0.603	-0.796	3.98	0.02
MSD 12	0.554	-0.674	4.28	0.01
MSD 13	0.738	-0.835	4.09	0.07
MSD 14	0.677	-0.635	4.45	0.05
MSD 15	0.628	-0.718	3.92	0.03
MSD 16	0.681	-0.568	4.27	0.07
BSD 1	0.608	-0.253	5.90	-0.07
BSD 2	0.677	-0.462	5.83	-0.10
BSD 3-1	0.660	-0.334	5.33	0.06
BSD 3-2	0.589	-0.423	4.05	-0.04
BSD 4	0.686	-0.729	4.09	-0.05
ÖBD 1	0.614	-0.692	4.23	-0.07
ÖBD 2	0.615	-0.658	4.24	-0.06

**Şekil 1.** Çarpıklık değerinin grafik olarak yorumlanması.

Kaynak: Tekin H. Eğitimde Ölçme Ve Değerlendirme, Yargı Yayınevi: 17. Baskı, Ankara; 2004.

Kaynaklar

- Collins J, Education Techniques For Lifelong Learning : Writing Multiple-Choice Questions For Continuing Medical Education Activities and Self-Assessment Modules, RadioGraphics 2006; 26: 543-51.
- Kehoe J, Basic Item Analysis for Multiple-Choice Tests, Practical Assessment, Research & Evaluation 1995. Erişim: <http://pareonline.net/getvn.asp?v=4&n=10>, Erişim Tarihi: 10.02.2008
- Tekin H. Eğitimde Ölçme Ve Değerlendirme, Yargı Yayınevi: 17. Baskı, Ankara; 2004.
- Atılğan H, Kan A, Doğan N. Eğitimde Ölçme Ve Değerlendirme, Anı Yayıncılık: Ankara; 2006.
- Eskiocak S, Süer Gökmen S, Erbaş H, Çakır E, Gülen Ş. Trakya Üniversitesi Tıp Fakültesinde Son 5 Yılda Yapılan Biyokimya Sınav Sorularının Analizi. Türk Biyokimya Dergisi 2004; 29: 273-6.
- Sarin YK, Khurana M, Natsu MV, Thomas AG, Singh T. Item Analysis of Published MCQ's. Indian Pediatrics 1998; 35:1103-5
- Bekiroğlu FO. Ne Kadar Başarılı? Klasik Ve Alternatif Ölçme Değerlendirme Yöntemleri: Fizikte Uygulamalar. Nobel Kitabevi: 1. Basım, Ankara; 2004.
- Kim Seock-Ho. A Computer Program for Classical Item Analysis. The University of Georgia; 1999. Erişim: <http://shkim.myweb.uga.edu/epm99.htm>, Erişim Tarihi: 10.02.2008

9. Ebel Robert L. The Relation of Item Discrimination to Test Reliability. Journal of Educational Measurement 1967; 4: 125-8.
10. Kelley T L. The selection of upper and lower groups for the validation of test items. The Journal of Educational Psychology 1939; 30: 17-24.
11. Kocabaşoğlu YE. Çoktan Seçmeli Soru Analizi ve Çoktan Seçmeli Soru Sınavları, Akdeniz Ü. Tıp Fak. Ölçme Değerlendirme Kursu: Antalya; 2002. Erişim: http://www.akdeniz.edu.tr/tip/tipegitimi/mcqanaliz_files/frame.htm, Erişim Tarihi: 10.02. 2008.
12. Kocabaşoğlu YE. Tıp Eğitiminde Ölçme Değerlendirme Teknikleri (Ön Okuma), Akdeniz Ü. Tıp Fak. Ölçme Değerlendirme Kursu: Antalya; 2002. Erişim: <http://www.akdeniz.edu.tr/tip/tipegitimi/on%20okuma.doc>. Erişim Tarihi: 10.02.2008
13. Arıcı H. İstatistik, Yöntemler ve Uygulamalar. Meteksan Basımevi: Ankara; 1991.
14. Ercan İ, Kan İ. Ölçeklerde Güvenirlilik ve Geçerlik. Uludağ Üniversitesi Tıp Fakültesi Dergisi 2004; 30: 211-6.
15. Özen Y, Gülaçtı F, Kandemir M. Eğitim Bilimleri Araştırmalarında Geçerlik Ve Güvenirlilik Sorunsalı. Erzincan Eğitim Fakültesi Dergisi 2006; 8: 69-89.
16. Durak Hİ. Ölçme-Değerlendirmede (Sınama) Yararlı Yenilikler. UTEK 2004 Ulusal Tıp Eğitimi Kongresi Kitabı: Şanlıurfa; 2004. s. 23-6.
17. Turhan K, Yarış F, Özen İ, Dikici MF. Karadeniz Teknik Üniversitesi Tıp Fakültesi Soru Bankası Projesi ve Dinamik Web Modülü. UTEK 2004 Ulusal Tıp Eğitimi Kongresi Kitabı: Şanlıurfa; 2004. s.162.
18. Crisp GT, Palmer EJ. Engaging Academics with a Simplified Analysis of their Multiple-Choice Question (MCQ) Assessment Results. Journal of University Teaching and Learning Practice 2007; 4: 88-102.