

159929

**GELENEKSEL YÖNTEMLE VE ELEME YÖNTEMİ İLE  
PUANLANAN ÇOKTAN SEÇMELİ TESTLERİN  
PSİKOMETRİK ÖZELLİKLERİNİN İNCELENMESİ**

Bayram ÇETİN

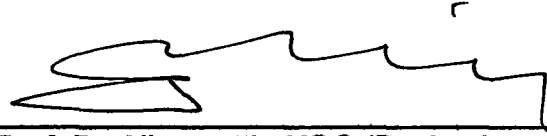
Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü  
Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı

Doktora Tezi

Ankara, 2005

## KABUL VE ONAY

Bayram ÇETİN tarafından hazırlanan "Geleneksel Yöntemle ve Eleme Yöntemi ile Puanlanan Çoktan Seçmeli Testlerin Psikometrik Özelliklerinin İncelenmesi" başlıklı bu çalışma, 10.06.20005 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Doktora Tezi olarak kabul edilmiştir.



Prof. Dr. Nizamettin KOÇ (Başkan)



Yrd. Doç. Dr. Hülya KELEÇİOĞLU (Danışman)



Prof. Dr. A. Ata TEZBAŞARAN



Doç. Dr. Aysun UMay



Doç. Dr. Selahattin GELBAL

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylıyorum.



Prof. Dr. Nuran ÖZYER  
Enstitü Müdürü

## BİLDİRİM

Hazırladığım tezin tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kağıt ve elektronik kopyalarının Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

- Tezimin/Raporumun tamamı her yerden erişime açılabilir.
- Tezim/Raporum sadece Hacettepe Üniversitesi yerleşkelerinden erişime açılabilir.
- Tezimin/Raporumun ...1... yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

Bayram ÇETİN



## ÖZET

ÇETİN, Bayram. *Geleneksel Yöntemle ve Eleme Yöntemi ile Puanlanan Çoktan Seçmeli Testlerin Psikometrik Özelliklerinin İncelenmesi*, Doktora Tezi, Ankara, 2005.

Bu araştırmada, çoktan seçmeli testlerin geleneksel yöntemle cevaplanmış formu ile cevaplayıcının tespit edebildiği çeldiricilerin elenmesi şeklinde cevaplanmış formunun (eleme puanlaması) madde ve test özellikleri Klasik Test Kuramı (KTK), Madde Tepki Kuramı (MTK) ve cevaplayıcı görüşlerine göre karşılaştırılmıştır. Araştırmada KTK'ya göre, geleneksel ve eleme puanlamaları madde ayırıcılıkları, testin göreceli etkilikleri, test güvenilirlikleri ve toplam puanları arasındaki ilişkiler karşılaştırılmıştır. MTK'ya göre, madde ve testin geleneksel ve eleme puanlamalarının göreceli etkililikleri, marjinal güvenilirlikleri ve kestirilen yetenek düzeyleri arasındaki ilişki açısından karşılaştırılmıştır. Bunun yanında, cevaplayıcıların bazı anket sorularıyla bu iki cevaplama şeklini karşılaştırmaları istenmiştir. Araştırmanın bulguları, KTK'ya göre madde ayırıcılık güçleri yönünden geleneksel puanlamanın 0.525, eleme puanlamasının 0.411 olarak belirlenmiştir. Eleme puanlamasının Cronbach  $\alpha$  güvenilirliği 0.83, geleneksel puanlamanın KR-20 güvenilirliği 0.74 olarak belirlenmiş ve eleme puanlamasının 0.01 düzeyinde anlamlı olarak daha yüksek güvenilirlik verdiğini ve uzunluk olarak daha etkili olduğunu göstermiştir. MTK karşılaştırmalarında ise, eleme puanlaması geleneksel puanlamaya göre çok daha etkili olduğu belirlenmiştir. Marjinal güvenilirlikler eleme puanlamasında 0.95, geleneksel puanlamada 0.74 olarak kestirilmiştir. Eleme puanlamasının 0.01 düzeyinde anlamlı olarak daha yüksek düzeyde güvenilirlik verdiği görülmüştür. Ayrıca, hem KTK'da hem de MTK'da kestirilen yetenek düzeyleri arasındaki ilişki yaklaşık 0.56 olarak kestirilmiştir. İki puanlama yönteminin birbirlerindeki değişkenliğin % 32'sini açıklayabildiği belirlenmiştir.

### **Anahtar Sözcükler**

Çoktan Seçmeli Testler, Eleme Puanlaması, Coombs, Kısmi Puan Modeli

## ABSTRACT

ÇETİN, Bayram. Ph.D. Dissertation, *Investigation of psychometric properties of multiple choice tests which scored with traditional method and elimination method*, Ankara, 2005.

In this research, item and test properties of form which is answered according to the traditional method and form which is answered as eliminating distracters that identified by examinee (elimination scoring) are compared within Classical Test Theory (CTT), Item Response Theory (IRT) and judgements of examinees. In respect of CTT, the item discriminations, relative length efficiency of methods, reliabilities and the correlation between total test scores estimated from traditional and elimination scoring are compared. In respect of IRT, the relative efficiency of items and test, marginal reliabilities and correlation between ability levels estimated from traditional and elimination scoring are compared. In addition, the examinees are asked to compare these two methods with some of survey questions. With respect to CTT, the findings of the research are indicated 0.525 in traditional scoring and 0.411 in elimination scoring in terms of median of the item-total correlations. Interconsistency reliability are obtained 0.83 in elimination scoring, 0.74 in traditional scoring and the difference of these reliability coefficients is significant at 0.01 level. Additionally, the elimination scoring is more efficient than traditional scoring according to length efficiency. The marginal reliabilities are estimated 0.95 in elimination scoring and 0.74 in traditional scoring in respect to IRT. The difference of marginal reliability coefficients between elimination and traditional scoring is significant at 0.01 level. Furthermore, correlation of ability levels estimated from elimination and traditional scoring is predicted as 0.56 both in CTT and IRT. The two scoring methods are explained % 32 of other's variance.

### Key Words

Multiple Choice Tests, Elimination scoring, Coombs, Partial Credit Model

<b>İÇİNDEKİLER</b>	<b>Sayfa</b>
ÖZET.....	i
ABSTRACT.....	ii
İÇİNDEKİLER.....	iii
TABLOLAR LİSTESİ.....	vii
ŞEKİLLER LİSTESİ.....	ix
ÖNSÖZ.....	xi

## BÖLÜM I

GİRİŞ.....	1
Çoktan Seçmeli Testlerde Alternatif Yöntemler.....	3
Objektif Ölçüte Göre Maddelerin Ağırlıklandırılması .....	5
Seçeneklerin Ağırlıklandırılması.....	6
Madde Yapısını Değiştirme.....	7
Cevaplama Biçimini Değiştirme.....	7
Problem Durumu.....	8
Eleme Puanlaması.....	9
Klasik Test Kuramı.....	12
KTK’da Maddelerin Seçilmesi.....	14
KTK’da Puan Verme.....	16
Güvenirlilik ve Ölçmenin Hatası.....	17
Geçerlilik ve Kestirimin Standart Hatası.....	19

Madde Tepki Kuramı.....	20
Lojistik Madde Tepki Kuramı Modelleri.....	24
Madde ve Test Bilgi Fonksiyonu.....	25
Kısmi Puan Modeli (Partial Credit Model).....	27
Genelleştirilmiş Kısmi Puan Modeli.....	29
Problem Cümlesi.....	30
Alt Problemler.....	30
Sayıtlar.....	31
İlgili Araştırmalar.....	31
<b>BÖLÜM II</b>	
<b>YÖNTEM.....</b>	<b>35</b>
Araştırmanın Türü.....	35
Verilerin Toplandığı Grup.....	35
Ölçme Araçları.....	36
Verilerin Çözümlemesi.....	38
Geleneksel Puanlama ve Eleme Puanlaması Dağılımlarının Özellikleri.....	38
Örtük Özellikler Kuramının Sayıtlarının Test Edilmesi.....	41
Tek boyutluluk.....	41
Yerel bağımsızlık.....	43
Model-veri uyumu.....	43
Alt Problemlerin Çözümlemesinde Kullanılan İstatistiksel Analizler.....	44

### BÖLÜM III

BULGULAR VE YORUMLAR.....	48
Alt Problem 1: Çoktan Seçmeli Testlerde Geleneksel Puanlama ve Eleme Puanlamasının Klasik Test Kuramına Göre Karşılaştırmalar.....	48
1.a) Madde ayıricılık güçlerinin karşılaştırılması.....	48
1.b) Belirli güvenilirliği sağlayan madde sayıları oranının belirlenmesi.....	51
1.c) Test güvenilirlikleri karşılaştırması.....	52
1.d) Toplam puanlar arasındaki ilişki düzeyinin belirlenmesi.....	53
Alt Problem 2: Çoktan Seçmeli Testlerde Geleneksel Puanlama ve Eleme Puanlamasının Örtük Özellikler Kuramına Göre Karşılaştırılması	54
2.a) Belirlenen 7 yetenek düzeyinde madde bilgi düzeylerinin göreceli etkililiğinin (relative efficiency) belirlenmesi.....	54
2.b) Marjinal güvenilirlik karşılaştırmaları.....	63
2.c) Çoktan seçmeli test maddelerinden oluşan aynı çoktan seçmeli testin kullanıldığı geleneksel ve eleme puanlaması uygulamalarından örtük özellikler kuramına dayalı olarak kestirilen yetenekler arasındaki ilişki düzeyinin belirlenmesi.....	64
Alt Problem 3: Çoktan Seçmeli Testlerde Geleneksel Puanlama ve Eleme Puanlaması Uygulamasının Karşılaştırılmasına İlişkin Öğrenci Görüşleri .....	66

### BÖLÜM IV

SONUÇ VE ÖNERİLER.....	72
Sonuçlar.....	72
Öneriler.....	76
KAYNAKÇA.....	78

## TABLOLAR

Tablo 1	Eleme Puanlamasında Elenen Seçenek Sayısına Göre Kazanılacak Puanlar	10
Tablo 2	Geleneksel ve Eleme Puanlamasına İlişkin Puan Dağılımlarının Özellikleri	39
Tablo 3	Sözel Yetenek Testinin Geleneksel ve Eleme Puanlaması Uygulamalarından Elde Edilen Madde Ayırıcılık Gücü Değerleri ve Farklarına İlişkin z Testi Sonuçları	49
Tablo 4	Geleneksel ve Eleme Puanlaması Madde Ayırıcılık Güçleri Sınıflaması	50
Tablo 5	Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Kestirilen Güvenirlilik Katsayılarının Karşılaştırılması	53
Tablo 6	Kullanılan Sözel Yetenek Testinin Geleneksel ve Eleme Puanlaması Uygulaması Yapılan Gruptan Elde Edilen Yetenek Ölçülerine İlişkin Betimsel İstatistikler	54
Tablo 7	Belirlenen Yetenek Düzeylerinde Geleneksel ve Eleme Puanlamalarından Kestirilen Madde ve Test Bilgi Düzeyleri	56
Tablo 8	Eleme Puanlaması'nın Geleneksel Puanlamaya göre Göreceli Etkiliği	58
Tablo 9	Sözel Yetenek Testinin Örtük Özellikler Kuramına Göre Geleneksel ve Eleme Uygulamalarından Kestirilen Marjinal Güvenirlilik Katsayılarının Karşılaştırılması	64

Tablo 10	Kullanılan Sözel Yetenek Testinin MTK'ya Dayalı Olarak Geleneksel ve Eleme Puanlaması Uygulamalarından Kestirilen Yetenek Ölçülerine İlişkin Betimsel İstatistikler	65
Tablo 11	Test Alan Bireylerin Geleneksel ve Eleme Uygulamalarını Karşılaştırmalarına İlişkin Görüşlerinin Frekans ve Yüzde Dağılımları	66



## ŞEKİLLER

Şekil 1	Bir parametrelî madde Karakteristik Eğrisi	24
Şekil 2	İki parametrelî madde Karakteristik Eğrisi	25
Şekil 3	Üç parametrelî madde Karakteristik Eğrisi	25
Şekil 4	Eleme Puanlaması Puan Dağılımı	40
Şekil 5	Geleneksel Puanlama Puan Dağılımı	40
Şekil 6	Eleme Puanlamasına Ait Öz Değer Grafiği	42
Şekil 7	Geleneksel Puanlamaya Ait Öz Değer Grafiği	42
Şekil 8	Geleneksel ve Eleme Uygulamalarından Kestirilen Test Bilgi Düzeyleri Grafiği	60
Şekil 9	Geleneksel ve Eleme Uygulamalarına İlişkin Madde Bilgi Grafikleri	61-62
Şekil 10	Cevaplayıcıların Cevaplama Metotlarından Hangisinin Cevaplamasının Daha Kolay Olduğuna İlişkin Görüşlerinin Dağılımları Grafiği	67
Şekil 11	Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisinin Cevaplamasının Daha Fazla Zaman Aldığına İlişkin Cevaplayıcı Görüşleri Dağılımı Grafiği	68
Şekil 12	Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisinin Başarıyı Belirlemede Daha Uygun Olduğuna İlişkin Öğrenci Görüşleri Dağılımı Grafiği	69

- Şekil 13 Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisinde Verilen Cevapların Doğruluğundan Emin Olunduğuna İlişkin Cevaplayıcı Görüşleri Dağılımı Grafiği 70
- Şekil 14 Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisini Test Alırken Tercih Edeceklerine İlişkin Cevaplayıcı Görüşleri Dağılımı Grafiği 71



## BÖLÜM I

### GİRİŞ

Eğitimde başarının, yeteneğin, becerinin, performansın vb. ölçülmesinde pek çok yol vardır. Bilişsel performansla ilgili olanlar: açık uçlu sorular, sözlü yoklamalar, kısa cevaplı testler, eşleştirme maddeleri, doğru-yanlış testleri, çoktan seçmeli testler vb. olarak sayılabilir. Bu ölçme araçlarından büyük gruplara en fazla uygulananı çoktan seçmeli testlerdir.

Çoktan seçmeli testler ülkemizde özellikle öğrenci veya personel seçme ve yerleştirme sınavlarında, ÖSYM'nin yaptığı onlarca sınavda kullanılan bir test türüdür. Ülkemizde neredeyse herkes çoktan seçmeli test maddeleriyle bir şekilde karşılaşmış ve bu yolla değerlendirilmiştir. Hatta birkaç kez karşılaşma olasılığı daha yüksektir. Çünkü çoktan seçmeli testler, büyük ölçekli değerlendirilmelerde, tek sınav türü imiş gibi algılanmaktadır.

Çoktan seçmeli testler en çok kullanılan ve belki de bilgi, yetenek ve başarının objektif olarak ölçülmesinde en iyi yöntemdir (Ben-Simon, Budescu ve Nevo, 1997). Çoktan seçmeli testlerin en iyi olarak değerlendirilmesindeki temel özellikler, objektiflik, kolay uygulanabilirliği, otomatik puanlanabilmesi ve testte deneysel kanıtlarla (ör. madde analizi) değişiklikler yapılmaya olanak sağlaması olarak sayılabilir. Kurz (1999), çoktan seçmeli testlerin geniş bir içerik örnekleme, yüksek güvenilirliği, kolay uygulanabilirliği, farklı içerikli sınavlarda kullanılabilirliği ve objektif olarak puanlanabilmesinden dolayı hem büyük sınav organizasyonlarında hem de sınıfta öğretmenler tarafından kullanımının uygun görüldüğünü belirtmektedir.

Puanlanmanın objektif yapılabilmesi, önemli değerlendirme kararları verilecek bir sınavda hayati önem taşıyan bir konudur. Objektif puanlamanın mümkün olmadığı sınavlarla yapılan değerlendirmelere olacak itirazlar büyük sorunlara yol açabilecektir. Çünkü yapılan bir yerleştirmenin geri alınması, verilen bir hakkın geri alınması anlamına gelebilir. Bu da telafisi olmayan hatalara sebep olabilir.

Çoktan seçmeli maddelerin cevaplandırılması, maddeyi okuma, anlama ve seçenekler arasından doğru cevabı seçmeyi gerektirir. Bu özellik, kısa sürede daha fazla madde cevaplanmasını sağlar. Çoktan seçmeli testlerin süre konusundaki tasarrufu, daha fazla madde uygulama imkanı sağlayacak ve ölçmeyi daha duyarlı hale getirerek güvenilirliğini artıracaktır. Madde sayısı artırılmasa bile, uygulama süresinin daha kısa olması, çoktan seçmeli bir testin diğer sınav türlerine göre daha kullanışlı olmasını sağlayacaktır.

Çoktan seçmeli testlerin uygulanmasında, cevapların kaydedildiği optik formlar kullanılması otomatik puanlama olanağı sağlar. Çoktan seçmeli testlerin geniş ölçekli test programlarında tercih edilmesinin en önemli sebeplerinden biri de bu özelliğidir. Bu özellik çoktan seçmeli testleri geniş ölçekli sınav uygulamalarında alternatifsiz duruma getirmektedir. Elde yapılan puanlamanın çok zaman alacağı, çok inceleyici kullanmanın pahalıya mal olacağı ve buna insanlardan kaynaklanan maddi hataların karşılabileceği de hesaba katıldığında, optik okuyucu yoluyla puanlamanın önemi daha iyi anlaşılabilir.

Çoktan seçmeli testlerin bu üstünlüklerinin olması, bu test türünün kusursuz olduğu anlamına gelmez. Yetenek veya başarının objektif olarak ölçülmesinde en iyi yol olarak çoktan seçmeli test tipinin genel kabulü, onun her yönden mükemmel olduğu sonucuna götürmez (Coombs, Milholland ve Womer, 1956). Çoktan seçmeli testlerin de önemli sınırlılıkları vardır. Bu sınırlılıklardan biri Çoktan seçmeli testlerin şans başarısına izin vermesi, bir diğeri de madde temelinde bireyler arasındaki ayrımı iki kategoride yapması olarak sayılabilir.

Çoktan seçmeli testler madde kökü ve seçeneklerden oluşmaktadır. Madde kökü asıl sorunun sorulduğu bölümdür, seçeneklerde ise anahtarlanmış bir doğru cevap ve çeldiriciler bulunur. Madde ile ölçülen özellik konusunda tam bilgiye sahip bireylerin doğru cevabı seçmesi, yanlış bilen bireylerin çeldiricilere yönelmesi ve hiç bilgiye sahip olmayanların ise boş bırakması beklenir. Fakat pratikte bu böyle olmayabilir. Bazı bireyler madde konusunda bilgiye sahip olmadıkları halde maddeyi cevaplandırma

yoluna gidebilirler. Bu da testin güvenilirliğini, geçerliğini ve maddenin psikometrik özelliklerini etkileyecektir.

Şans başarısı çoktan seçmeli testlerin uygulanmasında ve puanlanmasında karşılaşılan en büyük sorunlardan biridir. Şans başarısını düşürmek için çoktan seçmeli testlerle ilgili birçok çalışma yapılmıştır. Genelde şans başarısı kısmi bilgiden kazanılan puanlarla karıştırılır. Bir birey bazı seçenekleri eleyebiliyor, kalan seçenekler arasından tahminle doğru cevabı bulma yoluna gidiyorsa, burada kör tahminden kazanılan puandan farklı bir olgu vardır. Bu olgu önceleri çok iyi anlaşılmamış ve kısmi bilgiyi yansıtmalarına rağmen tahminden dolayı kazanılan puanlar “kural dışı” kazanılmış olarak görülmüştür (Frery, 2000). Fakat kısmi bilgiye sahip bireyler bazı seçenekleri eledikten sonra, kalan seçenekler arasından tahminle bir cevap seçme yoluyla, kısmi bilgilerinin gösterme şansı bulabilmektedirler. Eğer kısmi bilgiye sahip bireyler maddeyi boş bırakırlarsa, bu bilgileri hiç değerlendirilmemiş olacaktır. Bu da tahmini kullanmanın olumlu bir yönüdür. Fakat bu durum test istatistiklerini olumlu etkilese bile, madde istatistiklerini olumsuz etkileyecektir. Çünkü bazı maddelerde şans başarısına dönüşürken, bazı maddelerde şans başarısızlığına dönüşecektir.

Çoktan seçmeli testlerin madde ölçeğinde kısmi bilgiye duyarsız olması, belirtildiği gibi testi cevaplayanların cevabı tahmin etmeye yönelmelerini artırmaktadır. Çoktan seçmeli testler, sınananlara kısmi bilgilerinin gösterme olanağı verebilseydi, bazı seçenekleri eleyenlerin tahminle doğru cevabı aramalarına gerek bırakmayacak ve kör tahmin kullanımının cezalandırılması (puan kesme şeklinde olabilir) mümkün olabilecekti. Görüldüğü gibi çoktan seçmeli testlerin bu iki sınırlılığı birbirine bağlıdır. Kısmi bilgiye duyarsız madde puanı, tahminle doğru cevabı bulmaya çalışma ihtiyacını ortaya çıkarmaktadır.

## **ÇOKTAN SEÇMELİ TESTLERDE ALTERNATİF YÖNTEMLER**

Çoktan seçmeli testlerde tahminle cevaplamayı düşürmeyi ve kısmi bilginin ölçülmesini amaçlayan bir çok çalışma yapılmıştır. Bunlardan bazıları sadece kısmi bilginin ölçülmesine dayanırken, bazıları tahmini düşürmeyi ve bazıları da hem kısmi bilgiyi

ölçmeyi hem de tahmini azaltmayı amaçlamaktadır. Örneğin tahmin için düzeltme formülü kullanılması yalnızca tahminden gelen başarıyı engellemeyi amaçlamaktadır. Çoktan seçmeli testlerde tahminle cevaplamayı düşürmeyi ve kısmi bilgiyi ölçmeyi amaçlayan alternatif uygulama çalışmalarını Lord ve Novick (1968) üç farklı sınıfa ayırmaktadır. Bunlar:

1. Yeni madde yapısı ve/veya cevap metodu
2. Farklı madde puanlama formülü / metotları (seçenek ağırlıklandırması)
3. Madde ağırlıklandırma formülü / metotları.

Bunun yanında çoktan seçmeli testlerde alternatif uygulamaları farklı sınıflara ayıranlar da olmuştur. Örneğin, Ben-Simon, Budescu ve Nevo (1997) başka önemli değişkenleri de göz önüne alarak bir sınıflama yapmıştır. Bu değişkenler:

- Metodun temel hedefi: tahmini düşürmek, kısmi bilgiyi ölçmek veya cevaplardan daha fazla bilgi elde etmek.
- Cevap metodunun dayandığı ölçme şekli: objektif veya subjektif (sınavı alan bireyin yargısı gibi).
- Sınananın cevap süreci, puanlama modeli konusundaki bilgi düzeyi ve bu süreçteki kontrolü.

Ben-Simon, Budescu ve Nevo (1997) alternatif uygulamaları dört temel sınıfa ayırmaktadır ve bu temel sınıflar aşağıdaki şekildedir:

- a) Objektif Ölçüte Göre Maddelerin Ağırlıklandırılması
- b) Seçeneklerin Ağırlıklandırması
- c) Madde Yapısını Değiştirme
- d) Cevaplama Biçimini Değiştirme

### **Objektif Ölçüte Göre Maddelerin Ağırlıklandırılması**

Bu metotta temel düşünce, maddeleri psikometrik özelliklerine göre değerlendirmektir. Maddelerin psikometrik özelliklerine ve/veya madde güçlüğüne göre ağırlıklandırılmasına dayanır. Eğer madde psikometrik açıdan iyi bir madde ise daha yüksek ağırlık alacaktır. Bir diğeri ise madde güçlüğüne göre ağırlıklandırmak; daha zor maddelere kolay maddelere göre daha yüksek puan verilmesi şeklinde olur. Bu metotlarda maddelerin ağırlıkları olarak en çok kullanılan özellikler madde güçlüğü, madde geçerliği, tanılama yeteneği, madde varyansı veya uzmanlarca belirlenen ağırlıklardır (Ben-Simon, Budescu ve Nevo, 1997).

Bu metotta sınav uygulanan bireyler nasıl puanlandığı konusunda bilgi sahibi olmadıklarından süreci kontrol edemezler. Bireylere bunu açıklamak da zor olabilir. Ayrıca madde ağırlıklandırma metotları, uzman kanılarına göre ağırlıklandırma dışında, örneklem bağımlıdır. Bu açılardan madde ağırlıklandırma metodunun kullanılmasında sakıncalar vardır. Gulliksen madde ağırlıklandırma metotlarının, testin birbiriyle yüksek ilişkili 10 maddeyi geçmesi durumunda, geleneksel yaklaşıma üstünlüğünün kalmadığını söylemektedir (Akt: Ben-Simon, Budescu ve Nevo, 1997). Madde sayısı 10'dan daha az olan testlerde geleneksel yaklaşıma göre daha iyi sonuçlar verdiği anlamına gelmektedir.

Madde Tepki Kuramı kullanılarak yapılan puanlama ve işlemler de bu sınıfta sayılabilir. Çünkü bu metotta maddelerin güçlüklerine ve ayırıcılıklarına göre ağırlıklandırılarak yetenek düzeylerinin kestirilmesi söz konusudur.

Bu sınıftaki özgün metotlardan biri de şartlı olasılığa dayalı olarak puanlama yapan "Şartlı Bilgi Modeli"dir. Bu metot farklı cevap türleri (doğru, yanlış veya boş) üzerinde şartlı bilginin olasılık dağılımı üzerine inşa edilmiştir. Doğru cevapların olduğu kadar, yanlış cevapların ve boş bırakılan maddelerin de bireyin bilgi düzeyi konusunda bilgi içerdiği sayılıdır. Şartlı Bilgi Modeli, tahminle kazanılan puanı, kısmi bilgiyi, yanlış bilgiyi ve boş cevapları da hesaba katar (Strashny, 2002). Bireyin maddeye

verdiği cevaba göre maddeyi doğru cevaplama olasılığı, yanlış cevaplama olasılığı ve boş bırakma olasılığı hesaplanarak puanı belirlenir.

### **Seçeneklerin Ağırlıklandırması**

Genel olarak deneysel ağırlıklandırma ve önsel ağırlıklandırma olmak üzere iki farklı madde seçeneği ağırlıklandırma metodu vardır (Echternacht 1976). Deneysel ağırlıklandırma, test uygulandıktan sonra seçeneklerin dağılımına göre; önsel ağırlıklandırma ise test uygulanmadan önce uzman kanılarına dayanarak yapılan ağırlıklandırma yöntemidir.

Deneysel ağırlıklandırma maddelerin ön uygulanması sonucunda elde edilen deneysel kanıtlar kullanılarak yapılır. Bu tür ağırlıklandırmada temel düşünce bir çoktan seçmeli test maddesinin çeldiricilerinin doğru cevaba yakınlığı ve kendisine çektiği bireylerin yetenek dağılımlarına göre puan almasıdır. Yani daha üst düzeydeki bireylerin seçtiği yanlış seçenek daha alt düzeydeki bireylerin seçtiği yanlış seçenekten daha az önemli bir hatadır. Çoktan seçmeli bir test maddesinde, herhangi bir çeldiriciyi seçen bireylerin puan dağılımı diğer çeldiricilerden herhangi bir diğerini seçenlerin puan dağılımından farklı olacaktır (Bock, 1972). İki çeldiriciden üst yetenek grubu tarafından seçilen çeldiriciyi işaretlemek daha küçük bir yanlış olarak görülmektedir. Madde Tepki Kuramı'nda Bock'un Nominal Response Modeli ve Thissen'in Multiple Choice Modeli bu düşünce üzerine kurulmuştur. Yanlış cevapların da bireylerin yetenek düzeyleri hakkında bilgi içerebileceği düşüncesi vardır.

Önsel ağırlıklandırmada ise daha çok uzman kanılarına dayanarak ağırlıklandırma yapılır. Bu tür maddelerin seçeneklerinde tek doğru cevap değil, en doğrudan-en az doğruya doğru bir sıralanma vardır. Seçeneklerin ağırlıkları uzmanların veya testi hazırlayan kişilerin yaptıkları doğruluk sıralamasına göre belirlenir.

Seçenek ağırlıklandırma yönteminde de öğrenciler puanlama sürecini bilmezler ve buna göre bir cevap yöntemi geliştiremezler. Bu puanlama metodunun, sınav alanları rahatsız

edebilecek yönleri vardır, çünkü bir birey kendisi gibi maddeyi yanlış cevaplayan bireyden daha az puan alabilmektedir (Echternacht 1976). Bunun sınavı alan bireylere açıklanması çok zordur.

### **Madde Yapısını Değiştirme**

Bu sınıftaki metotlarda çoktan seçmeli maddelerin yapısı farklılaştırılmıştır. Birden çok doğru cevabı olan maddeler veya her bir seçeneğin bir doğru-yanlış tipi madde gibi sunulduğu örnekleri vardır. Bu türlerde puanlamalar da farklı şekillerde yapılabilmektedir. Genelde doğru cevaplanan seçenek sayısına göre puanlanır. Birden çok doğru cevabı olan maddeleri, Cronbach “Multiple Multiple Choice”, Dressel ve Schmidt “Multiple Answer” ve Poplun ve Omar ise “Multiple Mark” olarak isimlendirmişlerdir (Poplun ve Omar 1997). Birden çok doğru cevabı olan maddelerde de farklı puanlama metotları kullanılabilir. Bunlar: (1) Ya hep ya hiç temelinde puanlama (tüm doğru cevaplar seçilmiş ve hiçbir çeldirici seçilmemişse 1 puan, diğer durumlarda sıfır puan) ve (2) her bir seçeneği bağımsız puanlama (her doğru seçime 1 puan) (Xiaoying and Sireci, 2004).

### **Cevaplama Biçimini Değiştirme**

Bu sınıftaki uygulamalar çoktan seçmeli maddelerin yapısını değil, cevaplama biçimini değiştirir. Cevapları kaydetme biçimi değişebileceği gibi, cevapların yanında, cevapların hangi güven düzeyinde verildiği gibi ek bilgiler içerebilirler. Ben-Simon, Budescu ve Nevo (1997) bu tür uygulamaları aşağıdaki şekilde sınıflandırmıştır:

1. Seçeneklerin tam veya kısmi sıralanması
2. Güven ağırlıklandırması – en fazla doğru cevap olma olasılığı olan seçeneği seçme ve doğruluğundan emin olma derecesini belirtme
3. Olasılık ağırlıklandırması – doğru cevap olma olasılığına göre her bir seçeneğe ağırlık verilmesi

4. Kabul edilebilir cevapların bir alt grubunu seçme
5. Kabul edilemeyecek tüm seçenekleri reddetme.

Bunlara ek olarak özgün bir uygulama “Doğru Cevaba Bulana Kadar Yanıtlama” (Answer Until Correct) uygulamasıdır. Bu uygulamada, bireye doğru cevabı bulana kadar seçme olanağı verilir, sürekli dönüt sağlanır. Doğru cevabı bulana kadar yanıtlama'nın öğrenmeleri güçlendirmede ve yanlış bilgileri düzeltmede önemli katkılar sağladığı bildirilmiştir (Wood, 1977).

Yukarıda maddeler halinde verilen metotların çoğunda şans başarısını düşürme ve kısmi bilgiyi ölçme iddiası vardır. Kör tahminin beklenen değeri sıfır puana denk gelir, bu da yanlış bilgiye eksi puan verme şeklinde sağlanır. Yanlış bilginin düzeyi arttıkça eksi puan da artar. Eğer bir birey doğru bir cevabı yanlış olarak belirliyorsa ve yanlış bir cevabı da doğru olarak belirliyorsa, bu birey ölçülen konuda yanlış bilgiye sahiptir. Yanlış bilgi de bilginin yokluğundan farklı değerlendirilmelidir ve eksi puan almalıdır düşüncesi vardır. Bu metotlarda hangi tahmin stratejisi kullanılırsa kullanılsın şans başarısının beklenen değeri sıfır olur (Gibbons, Olkin ve Sobel, 1979; Ben-Simon, Budescu ve Nevo, 1997).

## **PROBLEM DURUMU**

Çoktan seçmeli testlerde kısmi bilginin ölçülmesine izin veren ve şans başarısını düşüren bir yöntem bulunabilirse ve bu yöntemle yapılan ölçümler gerek Klasik Test Kuramına (KTK) göre kestirilen madde istatistikleri ve güvenilirlik anlamında, gerekse Madde Tepki Kuramı'na (MTK) göre kestirilen madde parametreleri ve madde bilgi düzeyleri bakımından daha iyi sonuçlar verirse ölçmenin daha doğru sonuçlarla yapılmasına katkı sağlayacağı düşünülmüştür. Bu amaçla düşünülen, çoktan seçmeli testlerde çeldiricilerin elenmesi yoluyla cevapların kaydedilmesi ve elenen çeldirici sayısına göre puanlama yapılması yönteminin literatürde “Eleme Testi” (Elimination Testing) (Ben-Simon, Budescu ve Nevo, 1997) veya “Eleme Puanlaması” (Elimination

Scoring) (Jaradat ve Tollfson, 1988) olarak isimlendirildiği görülmüştür. Bu çalışmada “Eleme Puanlaması” adının kullanılması tercih edilmiştir.

Bu metot ilk defa 1953 yılında Coombs tarafından önerilmiştir (Jaradat ve Tollefson, 1988) ve daha sonra Coombs, Milholland ve Womer (1956) tarafından uygulanarak geleneksel puanlama ile karşılaştırılmıştır.

### **Eleme Puanlaması**

Bilindiği gibi çoktan seçmeli testlerin geleneksel uygulanmasında doğru cevap seçildiğinde 1 puan, çeldirici seçildiğinde ise 0 puan verilmektedir. Bir puan tam bilgiyi, sıfır puan ise bilginin yokluğunu temsil eder (Coombs, Milholland ve Womer, 1956). Cevaplayıcının kısmi bilgiye sahip olması durumunda ise ya 1 puana ya da 0 (sıfır) puan alabilecektir. Yani geleneksel çoktan seçmeli test uygulamasının tam bilgi, kısmi bilgi ve yanlış bilgi arasındaki ayrımı yapamama gibi bir sınırlılığı vardır (Coombs, Milholland ve Womer, 1956). Bu sınırlılık bireyleri tahminle doğru cevabı aramaya yöneltir. Tahmini kullanmanın tek sebebi bu sınırlılık olmasa bile büyük bir etken olarak karşımıza çıkar. Hatta test programını hazırlayanlar da kısmi bilgilerin bu şekilde gözlenebileceğini umdukları için test yönergesine “eğer bazı seçenekleri eleyebiliyorsanız, kalan seçenekler arasından birini seçmeniz sizin faydanıza olacaktır” benzeri ifadeler ekleme ihtiyacı duymaktadırlar (ÖSYM sınav yönergelerine bakılabilir).

Çoktan seçmeli testlerin geleneksel uygulamasının böyle bir sınırlılığının bulunması, bu sınırlılığı ortadan kaldırabilmeye olası çözüm üretebilecek yeni uygulamalara ihtiyaç yaratmıştır. Bu amaçla ortaya çıkan eleme puanlaması (EP), her maddede farklı bilgi düzeylerindeki bireyleri ayırt edebilmesi ve şans başarısına olan ihtiyacı kaldırması yönü ile geleneksel metoda üstünlük sağlayabilecek niteliklere sahiptir.

EP yöntemi, sorunun doğru cevabını bilmeyen bireylerin bazı seçeneklerin yanlış olduğunu bilebilecekleri düşüncesinden yola çıkmıştır (Coombs, Milholland ve Womer,

1956). Böylece kısmi bilgiyi de puanlayabilmektedir. Birey doğru cevabın doğru olduğunu, çeldiricilerin de yanlış cevap olduğunu biliyorsa tam bilgiye sahiptir. Eğer birey doğru cevabın yanlış olduğunu biliyor ve buna göre maddeyi cevaplandırıyorsa yanlış bilgiye sahiptir. Doğru cevabı yanlış olarak bilmesine rağmen, bazı çeldiricilerin de yanlış olduğunu biliyorsa, bu durum kısmi yanlış bilgi olarak düşünülmüştür (Coombs, Milholland ve Womer, 1956).

EP uygulaması, bireylerden doğru cevabı değil de tüm yanlış cevapları işaretlemeleri şeklinde yönlendirilerek yapılır (Jaradat ve Tollefson, 1988). Bireyler yanlış olduğunu belirleyebildikleri tüm çeldiricileri işaretlerler ve tespit edebildikleri çeldirici sayısına göre puan alırlar. Boş bıraktıklarında veya tüm seçenekleri elediklerinde, bilginin yokluğunu gösterdiği varsayılarak 0 (sıfır) puan alırlar. Maddelerin seçenek sayısına göre puan sayısı değişir. Seçenek sayısı  $k$  olan bir maddede  $-(k - 1)$  ile  $(k - 1)$  arasında değişen puan alınabilir. Eğer madde 5 seçenekli ise  $-4$  ile  $+4$  arasında değişen puan almak olasıdır. Bir birey sadece doğru cevabı çeldirici olarak belirleyip elemiş ise  $-4$  puan alacaktır, bunun yanında bazı çeldiricileri de eleyebilmişse eleyebildiği çeldirici sayısına göre  $-3$ ,  $-2$  veya  $-1$  puan alabilir. Eğer birey sadece bir çeldiriciyi eleyebilmişse  $+1$  puan, eleyebildiği çeldirici sayısına göre  $+4$  puana kadar alabilir. Beş seçenekli bir test maddesinde, elenen seçenek sayısına göre alınacak puanlar, doğru cevabı içermesi ve içermemesi durumuna göre Tablo 1’de verilmiştir.

**Tablo 1**

**Eleme Puanlamasında Elenen Seçenek Sayısına Göre Kazanılacak Puanlar**

Elenen Seçenek Sayısı	Elenen Seçeneklerin Doğru Cevabı İçermesi Durumu	Doğru Cevabı İçermemesi Durumu
0	–	0
1	– 4	1
2	– 3	2
3	– 2	3
4	– 1	4
5	0	–

Bu metotta tahmin için düzeltme formülü kullanmaya gerek yoktur. Bir birey dört seçenekli bir test maddesinde, iki seçeneğin yanlış olduğunu bilir ve kalan iki seçenekten birini daha şansla elemek isterse, bir puan kazanmaya karşı üç puan

kaybetmeyi %50 olasılıkla göze alması gerekir. Bu durum cevaplayıcı için kazançlı bir kumar değildir (Coombs, Milholland ve Womer, 1956). Madde hakkında hiç bilgi sahibi olunmadığında, kör tahmin kullanılırsa veya başka tahmin stratejileri geliştirilerek madde cevaplandığında beklenen şans başarısı sıfır olacaktır. Bu durum Gibbons, Olkin ve Sobel (1979)'in Altgrup Seçme tekniğinde verdiği örnekle şöyle açıklanabilir: bir bireyin beş seçenekli bir soruda, iki seçeneği tesadüfi olarak ellediğini düşünelim. Bu durumda, seçilen iki seçeneğin doğru cevabı içermeme olasılığı 3/5, doğru cevabı içermeme olasılığı 2/5'dir. Sonuç olarak, iki seçeneğin (Tablo 1'i kullanarak) doğruyu içermesi durumunda 3 puan kaybedileceği ve içermemesi durumunda 2 puan kazanılacağı düşünülerek beklenen puan:

$$E(P) = -3 (2/5) + 2(3/5) = 0 \text{ olacaktır.}$$

Çoktan seçmeli testlerde EP kullanılması, geleneksel doğru cevaba 1 puan yanlış cevaba 0 puan veren uygulamadan, madde ölçeğinde daha ayırcılığı yüksek sonuçlar vereceği düşünülmektedir. Ayrıca madde düzeyinde geleneksel uygulamada kaybolan kısmi bilgi düzeylerine ait bilgilerin EP'de ortaya çıkarılabileceği ve psikometrik açıdan daha olumlu sonuçlar vereceği umulmakta, daha duyarlı madde puanlarından daha doğru madde istatistikleri hesaplanabileceğinden, testin güvenilirliğinin artacağı beklenmektedir. Puanlamada geleneksel uygulamada iki kategorili (dichotomous) ve eleme puanlamasında çok kategorili (polytomous) veriler elde edilmektedir. Bu sebeple daha duyarlı ölçme sonuçları elde edileceği düşünülmüştür.

Test teorilerinin içeriği temel olarak (1) verilen bir durumda elde edilen ölçme sonuçlarını etkileyen problemleri kestirmek (2) bu problemleri en aza indirmek için metotlar geliştirmekle ilgilenir (Crocker ve Algina, 1986). KTK ve MTK temelinde çoktan seçmeli testlerin geleneksel ve eleme puanlamasından madde ve test istatistiklerinin karşılaştırıldığı bu çalışmanın yapılmasında, çoktan seçmeli testlerin problemleri olan şans başarısı ve kısmi bilgiye duyarsız olma durumlarının kontrol edilip edilemeyeceğini belirlemek amaçlanmıştır. MTK'da eleme puanlamasının

analizinde, çok kategorili veri elde edileceğinden, çok kategorili ve kısmi bilginin ölçülmesinde uygun bir model olan Kısmi Puan (Partial Credit) modeli kullanılmıştır.

## KLASİK TEST KURAMI

Bireyler üzerinde değişken olan ve bireylerin özelliklerini belirleyen yetenek, başarı, beceri, performans gibi kelimeler günlük dilde kullanıldığı anlamlarından daha dar anlamlara sahiptir. Bu kelimeler daha dar ve açık biçimde, sözel ve anlamsal olarak tanımlanarak teorik yapıları belirlenir (Lord ve Novick, 1968). Sözel ve anlamsal olarak tanımlanan bu yapılar matematiksel modellerle açıklanmaya çalışılır. Matematiksel modellerle, kuramsal olarak tanımlanmış yapıların açıklanması amaçlanır ve teorik yapıyı sağlamasına kanıtlar aranır. Tanımlanan psikolojik yapıların belirlenmesi amacıyla oluşturulan testlerden elde edilen verilerle teorik yapının uyum içinde olup olmadığını kontrol etme gereği vardır (Gulliksen, 1950).

Bu amaçla geliştirilen ve bu anlamda ilk olduğu için Klasik Test Kuramı olarak adlandırılan Gerçek Puan kuramıdır. KTK en basit haliyle aşağıdaki şekilde formüle edilir:

$$X_i = T_i + E_i \quad (1)$$

burada  $X_i$ : i bireyinin gözlenen test puanını

$T_i$ : i bireyinin gözlenemeyen ve varsayılan gerçek puanını

$E_i$ : i bireyinin puanına karışan hata miktarını göstermektedir.

Gerçek puan kavramı kuramsal bir kavramdır. Gerçekte böyle bir puan var mı bilinmiyor, fakat olduğu varsayılmaktadır. Gerçek puan bir bireyin performansını etkileyebilecek tüm hata kaynakları kontrol edildiğinde, bireyin kazanacağı puan olarak düşünülmüştür. Aynı özelliği ölçen testler için bireyin sahip olduğu gerçek puan testten teste değişmez (Gulliksen, 1950). Hata puanı ise bireyin puanının değişen kısmıdır, testten teste, durumdan duruma, bir çok etkiden dolayı değiştiği varsayılan kısımdır. Bir

birey bazen bildiği halde bir maddeyi doğru cevaplayamadığı gibi, bazen de bilmediği halde maddeyi doğru cevaplayabilir. Bu durumda bireyin aldığı puan hata puanıdır. Buradaki hata türü de tesadüfi, yani kaynağı ve yönü belli olmayan hatadır.

KTK gözlenen puanların birer parçası olarak kabul edilen gerçek puan (T) ve hata puanı (E)'ni tanımlamak ve aralarındaki ilişkileri açıklamak için bazı sayıltılar oluşturmuştur. Bu sayıltıları yukarıdaki iki bilinmeyenli eşitliği açıklamak için oluşturmak zorundadır. Çünkü bu eşitlikte belli olan sadece gözlenen puandır (X). Ayrıca testle ilgilenilen özellik bakımından tüm bireyler kadar gözlenen puan, gerçek puan ve hata puanı vardır. KTK'da oluşturulan sayıltıları aşağıdaki şekilde sayabiliriz:

$$\mu_E = 0 \quad (2)$$

Bu ifade tesadüfi hataların ortalamasının sıfır olduğunu göstermektedir. Bu ortalama tüm bireylerin hatasının ortalaması olabileceği gibi, bir birey için tüm ölçmelerin hatası da olabilir. Bu eşitlik sadece bir sayıltı olarak görülmemeli, tesadüfi hataların tanımının bir parçası olarak da görülmelidir (Gulliksen, 1950).

$$r_{TE} = 0 \quad (3)$$

Yukarıdaki ifade de gerçek puanlar ile hata puanları arasında ilişkinin olmadığını göstermektedir. Tesadüfi hatalar bazı bireylerin puanına yüksek oranda, bazı bireylere düşük oranda, bazılarında artı yönde veya bazılarında eksi yönde karışabilir. Bu hataların puanlara karışma miktarı ve yönü hiçbir şekilde gerçek puanların büyüklüğüyle ilgili değildir. Tesadüfi hataların tanımı gereği bir sistematığı olmadığı düşüncesinden kaynaklanır.

$$r_{E_1E_2} = 0 \quad (4)$$

Yukarıdaki eşitlik de aynı bireyler üzerinde birbirine paralel iki teste ilişkin hata puanları arasında ilişkinin olmayacağı sayıltısını göstermektedir. Bu sayıltı da tesadüfi

hataların dođaları geređi bir sistematiđi olmadıđı dűşüncesine dayanır ve aynı bireyin birinci test puanına karışan tesadűfi hatanın aynı miktarda ve aynı yönde karışabileceđini gösterebilecek bir nedenin olmamasından kaynaklanır.

Test kuramları testlerin oluřturulması, çözümlenmesi ve sonuçların yorumlanmasında ařađıdaki beř farklı probleme cevap arar (Gulliksen, 1950).

1. Test maddelerini yazma ve seçme.
2. Her bir bireye bir puan verme.
3. Test puanlarının dođruluk düzeyini (güvenirlik veya ölçmenin hatası) belirleme.
4. Test puanlarının kestiricilik gücünü (geçerlik veya kestirimin hatası) belirleme.
5. Sonuçları, başka testlerle veya başka grupların test sonuçlarıyla karşılaştırma. Bu karşılařtırmaları yaparken, test uzunluđu ve grup dađılımlarının göz önüne alınması gereklidir.

### **KTK'da Maddelerin Seçilmesi**

KTK'da maddelerin seçilmesinde kullanılan ölçütler genel olarak madde güçlüđu, madde ayırdediciliđi (madde geçerliliđi) ve madde güvenilirliđi olarak sayılabilir. Bu ölçütler göz önüne alınarak daha iyi psikometrik özelliklere sahip maddeler tercih edilmek suretiyle nihai teste alınacak maddeler belirlenir.

*Madde Güçlüđu.* KTK'da madde güçlüđu kavramı maddelerin grupta dođru cevaplanma oranıyla ilgili bir kavramdır. Maddeyi dođru cevaplayanların gruba oranlanması şeklinde hesaplanan çok basit bir istatistiktir. Yani madde puanları ortalamasıdır diyebiliriz. Matematiksel olarak ifade edersek:

$$P_j = \frac{n_D}{N} \quad (5)$$

burada  $n_D$  maddeyi doğru cevaplayanların sayısını,  $N$  ise cevaplayıcı sayısını göstermektedir.

Madde güçlüğü yönünden bir madde bir grupta en ideal olarak madde varyansının en yüksek olduğu zaman en iyi ayırıcılığa ulaşacağı düşüncesiyle testlerde genellikle 0.50 güçlükte olan maddeler tercih edilir. Maddeler arasındaki korelasyonun sabit olduğu varsayılırsa, madde güçlüğü 0.50 olduğunda test puanları varyansı en yüksek olacaktır (Crocker ve Algina, 1986). Test puanları varyansının yüksek olması güvenilirliği artıracaktır (Gulliksen, 1950; Lord ve Novick, 1968; Baykul, 2000). Fakat uçta bulunan bireyler arasında da iyi ayırım yapabilecek maddelerin bulunması özellikle seçme sınavlarında ve bireyler arası farkın önemli olduğu sınavlarda istenen bir durumdur. Bu sebeplerden dolayı bir testte madde güçlüklerinin ortalaması 0.50'ye yaklaşması istenen bir durumdur diyebiliriz. Bazı özel durumlarda bu geçerli olmayabilir, çünkü çok fazla kişinin katıldığı, fakat çok az kişinin seçileceği seçme sınavlarında genellikle madde güçlüğü düşük yani zor maddeler tercih edilmelidir.

*Madde Ayırdediciliği.* KTK'da madde seçilmesinde çok önemli bir ölçüt de madde ayırıcılık indeksidir. KTK'da madde ayırıcılık indeksini belirlemede farklı yollar vardır. Madde ayırıcılık gücünü belirlemek için alt-üst grup ortalamaları farkı ya da korelasyonel yöntemler kullanılır. Madde puanları iki kategorili olduğunda, her madde için alt ve üst grup ortalamalarının farkına dayalı olarak madde ayırıcılık gücü elde edilir. Maddelerin sürekli kabul edilebilecek bir şekilde puanlanması durumunda, madde ayırıcılığı üst grup – alt grup puan ortalamalarının farkı test edilerek hesaplanır. Fark üst grup lehine anlamlı bulunursa, madde ayırdedicidir denir. Korelasyonel yaklaşımda madde puanları ile test puanları arasında korelasyona bakılır. Madde puanları iki kategorili ise çiftserili veya nokta çiftserili korelasyon hesaplanır. Madde puanları çok kategorili ise polyserial korelasyon, sürekli kabul edilebilecek bir ölçekte ise Pearson Momentler çarpımı korelasyon katsayısı kullanılır.

Maddelerin ayırıcılık güçlerini hesaplariken, dış bir ölçüt puanı kullanmak daha doğrudur. Çünkü o teste ait toplam puanlar ölçüt alındığında, madde puanını da içeren bir ölçüt seçilmiş olur. Madde puanını da içeren toplam puanların ölçüt alınması sonucu yanlış etkileyebilir. Fakat dış bir ölçüt bulmak her zaman için mümkün olmayabilir. Bu gibi uygulamadan kaynaklanan sebeplerden, dış ölçüt yerine toplam test puanları ölçüt olarak alınabilmektedir.

*Madde Güvenirliği.* Madde seçilmesinde diğer bir ölçüt ise madde güvenirligidir. Madde güvenirligi, madde ayırdediciligi ile madde standart sapmasının çarpımı ile hesaplanmaktadır. Madde güvenirligi madde puanının deęişkenliginin bir fonksiyonudur (Crocker ve Algina, 1986).

### **KTK'da Puan Verme**

KTK'da puan verme çok basit bir şekilde yapılmaktadır. Bu basitlik kuramın üstünlüğü olarak sayılabilir. KTK'ya dayalı yöntemlerin anlaşılması, uygulanması ve savunulması kolaydır. Test puanı madde puanlarının toplamından ibarettir. İki kategorili puanlanan maddeler için ise test puanı sadece doğru cevaplanan madde sayısına eşittir.

Yukarıda madde puanlarının toplamı olarak verilen test puanları gözlenen test puanlarını ifade eder. KTK bu gözlenen puanları eşitlik 1'de verildiği gibi gerçek puan ve hata puanının toplamı olarak iki bileşene ayırır. KTK'nın gerçek puan sayılısı sonsuz sayıda testin ortalama puanı olarak verilir. K ölçme sayısını göstermek üzere K sonsuza giderken paralel testlerin puanlarının ortalaması bireyin gerçek puanını verir (Gulliksen, 1950). Bu tanımın matematiksel ifadesi aşağıdaki gibidir:

$$T_i = \lim_{K \rightarrow \infty} \frac{\sum_{g=1}^K X_{ig}}{K} \quad (6)$$

bu eşitlikte  $T_i$  : i bireyinin gerçek puanını

$K$  : madde sayısını

$X_{ig}$ : i bireyinin g testinden aldığı puanı göstermektedir.

Buradaki paralel testler her bir madde olarak düşünülebilir, madde sayısı arttıkça gerçek puana yaklaşıldığı sayılıdır. Ölçme sayısı sonsuza yaklaştığında, eşitlik 2’de verilen sayılı gereği tesadüfi hataların ortalamasının sıfıra yaklaşacak ve gerçek puan elde edilecektir. Ölçme sayısının sonsuza ulaşması mümkün değildir, fakat testteki maddelerin mümkün olduğunca artırılması, gerçek puana yaklaşılmasını sağlayacaktır.

### Güvenirlilik ve Ölçmenin Hatası

KTK’da güvenirlilik kavramı *gerçek puanların, gözlenen puanlar* üzerindeki regresyonunun *belirtme katsayısı* olarak tanımlanmıştır (Lord ve Novick, 1968), fakat gerçek puanlar doğrudan gözlenemediği için bunun başka pratik tanımları da yapılmıştır. Bu belirtme katsayısı gerçek puan varyansının gözlenen puan varyansına oranına eşittir ve eşitlik 7’de gösterilmiştir.

$$\rho_{XT}^2 = \sigma_T^2 / \sigma_X^2 \quad (7)$$

Fakat bu eşitlikteki gözlenen puan varyansı dışındaki ifadeler bilinmediği için eşitlik 8 türetilmiştir, böylece daha uygulanabilir bir güvenirlilik katsayısı hesaplama eşitliği elde edilebilmiştir. Bu pratik katsayı bir testin kendisine paralel olan bir başka testle ilişkisidir. Bu katsayı bir korelasyon katsayısı olmasına rağmen eksi değer almaz, sıfır ile bir arasında değer alır.

$$\rho_{XX'} = \sigma_T^2 / \sigma_X^2 = \rho_{XT}^2 \quad (8)$$

KTK’da bu uygulanması kolay olan güvenirlilik eşitliğinden başka güvenirlilik ifadeleri de türetilmiştir. Bu ifadeler güvenirliliğe yüklenen anlamlara göre yorumlanmaktadır ve uygulamada hangi açıdan bilgi toplanacağıyla ilgili olarak bunlardan birinin veya bir kaçının hesaplanması gerekebilir.

Güvenirlilik test puanlarının kararlılığı, tutarlılığı, duyarlılığı ve tekrar elde edilebilirliği anlamlarını içerir (Baykul, 2000). Bu anlamların her birine göre güvenirlilik hesaplama

yöntemleri değişmektedir. Ayrıca uygulama sayısına göre de farklı güvenilirlik hesaplama yöntemleri vardır. Güvenirlik hesaplamada iki uygulamaya dayanan yöntemler olarak test puanlarının kararlılığı anlamında *test – tekrar test* yöntemi, test puanlarının tutarlılığı anlamında *paralel formlar* yöntemleri sayılabilir. Crocker ve Algina (1986) ek olarak yukarıdaki iki yöntemin birleştirilmesinden *alternatif formlarla test – tekrar test* yöntemini de saymışlardır ve bu yöntemin hem kararlılık hemde tutarlılık anlamında yorumlanabileceğini belirtmişlerdir.

Güvenirlik hesaplama yöntemlerinden tek uygulamaya dayanan yöntemler ise maddeler arası tutarlılığa ilişkin bir ölçü veren KR-20, KR-21 ve Cronbach  $\alpha$ 'dır. Bu yöntemler maddelerin homojenliği konusunda bilgi verirler. İki yarı güvenirliliği de tek uygulamaya dayanan yöntemlerdendir, bu yöntem iki yarının tutarlılığı hakkında bilgi vermektedir.

Ayrıca yalnız bir uygulamaya dayanan, fakat ölçümlerin güvenirliliğini belirlemede birden fazla puanlama gerektiren durumlar da vardır. Eğer yalnızca bir test ve testi puanlayan birden fazla puanlayıcı varsa, güvenilirlik için puanlayıcılar arasındaki tutarlılığı belirlemek gerekir. Fiziksel performans testlerinin birden çok kişi tarafından derecelenmesinde, kompozisyon tipi sınavların değerlendirilmesinde, sözlü sınavların değerlendirilmesinde puanlayıcı güvenirliliği hesaplanabilir. İki puanlayıcı kullanıldığında iki puan seti arasında korelasyon hesaplama yoluyla belirlenebilir. İki'den fazla puanlayıcı varsa, sınıf içi (intraclass) korelasyon veya uyum (concordance) katsayısı hesaplanır (Aiken, 1971). Ayrıca puanlayıcı güvenirliliği hesaplamada varyans analizine dayanan yöntemler de kullanılmaktadır.

Ölçmenin standart hatası da güvenilirlik anlamında yorumlanabilir. Aynı ölçek boyutunda iki testten hangisinin standart hatası düşükse o test daha güvenilir olarak yorumlanır, çünkü aynı hata düzeyinde daha dar bir güven aralığı verir. Hesaplanması eşitlik 9'da gösterilmiştir.

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}} \quad (9)$$

### **Geçerlilik ve Kestirimin Standart Hatası**

Cronbach dar anlamda, geçerliliği belirli bir kestirimin veya test puanlarından yapılan çıkarımların doğruluğunun kontrolü olarak tanımlamıştır (Thorndike, 1971). Genel anlamda ise geçerlik, ölçme aracının ölçmeyi amaçladığı özelliği, başka özelliklerle karıştırmadan, doğru olarak ölçebilme derecesidir (Tekin, 1993). Geçerlilik ölçülmek istenen özelliğin doğasına veya ölçmenin kullanılacağı amaca veya özelliğe ait verilerin elde bulunup bulunmamasına göre farklı türlere sahiptir.

*Görünüş Geçerliliği.* Ölçme aracının genel görünümüyle ilgilidir. Bir testin, inceleyen kişiye, ölçülmek istenen alanla ilgili olduğu izlenimi vermesi istenir. Görünüş geçerliği psikometrik bir nitelikten çok ticari bir niteliktir. Bu geçerlilik türü özellikle testin pazarlanmasıyla ilgilidir ve pazarlanacağı zaman anlam kazanır (Aiken, 1971).

*Kapsam Geçerliliği.* Geliştirilen bir test ile ölçülmek istenen belirli bir kapsam ise o zaman hazırladığımız testin kapsamı yeterince örneklemesini isteriz. Geliştirilen testin kapsam geçerliliğine sahip olabilmesi için, yoklanacak içeriği veya hedef davranışları yeterince örneklemesi gereklidir. Genellikle başarı testlerinde kapsam geçerliliği önemlidir. Çünkü yoklanacak bir konu alanı vardır ve hazırlanacak testin konu alanını içermesi istenir. Yetenek, ilgi ve kişilik testlerinin de bir kapsamı olabilmektedir (Aiken, 1971). Fakat yetenek, ilgi veya kişilik testlerinde yoklanmak istenen içerikten daha çok psikolojik bir yapıdır. Kapsam geçerliliğini belirlemede çok kullanılan bir yöntem ünite analiz tablosu hazırlamaktır. Uzman kanılarına dayalı olarak da kapsam geçerliği belirleme çalışmaları yapılmaktadır.

*Ölçüt Dayanaklı Geçerlilik.* Bir çok durumda hazırlanan bir testten bireylerin test ile gözlenemeyen davranışları hakkında çıkarımlar yapmak amaçlanır (Crocker ve Algina, 1986). Örneğin yapılan ÖSS sınavından elde edilen puanlardan, yerleştirileceği alandaki üniversite başarısı hakkında çıkarımlar yapılmak istenir. Bu durumda ölçüt ölçüleri ile ölçme sonuçları arasındaki tutarlılık anlamında geçerlilik önemlidir. Bazen ölçüt

ölçüleri elde bulunmaz da, yordayıcıdan daha sonraki bir zamanda elde edilir. Gelecekteki bir durum hakkında çıkarım yapmak amaçlandığında, ölçeğin ölçütü ne derece yordadığı ile ilgili bilgi elde edilmek istenir, bu tür geçerliliğe *yordama geçerliliği* adı verilir. Eğer ölçüt ölçüsü ölçme sonuçlarıyla aynı zamanda ya da daha önce elde edilmiş ise, bu durumdaki geçerlik *uyum geçerliliği* adını alır (Magnuson, 1966). Ölçüt dayanaklı geçerlik katsayıları genellikle ölçme sonuçları ile ölçüt puanları arasında korelasyon ile belirlenir.

*Yapı Geçerliliği.* Psikolojik yapıların elde edilen verilerle sağlanıp sağlanmadığına ilişkin bilgi verir. Farazi yapı geçerliliği de denir. Psikolojik bir özelliğin kuramsal olarak düşünülen bir yapısı ve bunun gözlenebilir davranışlarla ilişkisi olduğu kabul edilir. Elde edilen veriler ne oranda bu yapıyı sağlıyorsa, yapı geçerliliğine sahiplik derecesini ona göre belirleyebiliriz. Yapı geçerliliği belirlemek bir dizi işlem gerektiren bir süreçtir, bunları Aiken (1971), (1) test içeriği ile ilgilenilen yapı hakkındaki uzman kanısı, (2) iç tutarlılık katsayıları, (3) bir ölçüt değişken ile belirlenen grup farklılıkları ile ilişkisi, (4) faktör analizi, (5) bireylere veya uzmanlara cevaplarının detayları hakkında soru sorulması veya dereceleme yaptırılması, olarak sıralamıştır.

*Kestirimin Standart Hatası.* Ölçmenin standart hatasının güvenilirlik anlamında bilgi verdiği gibi, kestirimin standart hatası da geçerlilik anlamında bilgi verir. Bir test herhangi bir konudaki performansın kestiricisi olarak kullanılacaksa, o testin kestirmedeki hatası önem kazanır. Kestirim hatası düşük olan test daha geçerlidir diyebiliriz (ölçek birimi aynı ise). Kestirimin standart hatası eşitlik 10 yardımıyla hesaplanabilir.

$$\hat{\sigma}_{Y.X} = \hat{\sigma}_Y \sqrt{1 - \hat{\rho}_{XY}^2} \quad (10)$$

## **MADDE TEPKİ KURAMI**

Klasik test kuramının zayıf sayılıtlara sahip olması ve sağlanması zor olan tamamen paralel testlerin üretilebileceği sayılıtlarının bulunması, psikometristleri bu alanda yeni

çalışmalar yapmaya itmiştir. Klasik test teorisinin madde ve test istatistikleri matematiksel totolojiler içerir (Lord, 1980). Bu matematiksel ifadelerin tümü yukarıda belirtilen sayılılarla anlam kazanır. Ayrıca Hambleton ve Swaminathan (1985) KTK'nın sınırlılıklarını: (1) madde ve test parametrelerinin örneklem bağımlı olması, (2) Maddelerin hep orta yetenek düzeyine göre düzenlenmesi, (3) güvenilirliğin paralel testlere bağlı tanımlanması ve uygulamada tamamen paralel testler oluşturmanın güçlüğü, (4) bir bireyin herhangi bir maddeyle karşılaştığında nasıl performans göstereceği konusunda bilgi vermemesi, (5) ölçmenin hata varyansının tüm bireyler için aynı olması, (6) test düzenleme, madde yanlılığı, test eşitleme ve bireye uyarlanmış testler gibi ölçme problemlerine tatmin edici çözümler üretmede yetersiz kalması olarak saymışlardır. KTK'nın bu sınırlılıkları psikometristleri yeni çalışmalar yapmaya itmiş ve MTK buradan doğmuştur.

Madde Tepki Kuramı farklı isimlerle anılabilmektedir. Bunlardan bazılarını Hambleton ve Swaminathan (1985) *Örtük Özellikler Kuramı* (Latent Trait Theory), *Madde Karakteristik Eğrisi Kuramı* (Item Characteristic Curve Theory), *Madde Tepki Kuramı* (Item Response Theory) olarak saymıştır. MTK'nın temeli Binet ve Simon'un kullandığı yaşlara göre maddenin doğru cevaplandırılma olasılığı grafiğinin kullanıldığı madde karakteristik eğrisine dayandırılır (Hambleton ve Swaminathan, 1985). Daha sonra Richardson, Lawley ve Tucker'in çalışmalarıyla devam etmiştir. Tucker "madde karakteristik eğrisi" kavramını ilk kullanan kişi olarak bilinir (Hambleton ve Swaminathan, 1985).

Bir madde tepki modeli bir bireyin gözlenebilen test performansı ile testteki performansının altında yatan gözlenemeyen özelliği veya yeteneği arasındaki ilişkiyi belirtir (Lord, 1980). MTK klasik test kuramının yukarıda sayılan sınırlılıklarını taşımadığı iddiasıyla ortaya çıkmıştır. Yetenek kestirimlerinin farklı madde setlerinden kestirilmesindeki değişimin yalnızca ölçme hatasından geldiği belirtilmektedir (Hambleton ve Swaminathan, 1985).

Hambleton ve Swaminathan (1985) MTK'nın üç önemli üstünlüğü olduğunu belirtmişlerdir. Bunlar: (1) aynı özelliği ölçen geniş bir madde havuzu olduğunu

varsayalım. Bireylerin yeteneğinin kestirimi bireylere uygulanan madde setinden bağımsızdır. (2) Çok geniş bir sınanan grubu olduğunu varsayalım. Test maddelerinin özellikleri (örneğin, madde güçlüğü, madde ayırdediciliği) maddenin kalibrasyonu için seçilen gruptan bağımsızdır ve (3) ölçmenin doğruluğunu gösteren bir istatistik elde edilir ve bu istatistik bir yetenek düzeyinden diğerine değişebilir. Ama bu özelliklerin gerçekleşmesi model-veri uyum istatistiğinin durumuna bağlıdır. Eğer veriler modele uymuyorsa, sağlayacağı düşünülen üstünlükler elde edilemeyebilir.

MTK'nın da güçlü sayıltılara sahip olduğu söylenmişti. Bu sayıltıları her türlü verinin modele uymasını zorlaştırır. MTK'nın sayıltıları (1) Tek boyutluluk, (2) Madde karakteristik eğrisinin modele uygunluğu, (3) Yerel bağımsızlık ve (4) Hızsızlık olarak sayılabilir (Hambleton ve Swaminathan, 1985).

Tek boyutluluk sayıltısı testte yer alan maddelerin tamamının aynı özelliği ölçmesi olarak tanımlanabilir. Bir testte cevaplayıcının maddeleri doğru cevaplama olasılığı, yalnızca testle ölçülen tek bir özelliğe ait yeteneğin düzeyine bağlı olmalıdır. Maddenin doğru cevaplanmasının test ile ölçülmek istenen özelliğin yanında başka özellikleri gerektirmesi, söz konusu özelliğe ait gerçek performansın görülmesine engel olacaktır. Bu durum maddelerin parametre ve bireyin yetenek kestiriminde hataya sebep olacaktır. Eğer maddelerden bazıları ilgilenilen özelliğin dışında başka şeyleri ölçüyorsa testte birden fazla boyut vardır. Madde karakteristik eğrisinin modele uygunluğu ise seçilen modelin parametre sayısının elde edilen verilerle uyumudur. Eğer bir parametrelili model kullanılacaksa, şans parametresinin olmaması ve tüm maddelerde ayırcılığın eşit olması gerekir. Eğer iki parametrelili model kullanılacaksa, madde karakteristik eğrisi dikey eksenini orijinde kesmelidir. Ya da veriler iki kategorili değil de çoklu puanlama şeklinde elde edilmiş ise, çok kategorili uygun bir model seçilmelidir. Yerel bağımsızlık ise, bir birey için maddelerin birinin doğru cevaplanması bir diğer maddeye bağlı olmaması anlamına gelir. Belirli bir yetenek düzeyinde ( $\theta$ ) bir maddeyi doğru cevaplayamayan bireyler bir diğer maddeyi de doğru cevaplayamıyor ise maddeler bağımlıdır denir. Zamanın yeterliliği sayıltısı, bir testte maddelerin tümüne ulaşılacak kadar yeterli zaman verilmesi anlamına gelir. Bireyin maddeden puan alamamasının tek nedeni, bireyin o maddeyi cevaplayabilecek yeterli bilgi düzeyine sahip olmaması olmalıdır.

MTK'da ana kavram madde karakteristik eğrisidir (Item Characteristic Curve). Bireylerin k boyutlu örtük uzaydaki yeri yetenek puanı vektörünün temsilcisi ( $\theta_1, \theta_2, \theta_3, \dots, \theta_k$ ) olarak ve bir bireyin i'nci özelliğinin düzeyi ise  $\theta_i$  olarak gösterilir (Hambleton ve Swaminathan, 1985). Madde karakteristik eğrisi test performansının altında yatan örtük özelliğın ( $\theta$ ) bir fonksiyonu olarak maddenin doğru cevaplanma olasılığını gösterir. Bu fonksiyon  $\theta$  üzerinde tekdüze artan bir yapıdadır (van der Linden ve Hambleton, 1997). ICC sınananların yetenekleri üzerinde, maddenin doğru cevaplanma olasılığının regresyonudur (Embretson ve Reise, 2000).

Madde karakteristik fonksiyonları kullanılan fonksiyonun normal ogive veya lojistik ogive olmasına göre farklılaşabilir. Bu farklılaşma oluşturulan fonksiyonun türüne göre olmaktadır. Normal ogive fonksiyonu fonksiyon eğrisinin altında kalan alana göre oluşturulmuştur. Üç parametrelı normal ogive modeli eşitlik 11'de gösterilmiştir.

$$P(\theta) = c + (1 - c) \int_{-\infty}^{a(\theta-b)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (11)$$

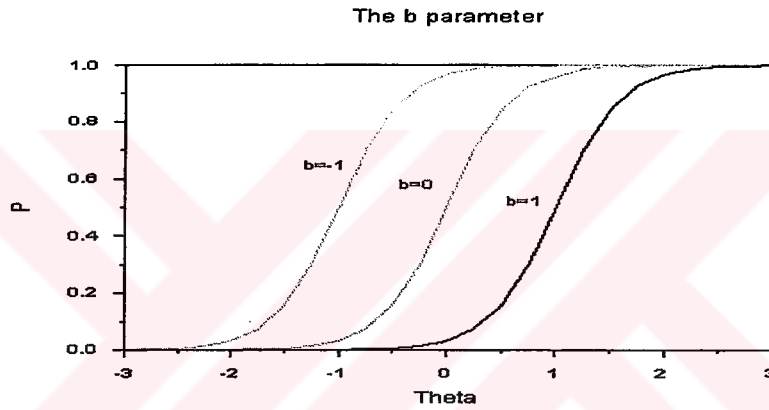
burada, c eğrinin dikey eksenı kestiğı noktayı, b eğrinin dönme noktasındaki yetenek düzeyini, a eğrinin dönme noktasındaki eğimini ve z ise  $b_i$  ortalama ve  $1/a_i$  standart sapmalı dağılımdaki birim standart normal sapma değerini göstermektedir (Lord, 1980).

Madde karakteristik fonksiyonu maddenin parametrelerinin sayısına göre de farklı isimler alır. Bunlar ayrıca, elde edilen madde puanlarının kategori sayısına göre iki kategorili ve çok kategorili modeller olarak farklılaşır. İki kategorili modeller bir parametrelı, iki parametrelı ve üç parametrelı lojistik modellerdir. Çok kategorili modeller: Graded Response Model, Modified Graded Response Model, Nominal Response Model, Multiple Choice Model, Kısmi Puan (Partial Credit) Modeli, Genelleştirilmiş Kısmi Puan Modeli olarak sayılabilir.

## Lojistik Madde Tepki Kuramı Modelleri

*Bir Parametrelili Lojistik Model (Rasch Modeli):* Bir parametrelili modelde sadece güçlük parametresi ve yetenek parametresi vardır. Güçlük parametresi ile yetenek parametresi  $\theta$  ile gösterilir. Tüm maddelerin ayırıcılık güçlerinin eşit olduğu varsayılır. Madde karakteristik fonksiyonunun ifadesi eşitlik 12’de, madde karakteristik eğrisi örneği şekil 1’de gösterilmiştir.

$$P_i(\theta) = \frac{1}{1 + \exp\{- (\theta - \delta_i)\}} \quad (12)$$

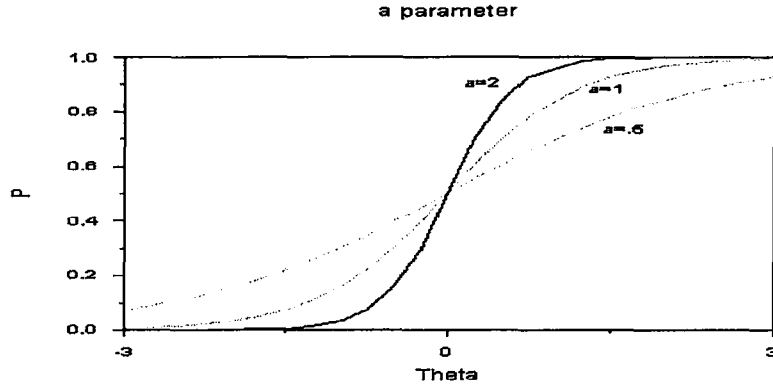


**Şekil 1: Bir parametrelili madde Karakteristik Eğrisi**

*İki Parametrelili Lojistik Model:* Bir parametrelili modele ek olarak ayırıcılık parametresinin de işe koşulduğu modeldir. Eğrinin dönme noktasındaki eğimini gösteren ayırıcılık parametresi  $a_i$  ile gösterilir. İki parametrelili lojistik fonksiyonun ifadesi eşitlik 13’de, madde karakteristik eğrisi örneği şekil 2’de gösterilmiştir.

$$P_i(\theta) = \frac{1}{1 + \exp\{-Da_i(\theta - \delta_i)\}} \quad (13)$$

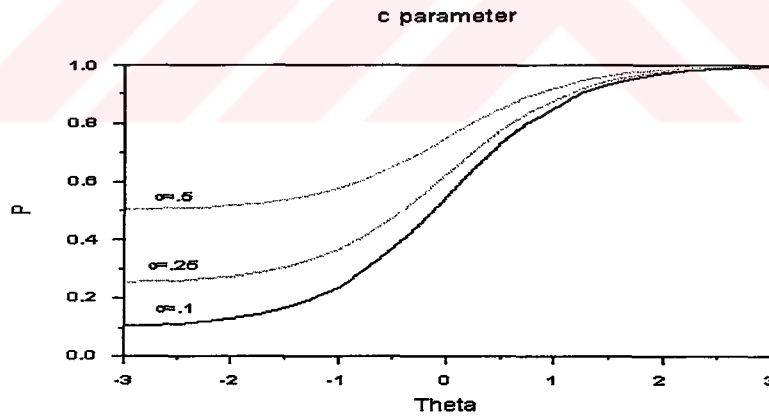
bu eşitlikte D sabit bir ölçekleme faktörü olarak alınır, değeri  $D=1,7$ ’dir.



**Şekil 2: İki parametrelı madde Karakteristik Eğrisi**

*Üç parametrelı Lojistik Model:* Bu modelde, iki parametrelı modele ek olarak  $c_i$  şans parametresi vardır. En düşük yetenek düzeyindeki bireyin maddeyi doğru cevaplama olasılığını ve eğrinin dikey eksenı kestiğı noktayı gösterir. Üç parametrelı lojistik modelin ifadesi eşitlik 14'de, madde karakteristik eğrisi de şekil 3'de verilmiştir.

$$P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + \exp\{-D\alpha_i(\theta - \delta_i)\}} \quad (14)$$



**Şekil 3: Üç parametrelı madde Karakteristik Eğrisi**

### Madde ve Test Bilgi Fonksiyonu

Madde bilgi fonksiyonu iki farklı yetenek ( $\theta_1$  ve  $\theta_2$ ) düzeyinde maddenin ayırdetmesindeki etkililiğinin değerlendirme fonksiyonu olarak kabul edilir

(Hambleton ve Swaminathan, 1985). Teknik tanım olarak da madde karakteristik eğrisinin belirlenen yetenek düzeyindeki eğiminin karesinin, madde varyansına oranı olarak tanımlanabilir. Madde bilgi fonksiyonunun ifadesi eşitlik 15’de verilmiştir.

$$I(\theta, u_i) = \frac{(P_i)^2}{P_i Q_i} \quad (15)$$

Bir ve iki parametrelî modellerde madde bilgi fonksiyonunun en yüksek olduđu yetenek düzeyi madde güçlüğüne karşılık gelen düzeydir. Üç parametrelî modelde madde bilgi fonksiyonunun en yüksek olduđu yetenek düzeyi ayırıcılık ve şans parametrelerinden de etkilenir (Hambleton ve Swaminathan, 1985). Üç parametrelî modelde en yüksek madde bilgi düzeyinin elde edildiđi yetenek düzeyi eşitlik 16’da verilmiştir.

$$\theta_{\max} = b_i \frac{1}{Da_i} \ln \left[ \left( \frac{1}{2} \right) + \left( \frac{1}{2} \right) \sqrt{1 + 8c_i} \right] \quad (16)$$

Test bilgi fonksiyonu madde bilgi fonksiyonlarının toplamlarından oluşur ve madde bilgi fonksiyonlarının birbirlerinden bağımsız olarak teste katkılarını içerir (Lord, 1980). Test bilgi fonksiyonu madde bilgi fonksiyonlarının toplamından oluştuđu için belirli bir yetenek düzeyinde hedeflenen bilgi düzeyine ulaşmak için hangi bilgi düzeyinde kaç madde alınması gerektiđi belirlenebilir. Test bilgi fonksiyonunun ifadesi eşitlik 17’de verilmiştir.

$$I(\theta) = \frac{\sum (P_i)^2}{\sum P_i Q_i} \quad (17)$$

İki ayrı testten test bilgi fonksiyonu yüksek olan testin daha güvenilir sonuçlar vereceđi söylenebilir. Çünkü herhangi bir  $\theta$  yetenek düzeyi için, kestirilen yetenek düzeyinin standart hatası test bilgi düzeyi ile ters orantılıdır (Hambleton ve Swaminathan, 1985). Bir test için kestirilen yetenek düzeyinin standart hatası eşitlik 18’de verilmiştir.

$$SE(\theta) = \frac{1}{\sqrt{\sum I(\theta)}} \quad (18)$$

Aynı yeteneği ölçen iki testin veya aynı testin iki farklı puanlama yönteminin birbirine göre verimliliği test/puanlama bilgi fonksiyonlarına bağlı olarak belirlenebilmektedir. Bilgi fonksiyonu daha yüksek olan test daha verimli olarak kabul edilirse, test bilgi fonksiyonlarının oranı görece verimliliği (relative efficiency) verecektir. Göreceli verimlilik fonksiyonu örtük özellik ölçeğinde her noktada ayrı değerlendirilir (Crocker ve Algina, 1986). Göreceli verimliliğin ifadesi eşitlik 19'da verilmiştir.

$$RE(\theta, X, Y) = \frac{I(\theta, X)}{I(\theta, Y)} \quad (19)$$

MTK'de veriler iki kategorili (dichotomous) yerine çok kategorili (polytomous) elde ediliyorsa, çok kategorili modellerden uygun olanını kullanmak gerekecektir. Bu araştırmada kullanılan eleme puanlaması kısmi bilgiyi ölçmeyi amaçladığından ve çok kategorili puanlama yapıldığından uygun model olan Kısmi Puan (Partial Credit) modeli olacaktır.

### **Kısmi Puan (Partial Credit) Modeli**

Kısmi Puan Modeli (Partial Credit Model veya PCM) orijinal olarak kısmi doğru cevabın mümkün olduğu madde tiplerinin bulunduğu testlerin analizinde kullanılmak üzere geliştirilmiş bir modeldir (Embretson ve Reise, 2000). PCM iki veya daha fazla sıralı kategoride elde edilen cevap puanlarını analiz etmek için geliştirilmiş tek boyutlu bir modeldir (Masters ve Wright, 1997). Testteki maddelerin her biri iki veya ikiden çok kategoride madde puanları verdiği zaman Partial Credit Model veya başka çok kategorili modeller kullanılabilir (örneğin, Graded Response Model ve Modified Graded Response Model gibi), fakat PCM bu modellerin en basitidir (Masters ve Wright, 1997).

PCM model olarak iki kategoride ele alındığında Rasch modeline dönüşür. PCM Rasch modelinin çok kategorili verilere uyarlanmış şeklidir denilebilir. Rasch modelinin taşıdığı tüm özellikleri taşır, örneğin ayrılabilir birey ve madde parametrelerine sahiptir (Embretson ve Reise, 2000). PCM iki kategorili Rasch modelinin sıralı kategorilere genelleştirilmiş halidir. İki parametre seti içerir, bunlardan birincisi bireylerin yetenek parametresi ( $\theta_j$ ) ve diğeri maddelerin güçlük parametresi ( $\delta_i$ ). Güçlük parametresi her kategori için bir güçlük parametresi olacak şekilde hesaplanır.  $x=0, 1, \dots, m_i$  kategori sayısı olacak biçimde  $K=m_i+1$  olası puan mümkün olacaktır. Bu durumda her madde için  $\delta_1, \delta_2, \dots, \delta_{m_i}$  sayıda kategori güçlük parametresi olacaktır. Bunlara madde kategori güçlüğü ve bazen de madde basamak güçlüğü denir (Embretson ve Reise, 2000).

Bireyin  $x-1$  kategorisinden ziyade  $x$  kategorisinde cevaplama olasılığını gösteren şartlı olasılık modeli eşitlik 20'de verilmiştir.

$$\frac{P_{ijx}}{P_{ijx-1} + P_{ijx}} = \frac{\exp(\theta_j - \delta_{ix})}{1 + \exp(\theta_j - \delta_{ix})} \quad x=1,2,\dots,m_i \quad (20)$$

burada  $P_{ijx-1}$  j bireyinin i maddesinde  $x-1$  puanı alma olasılığını,  $P_{ijx}$  j bireyinin i maddesinde  $x$  puanı alma olasılığını,  $\theta_j$  j bireyinin yetenek düzeyini,  $\delta_{ix}$  x kategorisinin güçlük düzeyini göstermektedir.

Kısmi Puan Modelini Embretson ve Reise (2000) toplamla bölünme (divide-by-total) ya da doğrudan (direct) model olarak ifade etmişlerdir. Bunun anlamı belirli bir kategoride cevaplama olasılığı doğrudan o kategoriye ait üstel'in (exponential  $e$ ) toplam üstel'e bölümü olarak ifade edilmesidir.  $x=0, 1, \dots, m_i$  kategoriler olacak biçimde modelin şartsız olasılık ifadesi eşitlik 21'de verilmiştir.

$$P_{ix}(\theta_j) = \frac{\exp\left(\sum_{k=0}^x (\theta_j - \delta_{ik})\right)}{\sum_{h=0}^{m_j} \left[ \exp\left(\sum_{k=0}^h (\theta_j - \delta_{ik})\right) \right]} \quad (21)$$

burada  $P_{ijx}$  j bireyinin i maddesini  $x$  kategorisinde cevaplama olasılığını,  $\theta_j$  j bireyinin yetenek düzeyini,  $\delta_{ik}$  k kategorisinin güçlük düzeyini göstermektedir. Burada,  $\delta_{ik}$  bazen

madde basamak güçlüğü olarak isimlendirilir; basamağın  $\delta_{ik}$  değeri daha yüksek olan daha fazla güçlük içerdiği anlamına gelir.

### Genelleştirilmiş Kısmi Puan Modeli (Generalized Partial Credit Model)

Genelleştirilmiş Kısmi Puan Modeli Muraki tarafından geliştirilmiştir. Kısmi Puan Modeli'nden farkı, eğim parametrelerinde farklılaşmaya izin vermesinden kaynaklanmaktadır.

Genelleştirilmiş Kısmi Puan Modeli'nde her bir kategorinin olasılığını belirlemek eşitlik 21'e eğim parametresinin eklenmesi yoluyla elde edilir. Bunun ifadesi eşitlik 22'de verilmiştir.

$$P_{ix}(\theta_j) = \frac{\exp\left(\sum_{k=0}^x \alpha_{ik}(\theta_j - \delta_{ik})\right)}{\sum_{h=0}^{m_j} \left[ \exp\left(\sum_{k=0}^h \alpha_{ik}(\theta_j - \delta_{ik})\right) \right]} \quad (22)$$

Bu modelde Kısmi Puan modelinden farklı olarak sadece  $\alpha_{ik}$  i kategorisinin eğimini göstermektedir. Bu modelde de  $\delta_{ik}$  parametresi Kısmi Puan modelinde olduğu gibi basamak güçlüğü olarak yorumlanır. Fakat,  $\alpha_{ik}$  parametresi 2PL modelden farklıdır. Bu modelde  $\alpha_{ik}$  parametresi her bir kategorideki eğim parametresini gösterir ve madde ayıricılığı eğim parametrelerinin bir kombinasyonuna bağlıdır (Embretson ve Reise 2000).

Çoktan seçmeli testlerin geleneksel uygulamasında genellikle iki kategorili modeller kullanılmasına rağmen, eleme uygulamasında çok kategorili modellerden birinin kullanılması gerekmektedir. Çoktan seçmeli testlerin geleneksel uygulaması ile eleme uygulamasının MTK temelinde de karşılaştırıldığı bu çalışmada model veri uyumundan elde edilen sonuçlara göre iki parametrelili model ile genelleştirilmiş kısmi puan modeli kullanılacaktır.

## PROBLEM CÜMLESİ

Geleneksel yöntem ve eleme yöntemiyle puanlanan çoktan seçmeli testlerin psikometrik özellikleri arasında fark var mıdır?

### Alt Problemler

1. Geleneksel puanlama yöntemi ve eleme puanlaması yöntemi ile puanlanan çoktan seçmeli testlerin Klasik Test Kuramı'na göre,
  - a) Maddelerin ayırıcılık güçleri farklılaşmakta mıdır?
  - b) Belirli bir güvenilirliği sağlayan madde sayıları oranı nasıldır?
  - c) Test güvenilirlikleri farklılaşmakta mıdır?
  - d) Toplam puanları arasında nasıl bir ilişki vardır?
2. Geleneksel puanlama yöntemi ve eleme puanlaması yöntemi ile puanlanan çoktan seçmeli testlerin Madde Tepki Kuramı'na göre,
  - a) Belirlenen 7 yetenek düzeyinde madde bilgi düzeylerinin göreceli etkililiği (relative efficiency) nasıldır?
  - b) Marjinal güvenilirlikleri farklılaşmakta mıdır?
  - c) Kestirilen yetenek düzeyleri arasındaki ilişki nasıldır?
3. Test uygulanan bireylerin geleneksel uygulamayla eleme puanlaması uygulamasına ilişkin görüşleri nasıldır?

## SAYILTILAR

1. Bu çalışmada kullanılan örnekleme dahil olan bireylerin ciddiyetle test sorularını cevapladıkları varsayılmaktadır.
2. Geleneksel ve eleme yönteminin yönergelerini öğrenciler aynı düzeyde algılamaktadırlar.
3. Her bir maddedeki her seçeneğin birbirinden bağımsız olduğu varsayılmaktadır.
4. Testi uygulayanların uygulama yönergelerini kavramış olduğu ve öğrencilere açıklayabildikleri varsayılmıştır.

## İLGİLİ ARAŞTIRMALAR

Coombs, Milholland ve Womer (1956) çoktan seçmeli testlerde eleme puanlamasını geleneksel puanlama ile karşılaştırmıştır. Yapılan çalışmada, KR-20 güvenilirliği ve geçerlik yönlerinden her iki uygulamadaki sonuçları karşılaştırılmıştır. İki yöntemin güvenilirlikleri arasında eleme puanlaması lehine küçük bir fark bulunmuştur. Fakat bu fark daha zor maddelerin oluşturduğu testlerde, açık olarak görülebilmektedir. Daha kolay maddelerin oluşturduğu testte fark anlamlı değildir. Ayrıca güvenilirliğe göre göreceli uzunluk etkililiği incelenmiş ve eleme puanlamasının %20 oranında daha etkili olduğu görülmüştür. Geçerlik karşılaştırılmasında ise ölçüt alınan teste göre her iki yöntemin de geçerliği sağlanamamıştır.

Madde analizlerinin karşılaştırılmasında ise her iki yöntem ile yapılan madde analizi sonuçlarının bu çalışmada farklılaşmadığı görülmüştür. Fakat test uygulanan bireylere yapılan anketin sonuçları eleme puanlaması uygulamasının daha çok tercih edilebileceğini göstermiştir.

Collet (1971) yaptığı çalışmada eleme puanlaması ile geleneksel puanlama uygulamalarını karşılaştırmıştır. Bu çalışmada her iki uygulamanın test-tekrar test güvenilirliği ile yordama geçerlikleri karşılaştırılmıştır. Test-tekrar test güvenilirlikleri eleme puanlamasında 0.86, geleneksel puanlamada 0.81 olarak bulunmuştur. Buradan eleme puanlamasının güvenilirliği küçük miktarda artırdığı söylenebilir. Yordama geçerliğinde ise benzer şekilde sırasıyla 0.77 ve 0.67 olarak sonuçları elde edilmiştir. Güvenirlikte olduğu gibi yordama geçerliğinde de eleme puanlamasının daha iyi sonuçlar verdiği gözlenmiştir.

Hakstian ve Kansup (1975) kelime bilgisi ve matematiksel düşünme becerisi testlerini eleme puanlaması, geleneksel yöntem ve olasılık testi şeklinde uygulamışlar ve sonuçları alpha güvenilirliği, test-tekrar test güvenilirliği ve yordama geçerlikleri yönünden karşılaştırmışlardır. Alpha güvenilirliği birinci kelime bilgisi testinde eleme puanlaması 0.65, olasılık testi 0.74 ve geleneksel uygulama 0.63 bulunmuştur, yordama geçerlikleri ise sırasıyla 0.67, 0.65 ve 0.59 olarak bulunmuştur. İkinci kelime testinde alpha güvenilirliği sırasıyla 0.69, 0.72 ve 0.67 hesaplanmıştır. Birinci matematiksel düşünme becerisi testinde ise alpha güvenilirlikleri eleme puanlamasında 0.56, olasılık testinde 0.69 ve geleneksel uygulamada 0.60 olarak hesaplanmıştır. Yordama geçerlikleri ise sırayla 0.38, 0.49 ve 0.50 hesaplanmıştır. İkinci matematiksel düşünme testinde alpha güvenilirlikleri sırayla 0.64, 0.69 ve 0.64 bulunmuştur. Kelime bilgisi testinin test-tekrar test güvenilirlikleri eleme puanlamasında 0.67, olasılık testinde 0.76 ve geleneksel uygulamada 0.68 bulunmuştur. Matematiksel düşünme testinde test-tekrar test güvenilirliği sırayla 0.60, 0.67 ve 0.61 olarak hesaplanmıştır. Bu çalışmada eleme puanlaması geleneksel uygulamadan anlamlı derecede farklı bulunmazken, olasılık testi puanlaması daha iyi sonuçlar vermiştir.

Jaradat ve Tollefson (1988), yaptıkları çalışmada, eleme puanlamasını, altset seçme puanlamasını ve tahmin düzeltmeli geleneksel metodun güvenilirliklerini ve ölçüt dayanaklı geçerliklerini karşılaştırmışlardır. Alpha güvenilirlikleri eleme puanlamasında 0.67, altset seçme tekniğinde 0.71 ve tahmin düzeltmeli geleneksel metotta 0.57 olarak hesaplanmıştır. Güvenirlik farkları her üç metotta anlamlı bulunmamıştır. Ölçüt

dayanaklı geçerliklerde ise, en iyi sonuçlar eleme puanlamasında elde edilmesine rağmen, farklar anlamlı bulunmamıştır.

Ben-Simon, Budescu ve Nevo (1997) yaptıkları çalışmada, çoktan seçmeli testlerde kısmi bilginin ölçülmesini amaçlayan yöntemlerle geleneksel puanlamaların karşılaştırılmasını incelemişlerdir. Bu çalışmada, geleneksel puanlama (NC), geleneksel puanlamanın tahmin için düzeltme formülü kullanılması durumu (CG), eleme puanlaması (ET), olasılık testi (PT), güven işaretlemesi (CM), tam sıralama (CO) ve kısmi sıralama (PO) yöntemleri kullanılmıştır. Elde edilen katsayılar ise NC (Doğru cevap sayısı puanlaması) için 0.717, CG (Tahmin için düzeltme formülü uygulama) için 0.714, ET (eleme puanlaması) için 0.762, PT (Doğru cevabı olasılıkla cevaplama) için 0.819, CM (doğru cevabı belirlemede cevaplayıcının güveni) için 0.80, CO (Tam sıralama) için 0.726 ve PO (kısmi sıralama) için 0.75 olarak belirlenmiştir.

Bu yöntemlerin alpha güvenilirlikleri karşılaştırmasında, dört test ortancası olarak elde edilen değerler aşağıda verilmiştir. Yukarıda görüldüğü gibi, kısmi puanı ölçme iddiasında olan metotlar daha yüksek güvenilirlik vermiştir, fakat ayrılan zaman için düzeltme formülü kullanıldığında, sadece CM ve PT anlamlı olarak diğer kısmi bilgiyi ölçmeyi amaçlayan metotlara göre daha yüksek güvenilirlik vermiştir.

Bradbard, Parker ve Stone (2004) yaptıkları çalışmada, geleneksel doğru cevap sayısı puanlaması (NC) ile eleme puanlamasını (ET) karşılaştırmışlardır. Bu çalışmada Macroeconomics alanında dört testin NC ve ET uygulamaları kullanılmıştır. ET ve NC uygulamalarının puanları, güvenilirlikleri ve madde güçlükleri karşılaştırılmıştır. Bu çalışmada NC puanlamasının doğru cevap sayısı ET puanlamasının doğru cevap sayısından tüm testlerde yüksek bulunmuştur. Bu bulgu ET uygulamasının tahminden gelen puanları engellediği anlamında yorumlanmıştır. Uygulanan 4 testte öğrencilerin yarısına öncelikle ET, diğer yarısına öncelikle NC verilmiştir. Güvenirlik karşılaştırmalarında test istatistiği kullanılmamış, sadece üçüncü testte ilk yarıda ET uygulanmasında yüksek fark bulunmuş, diğer yedisinde güvenilirlik katsayıları arasındaki farkın 0.05'den daha küçük olduğu belirtilmiştir. Bu bulgu ET uygulamasının NC uygulaması kadar iyi sonuçlar verdiği şeklinde yorumlanmıştır.

Ayrıca NC ve ET uygulamalarının madde güçlükleri arasında korelasyonlar Pearson ve Spearman katsayıları olarak belirlenmiştir. Kullanılan 4 test için madde güçlükleri arasında Pearson korelasyonlar 0.734 ile 0.852 arasında, Spearman korelasyonlar ise 0.694 ile 0.871 arasında bulunmuştur. Bu bulgu iki test sürecinin tutarlı madde güçlüğü ürettiği şeklinde yorumlanmıştır.



## BÖLÜM II

### YÖNTEM

Bu bölümde, araştırmanın türü, verilerin toplandığı grup, ölçme araçları ve verilerin çözümlenmesinde kullanılan yöntemler açıklanacaktır.

#### ARAŞTIRMANIN TÜRÜ

Bu araştırmada, çoktan seçmeli testlerin geleneksel uygulanma yöntemi ile eleme puanlaması uygulaması yöntemlerinin karşılaştırılması amaçlanmaktadır. Her iki yöntemde elde edilen verilerin analizi sonucunda madde ve test istatistikleri elde edilerek karşılaştırılmıştır.

Bu araştırmada, bir örneklemden elde edilen bulguların örneklemin temsil ettiği evrene genellenmesi gibi bir amaç taşımadığı için ve eleme puanlaması ile geleneksel puanlama metotlarının karşılaştırılması yoluyla alana yeni bilgiler katmayı amaçladığı için temel araştırma niteliği taşıyabilir.

#### VERİLERİN TOPLANDIĞI GRUP

Araştırmanın bir temel araştırma niteliği taşıması ve örneklemden elde edilecek verilerin evrene genellemesi gibi bir çaba olmadığı için herhangi bir örnekleme tekniği ile örneklem alınmamıştır. Verilerin toplanması için LES sınavının sözel bölümünde yer alan maddelere benzer 30 çoktan seçmeli madde içeren sözel yetenek testi kullanılmıştır. Oluşturulan sözel yetenek testi hem geleneksel çoktan seçmeli madde cevaplama şekliyle, hem de eleme puanlamasına uygun şekilde 2004-2005 öğretim yılında Hacettepe Üniversitesi Yabancı Diller Yüksekokulunda İngilizce hazırlık okuyan 30'a yakın farklı bölümden 470 öğrenciye uygulanmıştır. İngilizce seviyeleri bakımından her gruptan öğrenci alınmıştır. 470 öğrencinin bazıları ilk uygulamaya, bazıları ikinci uygulamaya katılmadığından, bazılarının da ciddiyetle

cevaplamamasından, kalan 315 öğrenci üzerinden analizler yapılmıştır. Araştırma kapsamına dahil olan 315 öğrencinin cinsiyete göre dağılımı 155'i kız, 160'ı erkek olacak şekildedir.

## ÖLÇME ARAÇLARI

Araştırma problemlerine çözüm bulmak amacıyla bir çoktan seçmeli test geliştirilmiştir. Bu amaçla ÖSYM'nin yaptığı LES sınavının madde tiplerine benzeyen sözel akıl yürütme ve mantıksal akıl yürütme sorularından oluşan 40 madde yazılmış, yazılan maddeler redaksiyondan ve ön incelemeden geçirilerek 30'a düşürülmüştür. Maddelerin azaltılmasının amacı bir ders saati içerisinde uygulanabilecek bir test oluşturmaktır. Araştırma problemlerinin amacı maddelerin ve testin istatistiklerini karşılaştırmak olduğu için deneme uygulaması yapılmamıştır.

Oluşturulan sözel yetenek testi her iki yöntemle aynı gruba uygulanmıştır. Testin geleneksel uygulamasında ve eleme puanlaması uygulamasında farklı yönergeler kullanılmıştır. Geleneksel uygulamaya ve eleme uygulamasına ait yönergeler şekil 4'te verilmiştir.

Aynı gruptan elde edilen geleneksel uygulamaya ait puanlar ile eleme uygulamasına ait puanların madde ve test istatistikleri hesaplanarak karşılaştırılmıştır. İki yöntemin uygulanma sırasından kaynaklanan etkilenme durumlarını kontrol edebilmek için grup ikiye ayrılarak yarısına önce geleneksel uygulama, sonra eleme puanlaması uygulaması yapılırken, diğer yarıya önce eleme puanlaması uygulaması daha sonra geleneksel uygulama yapılmıştır. İki uygulama arasına, ilk uygulamada verilen cevapların hatırlanmayacağı kadar bir süre olarak, yaklaşık 40 günlük ara verilmiştir.

Eleme puanlamasının uygulanması aşamasında yönergesinin farklı olması ve bu yönergenin yöntemin tanıtımı için önemli olması açısından aşağıda verilmesi uygun görülmüştür.

## ELEME PUANLAMASI YÖNERGE

Değerli Öğrenciler,

Bu test bilimsel bir çalışma yapmak amacıyla uygulanmaktadır. Bu testin sonuçlarından elde edilen bilgiler başka herhangi bir amaçla kullanılmayacaktır. Test sonuçları hiçbir kişi veya kuruluşla paylaşılmayacaktır. Bu bilimsel çalışmadan doğru sonuçların elde edilebilmesi için testi ciddiyetle cevaplamamız çok önemlidir.

Bu testte 30 tane çoktan seçmeli soru vardır. Testi cevaplarırken bildiğiniz çoktan seçmeli test uygulamalarından farklı bir cevaplama şekli kullanılacaktır. Bu uygulamada, doğru cevabı bulup işaretleme yerine sadece **yanlış olduğundan emin olduğunuz** seçenekleri işaretlemeniz gerekmektedir. Sizden istenen, soruları okuyup her sorudaki yanlış seçenekleri belirlemenizdir. Cevap kağıdına yanlış olduğunu düşündüğünüz seçenekleri işaretleyiniz. Bu sebeple tek seçenek işaretleme şeklinde değil de 4 seçeneğe kadar çıkabilen yanlış seçenekleri belirlemeniz gerekmektedir. **Yanlış olduğundan emin olmadığınız hiçbir seçeneği işaretlemeyiniz. Tüm yanlış seçenekleri bulmanız beklenmemekte**, bulabildiğiniz yanlış seçenekleri işaretlemeniz istenmektedir. Sadece doğru cevabı kesinlikle bulduğunuzdan emin olduğunuz zaman, kalan 4 yanlış seçeneği işaretleyiniz. Diğer durumlarda yalnız belirleyebildiğiniz yanlış seçenekleri işaretlemeniz yetecektir. Yanlış olduğunu bulduğunuz her seçenek size o sorudan 1 puan kazandıracak, doğru cevabın yanlış olarak işaretlenmesi durumunda ise 4 puan kaybettirecektir. Bir sorudan alabileceğiniz puan +4 ile -4 arasında değişebilir.

**Örnek Soru:** "Kokusu çıkmak" deyimiyle dile getirilen aşağıdakilerden hangisidir?

- A) Artan bir yemeğin bozulması
- B) Bir kişinin insanları bıktırması
- C) Bir şeyin işe yaramaz hale gelmesi
- D) Gizli yapılan bir işin duyulması
- E) Bir çiçeğin kokusunu kaybetmesi

Yukarıdaki sorunun doğru cevabı D seçeneğidir. B, C ve E seçeneklerinin yanlış olduğunu bilen, fakat A ve D seçeneklerinin birinin veya her ikisinin birden doğru olabileceği konusunda şüpheleri olan kişi B, C ve E seçeneklerini işaretleyerek cevabını kaydedecektir. Bu işaretlemeleri yapan kişi 3 yanlış seçeneği işaretlediği için 3 puan alacaktır.

Durum 1:            A            B            C            D            E

○            ●            ●            ○            ●

Aynı soruda, B ve D seçeneklerinin yanlış olduğunu düşünerek işaretleyen başka bir kişi, bu seçenekleri işaretlediği zaman, doğru cevabı işaretlediği için eksi 4 puan ve bir yanlış seçeneği işaretlediği için de artı 1 puan kazanacak, bu sorudan toplam eksi 3 puan alacaktır.

Durum 2:            A            B            C            D            E

○            ●            ○            ●            ○

Katılımınızdan dolayı teşekkür ederim.  
Bayram ÇETİN.

Eleme puanlaması yapılırken KTK analizleri için -4 ile +4 arasında bir puanlama yapılırken, MTK analizlerinde puanlama eksi puan olmayacak şekilde dönüştürülerek yapılmıştır. MTK'da eleme puanlaması 1'den 9'a kadar bir aralığa dönüştürülerek puanlanmış ve bu şekilde analize katılmıştır. Multilog veya diğer MTK modellerinin kullanıldığı programların yapısı bu zorunluluğu gerektirmiştir.

## VERİLERİN ÇÖZÜMLENMESİ

Bu araştırma için oluşturulan test aynı formda yönergeler değiştirilerek iki form halinde düzenlenerek uygulanmış ve öğrencilerin tepkileri geleneksel uygulamada doğru cevaba 1 puan yanlış cevaplara ve boş bırakılanlara 0 puan şeklinde puanlanarak, eleme uygulamasında ise elenen seçeneklerin sayısına ve doğru cevabı içermesine göre -4 ile +4 arasında puanlanarak Excel 2000 paket programına girilmiştir.

Elde edilen veriler daha sonra SPSS 11.5 paket programına aktararak puan dağılımlarına ilişkin özellikler belirlenmiştir. Yine SPSS programında madde puanlarının tek boyutluluğu ve yerel bağımsızlığı test edilmiştir. Tek boyutluluk ve yerel bağımsızlığın test edilmesinde faktör analizi tekniği kullanılmıştır. Verilerin “İki Parametrelili” madde tepki modeline ve “Genelleştirilmiş Kısmi Bilgi Modeli”ne uyumu ise  $-2 \log \lambda$  istatistiği ile test edilmiştir. Puan dağılımlarının özellikleri, tek boyutluluk ve yerel bağımsızlığa ve model-veri uyumuna ilişkin sonuçlar ileride ilgili başlıklarla verilmiştir.

### Geleneksel Puanlama ve Eleme Puanlaması Dağılımlarının Özellikleri

Bu çalışmanın amacı çoktan seçmeli testlerde kısmi puanlama olanağı sunan ve şans başarısı etkisini yok edebileceği düşünülen eleme puanlaması yönteminin geleneksel puanlama ile karşılaştırmak olduğundan, iki yöntemle elde edilen puanların dağılımları aşağıda verilmiştir.

Dağılımlara ilişkin bilgiler olarak merkeze yığılma ölçülerinden aritmetik ortalama ve ortancalar, değişim ölçülerinden varyans, standart sapma ve ranjlar, en küçük ve en büyük puanlar, ayrıca dağılımın normalde sapmasına ilişkin çarpıklık ve basıklık katsayıları Tablo 2’de verilmiştir.

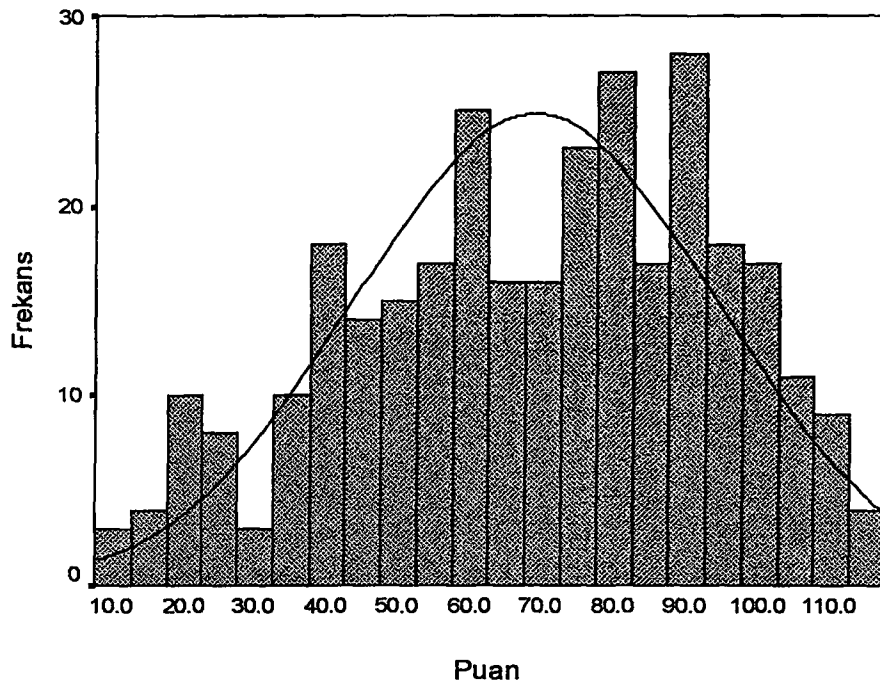
Tablo 2

## Geleneksel ve Eleme Puanlamasına İlişkin Puan Dağılımlarının Özellikleri

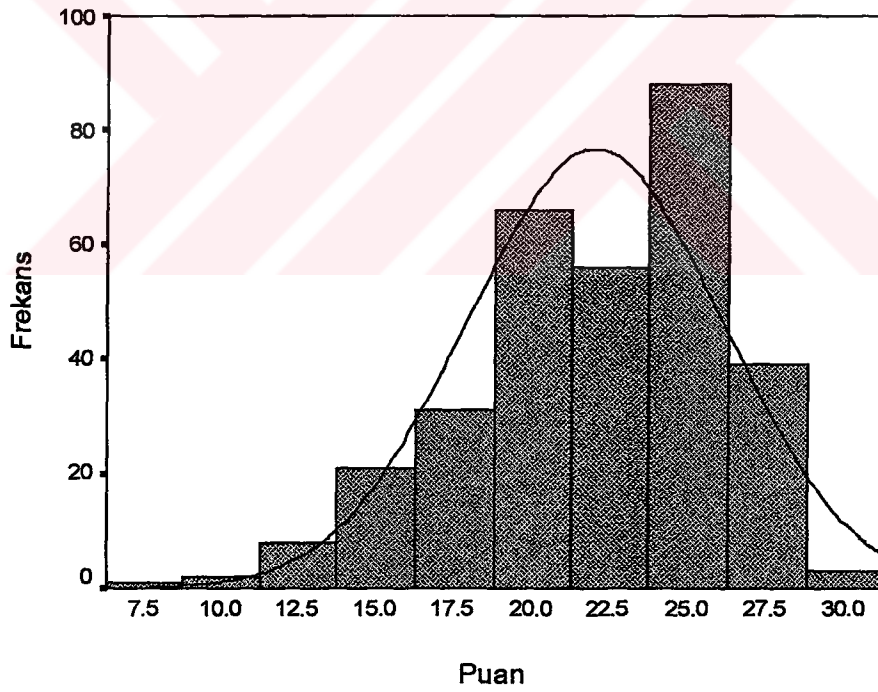
Betimsel İstatistikler	Eleme	Geleneksel
Aritmetik Ortalama	68.71	21.97
Ortanca	72	23
Varyans	628.89	16.74
Standart Sapma	25.08	4.09
En Küçük	10	8
En Büyük	116	29
Ranj	106	21
Çarpıklık	-0.32	-0.61
Basıklık	-0.68	-0.05
En yüksek Frekans	9	37
Madde sayısı	30	30
Birey sayısı	315	315

Tablo 2’de verilen puan dağılımlarının özellikleri incelendiğinde geleneksel puanlamanın eleme puanlamasına göre daha homojen puanlar verdiği ortalama ve standart sapma değerlerinden anlaşılabilmektedir. Her iki dağılımda da ortanca değerlerinin aritmetik ortalamadan büyük olduğu gözükmemektedir. Bu çarpıklık katsayısının eksi değerlerde olmasından da görülebilmektedir. Geleneksel puanlamanın eleme puanlamasına göre biraz daha fazla sola çarpık olduğu katsayıların büyüklüklerinden anlaşılabilmektedir. Geleneksel puanlamanın basıklığı normal dağılımınkine yakın iken, eleme puanlaması daha basıktır. Eleme puanlamasının daha basık olması en yüksek frekansın daha küçük olmasından kaynaklanmaktadır. Bu şekil 4 ve şekil 5’te verilen grafiklerde açıkça görülmektedir. Grafikler incelendiğinde, geleneksel uygulamada en yüksek frekans değeri 90’a yaklaşırken eleme uygulamasında 30’a yaklaşmaktadır.

Puanların dağılımına ilişkin histogram grafikleri dağılımların görüntüsünü vermesi açısından önemli olduğu düşünülmüş ve aşağıda şekil 4 ve şekil 5’te verilmiştir.



**Şekil 4: Eleme Puanlaması Puan Dağılımı**



**Şekil 5: Geleneksel Puanlama Puan Dağılımı**

## Madde Tepki Kuramının Sayıtlarının Test Edilmesi

Araştırmanın ikinci alt probleminde, MTK'ya dayalı olarak çoktan seçmeli sözel yetenek testinin geleneksel ve eleme puanlamalarının psikometrik özelliklerinin karşılaştırılması yapıldığı için, verilerin sayıtları sağlayıp sağlamadığına ilişkin *tek boyutluluk, yerel bağımsızlık ve model veri uyumu* kontrol edilmiştir. Bu sayıtlara ilişkin bulgular ilgili başlıklar altında aşağıda verilmiştir.

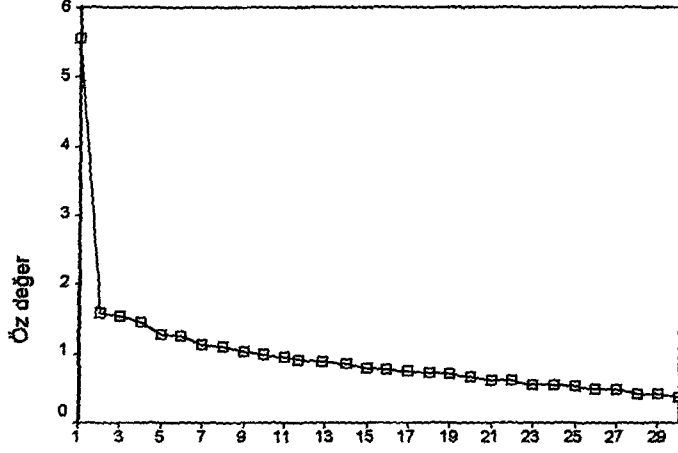
### Tek Boyutluluk

Madde Tepki Kuramının varsayımlarından biri tek boyutluluğun sağlanmasıdır. Bu araştırma için hazırlanan sözel yetenek testinin tek boyutluluk özelliğini sağlayıp sağlamadığını belirlemek amacıyla temel bileşenler faktör analizi tekniği kullanılmıştır. Faktör analizi sözel yetenek testinin hem eleme puanlaması hem de geleneksel puanlama uygulamalarından elde edilen verilere uygulanmıştır.

Faktör analizi sonucunda ölçeğin tek boyutluluk özelliğini sağlayıp sağlamadığına karar vermek için Lord (1980) tarafından önerilen (1) ilk boyut ikinciyle karşılaştırıldığında büyük, (2) ikinci boyut diğerleriyle karşılaştırıldığında aralarında fazla büyük değilse ölçeğin yaklaşık olarak tek boyutlu olduğuna karar verilebileceği ölçütleri alınmıştır. Araştırmada kullanılan sözel yetenek testinin geleneksel ve eleme uygulamalarından elde edilen verilerin tek boyutluluğuna karar vermede, yukarıda verilen ölçütler kullanılmıştır.

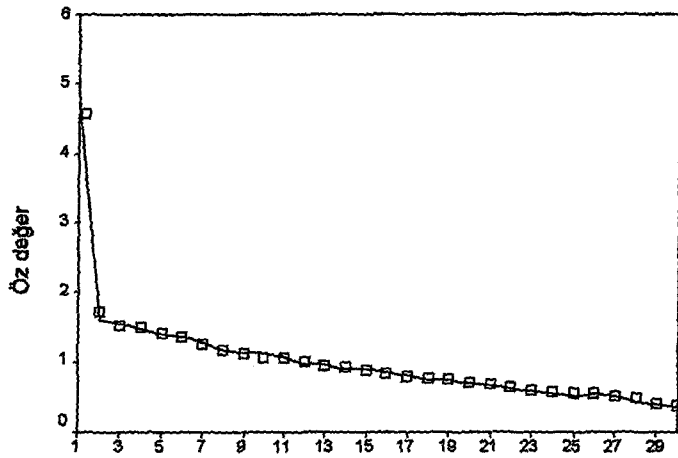
Eleme puanlamasının verilerine uygulanan faktör analizi sonucunda birinci faktörün öz değeri 5.62, ikinci faktörün öz değeri 1.56, üçüncü faktörün öz değeri 1.54, dördüncü faktörün öz değeri 1.46 ve bu değerlere yakın olarak devam etmektedir. Birinci faktör varyansın %19.05'ini, ikinci faktör %5.20, üçüncü faktör %5.14'ünü, dördüncü faktör %4.89'unu açıklamaktadır. Birinci faktörün ikinci faktörün üç katı büyüklüğünde olması ve diğer faktörlerin öz değerlerinin ikinci faktöre yakın olması bu ölçeğin tek

boyutlu olarak kabul edilebileceğini gösterir. Eleme puanlamasının faktör analizi sonucu elde edilen öz değerlere ait grafik şekil 6'da verilmiştir.



**Şekil 6: Eleme puanlamasına Ait Öz Değer Grafiği**

Yapılan faktör analizi sonucunda geleneksel puanlama için birinci faktör öz değeri 4.65, ikinci faktör öz değeri 1.505, üçüncü faktör öz değeri 1.482 ve diğer faktörlerin öz değerleri de ikinci ve üçüncü faktörlere yakın olarak devam etmektedir. Burada birinci faktör varyansın %15.22'sini, ikinci faktör %4.33'ünü, üçüncü faktör % 4.02'sini ve bu şekilde devam etmektedir. Birinci faktörün öz değeri ikincinin üç katını geçmekte ve ikinci faktörün diğerlerinden büyük olmaması bu ölçeğin tek boyutlu olduğunu kabul edebileceğimizi gösterir. Geleneksel puanlamanın faktör analizi sonucu elde edilen öz değerlere ait grafik şekil 7'de verilmiştir.



**Şekil 7: Geleneksel puanlamaya Ait Öz Değer Grafiği**

Araştırmada kullanılan sözel yetenek testinin geleneksel ve eleme uygulamalarından elde edilen verilerin Lord (1980) tarafından önerilen ölçütleri sağlaması dolayısıyla, bu ölçek tek boyutlu olarak kabul edilmiş ve MTK'nın tek boyutluluk sayılığını karşıladığına karar verilmiştir.

### **Yerel Bağımsızlık**

Madde Tepki Kuramında test maddelerinin cevaplanması birbirinden bağımsız olmalıdır. Yerel bağımsızlığın sağlanabilmesi için testin tek boyutlu olması gerekir (Hambleton and Swaminathan, 1985). Sabit bir yetenek düzeyinde, ölçülmek istenen özelliğe iki veya daha fazla özellik karışıyorsa, maddelere verilen cevaplar birbiriyle ilişkili olacaktır. Bir özellik yönünden sabit olan yetenek, diğer özellikler yönünden sabit olmayabileceği için maddelerin cevaplanması bağımsızlığını yitirecektir. Bu sebeple tek boyutluluğun sağlanması yerel bağımsızlık için yeterli görülmüştür.

### **Model – Veri Uyumu**

Geleneksel puanlamada doğru cevaba 1, yanlış ve boş cevaplara 0 puan verilmesi suretiyle puanlama yapıldığından hangi iki kategorili modelin kullanılması gerektiğinin belirlenmesi gerekmektedir. Bir parametrelili, iki parametrelili veya üç parametrelili lojistik modelin hangisinin kullanılacağını belirlemek için  $-2 \log$  Likelihood istatistikleri karşılaştırılmıştır. Bu istatistiklerin farkı 30 serbestlik derecesinde, 0.01 düzeyinde  $\chi^2$  kritik değeri ile karşılaştırılmış ve fark anlamlı ise bir üst modelin tercih edilmesi şeklinde belirlenmiştir (Embretson and Reise, 2000).

Yapılan analiz sonucunda  $-2 \log$  Likelihood değerleri Rasch modeli için 4846.2, 2PL model için 4763.7 ve 3PL model için 4728.2 olarak hesaplanmıştır. Rasch modeli ile 2PL modelinin  $-2 \log$  Likelihood farkı 82.5 bulunmuştur. Bu fark  $\chi_{(30;0.01)}^2 = 50.89$  değerinden yüksek olduğu için 2PL modelin Rasch modelinden daha iyi uyum verdiği kararına varılmıştır. 2PL model ile 3PL model için fark alındığında 35.5 olarak

hesaplanmış ve fark anlamlı bulunmamıştır. 3PL modelin 2PL modelden daha iyi olmadığı kararına varılarak 2PL model kullanılmasına karar verilmiştir.

Eleme puanlaması için 2PL modele uygun olarak iki parametrelî Genelleştirilmiş Kısmî Puan modeli kullanılmıştır. Genelleştirilmiş Kısmî Bilgi Modeli için hesaplanan  $-2 \log$  Likelihood değeri 14523.7 olarak bulunmuştur.

### **Alt Problemlerin Çözümlemesinde Kullanılan İstatistiksel Analizler**

Araştırmada, geleneksel puanlama ve eleme puanlamasının puan dağılımları ve varsayımların test edilmesinin ardından alt problemlerin çözümlenmesi işlemlerine başlanmıştır. Alt problemlerin çözümlenmesi süreci I. bölümdeki sıralamaya göre aşağıda verilmiştir.

Hazırlanan sözel yetenek testinin geleneksel uygulamasından elde edilen verilerin Klasik Test Kuramı analizleri ITEMAN programı kullanılarak yapılmıştır. Eleme puanlamasından elde edilen verilerin Klasik Test Kuramı analizleri ise LISREL 8.54 paket programı kullanılarak analiz edilmiştir. Madde Tepki Kuramı analizlerinin yapılmasında ise hem geleneksel puanlama verileri için hem de eleme puanlaması verileri için MULTILOG (Thissen, 1991) programı kullanılmıştır. MULTILOG 7.03 programı hem 1-0 şeklinde puanlanan verilerle 1, 2 ve 3 parametrelî lojistik modellerin analizi, hem de çok kategorilî IRT modelleri için uygundur.

Araştırmanın birinci alt problemi olan çoktan seçmeli testlerin geleneksel uygulaması ile eleme puanlaması uygulamasından elde edilen puanlamaların a) madde ayırıcılık indeksleri, b) belirli bir güvenilirliği sağlayan madde sayıları oranı, c) test güvenilirlikleri karşılaştırması ve d) iki puanlama türünde elde edilen yetenek düzeyleri arasındaki tutarlılık için öncelikle maddelerin ve testin istatistikleri kestirilmiştir.

Araştırmanın birinci alt probleminin a maddesinde, hazırlanan sözel yetenek testinin geleneksel uygulaması ile eleme uygulamasından elde edilen verilerin analizinden elde

edilen madde ayırdedicilik indekslerinin farklılaşmasının incelenmesinde, öncelikle her iki puanlama türünde elde edilen madde ayırıcılık indeksleri hesaplanmıştır. Madde ayırıcılık indekslerinin hesaplanmasında geleneksel puanlama için çift serili korelasyon katsayıları, eleme puanlaması için çok serili (polyserial) korelasyon katsayıları hesaplanmıştır. Elde edilen madde ayırıcılık katsayıları “çok yüksek”, “yüksek”, “orta”, “düşük” ve “çok düşük” olarak gruplanmıştır. İki puanlama türünde sınıflama sonuçları karşılaştırılmıştır. Sınıflama için 0.00-0.20 arası katsayılar “çok düşük”, 0.21-0.40 arası “düşük”, 0.41-0.60 “orta”, 0.61-0.80 arası “yüksek”, 0.81-1.00 arası “çok yüksek” ölçütleri kullanılmıştır (Tekin, 1993). Sınıflama sonucunda hangi puanlama türünde daha yüksek sınıflara daha çok madde girdiğine göre yorum yapılacaktır. Bunun yanında, her iki puanlama türüne ait katsayıların farkının anlamlılığı Fischer’in z istatistiği ile test edilmiştir. Test sonucunda geleneksel puanlama ve eleme puanlaması verilerinden elde edilen madde ayırıcılık katsayıları farklı bulunan maddelerde farkın hangi puanlama türünün lehine olduğu belirlenmiştir. Ayırıcılık güçleri arasında anlamlı fark bulunan maddelerin kaç tanesinde geleneksel puanlama , kaç tanesinde eleme puanlaması lehine olduğuna göre yorum yapılmıştır.

Araştırmanın birinci alt probleminin b maddesinde, hazırlanan sözel yetenek testinin geleneksel puanlaması ile eleme puanlaması uygulamasında aynı güvenilirliğe ulaştıkları madde sayıları oranının belirlenmesinde, öncelikle puanlama türlerinden elde edilen madde ayırıcılık indeksleri en yüksekten en düşüğe doğru sıralanmıştır. Alt problemde belirlenen güvenilirlik düzeyi olarak belirtilen ifade her iki puanlama türünün 30 maddelik test güvenilirliğinden düşük olan 0.70 düzeyi belirlenmiştir. ayırıcılık gücü en yüksek maddeden başlayarak 0.70 güvenilirlik elde edilinceye kadar madde eklenmiştir. Her iki puanlama türünde 0.70 güvenilirlik katsayıları elde edilen madde sayıları belirlenmiş ve birbirine oranlanmıştır. Bu oran göreceli uzunluk etkililiği olarak değerlendirilmiştir.

Araştırmanın birinci alt probleminin c maddesinde, hazırlanan sözel yetenek testinin geleneksel puanlama ve eleme puanlaması uygulamalarının iç tutarlılık güvenilirlik katsayıları karşılaştırılmıştır. puanlamalarının doğasından dolayı, geleneksel puanlama için KR-20, eleme puanlaması için Cronbach  $\alpha$  güvenilirlik katsayıları hesaplanmıştır.

Hesaplanan güvenilirlik katsayılarının farkının anlamlılığı için Fischer'in z istatistiği kullanılmıştır. İki güvenilirlik katsayısının farkının anlamlılığı için 0.01 düzeyi belirlenmiştir.

Araştırmanın birinci alt probleminin d maddesinde aynı sözel yetenek testinin geleneksel uygulanmasından elde edilen toplam puanlar ile eleme uygulamasından elde edilen toplam puanlar arasında ne derece tutarlılık olduğu incelenmiştir. Bu tutarlılığın belirlenmesinde, iki puanlama türünde elde edilen toplam puanlar arasında Pearson korelasyon katsayısı hesaplanmıştır. Korelasyon katsayısının 0.01 anlamlılık düzeyinde sıfırdan farklı olup olmadığı Fischer'in z istatistiği ile test edilmiştir. Bulunan korelasyon katsayısı iki puanlama türünün tutarlılığı olarak yorumlanmıştır.

Araştırmanın ikinci alt problemde çoktan seçmeli testlerin geleneksel puanlama ve eleme puanlaması uygulamalarından elde puanların Madde Tepki Kuramına göre kestirilen a) madde ve test bilgi düzeylerinin göreceli etkililikleri (Relative Efficiency) b) marjinal güvenilirlikleri c) kestirilen yetenek düzeylerinin ne derecede tutarlı olduğu incelenmiştir.

Araştırmanın ikinci alt probleminin a maddesinde çoktan seçmeli testlerin geleneksel puanlama ve eleme puanlaması verileri Madde Tepki Kuramına göre analizleri yapılmış ve madde ve test bilgi düzeyleri kestirilmiştir. Geleneksel puanlama için iki parametrelili lojistik model, eleme puanlaması için ise iki parametrelili Genelleştirilmiş Kısmi Puan modeli kullanılmıştır. Elde edilen madde ve test bilgi düzeyleri -3 ile +3 arasında tam sayılardaki yetenek düzeylerinde madde bilgi düzeyleri oranı hesaplanmıştır. Bu oran eşitlik 19'da verildiği gibi göreceli etkililik olarak alınmış ve yorumlanmıştır. Bu oran, eleme puanlamasının ilgili yetenek düzeyinde sağladığı bilginin geleneksel puanlamanın ilgili yetenek düzeyindeki sağladığı bilgi düzeyine bölünmesiyle elde edilmiştir. Dolayısıyla eleme puanlamasının geleneksel puanlamaya göre sağladığı göreceli etkililik olarak yorumlanmıştır. Ayrıca yorumu zenginleştirebilmek için grafiklerle karşılaştırma yapılmıştır..

Araştırmanın ikinci alt probleminin b maddesinde aynı sözel yetenek testine ait geleneksel puanlama ve eleme puanlamasının Madde Tepki Kuramına göre kestirilen marjinal güvenilirlik katsayıları karşılaştırılmıştır. Marjinal güvenilirlik katsayısı kestiriminde MULTİLOG 7.03 programı kullanılmıştır. Her iki puanlama türüne göre elde edilen güvenilirlik katsayılarının farklı olup olmadığı test edilmiştir. Güvenirlik katsayılarının farklı olup olmadığına ilişkin test için Fischer'in z istatistiği kullanılmıştır.

Araştırmanın ikinci alt probleminin c maddesinde aynı sözel yetenek testinin geleneksel puanlama uygulanmasından elde edilen yetenek düzeyleri ile eleme puanlaması uygulamasından elde edilen yetenek düzeyleri arasında ne derece tutarlılık olduğu incelenmiştir. Bu tutarlılığın belirlenmesinde, iki puanlama türünde kestirilen yetenek düzeyleri arasında Pearson korelasyon katsayısı hesaplanmıştır. Korelasyon katsayısının 0.01 anlamlılık düzeyinde sıfırdan farklı olup olmadığı Fischer'in z istatistiği ile test edilmiştir. Bulunan korelasyon katsayısı iki puanlama türünün tutarlılığı olarak yorumlanmıştır.

Araştırmanın üçüncü alt probleminde aynı sözel yetenek testinin geleneksel ve eleme uygulamasına katılan öğrencilerin bu iki uygulama türünün karşılaştırılmasına ilişkin görüşleri belirlenmeye çalışılmıştır. Görüşlerini belirlemek için öğrencilere ek 1'de verilen beş maddelik bir anket uygulanmıştır. Ankete verilen cevapların frekans ve yüzdelerine bakarak cevaplayıcıların iki uygulamaya ilişkin görüşleri belirlenmeye çalışılmıştır.

## BÖLÜM III

### BULGULAR VE YORUMLAR

Bu bölümde, araştırmanın alt problemlerinin birinci bölümdeki verilmiş sırası dikkate alınarak elde edilen bulgularına ve bu bulgulardan yapılan yorumlara yer verilmiştir.

#### ALT PROBLEM 1: ÇOKTAN SEÇMELİ TESTLERDE GELENEKSEL PUANLAMA VE ELEME PUANLAMASININ KLASİK TEST KURAMINA GÖRE KARŞILAŞTIRMALAR

##### a) Madde Ayırıcılık Güçlerinin Karşılaştırılması

Araştırmanın birinci alt probleminin a maddesinde, çoktan seçmeli testlerde geleneksel puanlama ve eleme puanlaması uygulamalarının Klasik Test Kuramına göre kestirilen madde ayırıcılık güçleri arasındaki farklar açıklanmaya çalışılmıştır.

Madde ayırıcılık güçlerinin karşılaştırılmasında farklı yollar denenmiştir. Bunlardan birincisi ayırıcılık güçlerinin aynı zamanda korelasyon katsayısı olması sebebiyle, iki korelasyon katsayısının farkının anlamlılığı testidir. Bu fark Fischer'in z istatistiği ile test edilmiştir. İkinci karşılaştırma madde ayırıcılık güçleri sınıflandırılarak yapılmıştır. Geleneksel ve eleme uygulamalarından elde edilen ayırıcılık güçlerinin büyüklüğüne göre sınıflamaları yapılmış ve her iki puanlama türünde sınıflara dağılım incelenmiştir. Üçüncü olarak ise, test geliştirmede genel olarak ayırıcılık için alınan 0.30'dan büyük olma ölçütünü her iki puanlama türünde kaç maddenin sağladığına bakılmıştır.

Geleneksel ve eleme puanlamasına ilişkin madde test korelasyonları ve bu korelasyonların z değeri ile farklarına ilişkin z istatistiği sonuçları tablo 3'te verilmiştir.

Tablo 3'ün ikinci sütununda görülebileceği gibi bu çalışmada kullanılan sözel yetenek testinin maddelerinin geleneksel puanlamada madde – test korelasyonuna dayalı olarak kestirilen madde ayırıcılık güçleri ortancası 0.525 olmak üzere, 0.289 (madde 2) ile 0.783 (madde 16) arasında değişmektedir. Çoktan seçmeli sözel yetenek testinin geleneksel puanlamada madde ayırıcılık gücü olarak sadece ikinci madde 0.30'un altında kalmaktadır.

Tablo 3

**Sözel Yetenek Testinin Geleneksel ve Eleme Puanlaması Uygulamalarından Elde Edilen Madde Ayırıcılık Gücü Değerleri ve Farklarına İlişkin z Testi Sonuçları**

Maddeler	Geleneksel		Eleme		Z	p
	$r_{jx}$	$z_r$	$r_{jx}$	$z_r$		
1	0.446	0.479	0.461	0.498	-0.238	0.812
2	0.289	0.298	0.570	0.648	-4.392**	0.000
3	0.445	0.478	0.418	0.446	0.402	0.688
4	0.555	0.626	0.324	0.336	3.639**	0.000
5	0.572	0.65	0.532	0.592	0.728	0.467
6	0.545	0.611	0.524	0.582	0.364	0.716
7	0.664	0.800	0.565	0.64	2.008*	0.045
8	0.590	0.678	0.288	0.479	2.497*	0.013
9	0.402	0.426	0.542	0.606	-2.259*	0.024
10	0.639	0.757	0.300	0.31	5.610**	0.000
11	0.603	0.699	0.659	0.792	-1.167	0.243
12	0.600	0.693	0.499	0.548	1.820	0.069
13	0.602	0.695	0.383	0.404	3.652**	0.000
14	0.583	0.667	0.331	0.344	4.054**	0.000
15	0.414	0.441	0.397	0.42	0.264	0.792
16	0.783	1.052	0.491	0.537	6.463**	0.000
17	0.353	0.369	0.608	0.707	-4.242**	0.000
18	0.548	0.616	0.404	0.429	2.347*	0.019
19	0.631	0.742	0.376	0.395	4.355**	0.000
20	0.591	0.679	0.480	0.523	1.958	0.050
21	0.395	0.418	0.387	0.408	0.125	0.900
22	0.515	0.57	0.425	0.454	1.456	0.145
23	0.419	0.447	0.278	0.286	2.021*	0.043
24	0.415	0.442	0.377	0.396	0.577	0.564
25	0.481	0.524	0.286	0.294	2.886**	0.004
26	0.534	0.596	0.537	0.599	-0.038	0.970
27	0.479	0.522	0.552	0.62	-1.230	0.219
28	0.335	0.348	0.194	0.197	1.895	0.058
29	0.297	0.42	0.352	0.367	0.665	0.506
30	0.406	0.431	0.244	0.64	-2.623**	0.009
Ortanca	0.525	0.583	0.411	0.437	1.431	0.153

Tablo 3'ün dördüncü sütununda görüldüğü gibi, bu araştırmada kullanılan sözel yetenek testinin eleme puanlaması uygulamasında madde – test korelasyonuna dayalı olarak

kestirilen madde ayırıcılık güçleri ortancası 0.411 olmak üzere, 0.194 (madde 28) ile 0.659 (madde 11) arasında değişmektedir. Madde ayırıcılık gücü 0.30'un altında kalan madde sayısı 5'tir.

Tablo 3'te görülebileceği gibi geleneksel ve eleme puanlaması uygulamalarının sonucunda 15 maddenin ayırıcılık güçleri 0.05 düzeyinde anlamlı olarak farklılaşmaktadır. 15 maddenin 11'inde geleneksel puanlama lehine bir fark bulunurken, 4 maddede eleme puanlaması lehine bir fark bulunmuştur. Bu araştırmada kullanılan sözel yetenek testinin 30 maddeden oluştuğu göz önüne alındığında, madde ayırıcılık güçleri arasında maddelerin yarısında farklılık anlamlı iken diğer yarısında anlamlı bulunmamıştır. Genel anlamda, geleneksel puanlama metodunun kullanıldığı uygulamada eleme puanlaması uygulamasından daha yüksek ayırıcılık gücü değerleri kestirilmiştir.

Eleme uygulamasının önerilmesindeki gerekçelerden biri bireyler arasındaki ayrımı daha iyi yapabileceği düşüncesi idi. Yüksek yetenek seviyesindeki bireyleri daha düşük yetenek seviyesindeki bireylerden madde ölçeğinde daha iyi ayırt edebileceği düşünülmekteydi. Yukarıda verilen bulgular bu düşüncüyü desteklememektedir. Geleneksel puanlama, eleme puanlamasına göre daha düşük ayırıcılık katsayıları vermiştir.

Geleneksel ve eleme puanlaması uygulamalarından elde edilen madde ayırıcılık güçlerinin "Çok Düşük", "Düşük", "Orta", "Yüksek" ve "Çok Yüksek" sınıflamasına ilişkin sonuçlar tablo 4'te verilmiştir.

**Tablo 4**

**Geleneksel ve Eleme Puanlaması Madde Ayırıcılık Güçleri Sınıflaması**

Sınıf	Madde Sayısı	
	Geleneksel	Eleme
Çok Düşük (0.00-0.19 Arası)	-	1
Düşük (0.20-0.39 Arası)	4	11
Orta (0.40-0.59 Arası)	19	14
Yüksek (0.60-0.79 Arası)	7	4

Tablo 4'te görülebileceği gibi geleneksel puanlamada madde ayırıcılık güçleri 4 “Düşük”, 19 “Orta” ve 7 “Yüksek” olarak sınıflanmaktadır. Eleme puanlamasında ise 1 “Çok düşük”, 11 “Düşük”, 14 “Orta” ve 4 “Yüksek” olarak sınıflanmaktadır. Geleneksel puanlamada hiç “Çok Düşük” ayırıcılık gücüne sahip madde yok iken eleme puanlamasında bir tane vardır. Geleneksel puanlamada “Düşük” ayırıcılık gücüne sahip madde 4 tane iken eleme puanlamasında bu sayı 11'dir. Açıkça görüldüğü gibi geleneksel puanlamada eleme puanlamasına göre ayırıcılık güçleri bakımından daha üst sınıflarda daha fazla madde içermektedir.

Test geliştirmede ayırıcılık olarak maddenin üst ve alt gruptaki bireyleri birbirinden ayırmada yeterli olduğunu gösteren ölçüt olarak kabul edilebilen 0.30 ve daha yukarı olma (Crocker and Algina, 1986) yönünden de karşılaştırma yapılmıştır. Geleneksel puanlamada sadece 2 madde 0.30 ölçütüne ulaşamazken, bu sayı eleme puanlamasında 5 madde olarak görülmektedir.

Madde ayırıcılık gücü yönünden farklı karşılaştırma sonuçları dikkate alındığında geleneksel puanlama eleme puanlamasına göre daha yüksek ayırıcılık verdiği açıkça görülmektedir.

#### **b) Belirli Bir Güvenirliği Sağlayan Madde Sayıları Oranının Belirlenmesi**

Geleneksel ve eleme puanlaması uygulamalarının Klasik Test Kuramı'nda birbirine göre etkililiğini belirlemek amacıyla belirli bir güvenirliği sağlayan madde sayıları oranı hesaplanmıştır. Belirlenecek güvenirlilik katsayısının her iki uygulamada tüm testten elde edilen güvenirlilik katsayılarından küçük bir değer olması gerektiği için bu değer 0.70 olarak seçilmiştir. Çoktan seçmeli testlerin geleneksel ve eleme uygulamasında 0.70 güvenirlilik katsayısını sağlayan madde sayıları, ayırıcılık yönünden en iyi maddeler seçilmek suretiyle belirlenmiştir. Güvenirlilik katsayısının belirlenmesinde iç tutarlılık anlamında güvenirliği veren KR-20 ve Cronbach  $\alpha$  katsayıları kullanılmıştır.

Geleneksel puanlamada en iyi 10 maddeden başlayarak, birer madde eklenmiş ve 0.70 güvenirlilik düzeyine ulaşınca kadar madde eklenmeye devam edilmiştir. Eklemeler sonucunda 15 maddede 0.704 güvenirlilik düzeyine ulaşılmıştır. Bu madde sayısı geleneksel puanlamanın 0.70 güvenirliği sağlayan madde sayısı olarak alınmıştır.

Eleme puanlamasında, tüm test için daha yüksek güvenilirlik verdiği için, en iyi 5 maddeden başlayarak, birer madde eklenmiş ve 0.70 güvenilirlik düzeyine ulaşınca kadar eklenmeye devam edilmiştir. Eklemeler sonucunda 8 maddede 0.709 güvenilirlik düzeyine ulaşılmıştır. Bu madde sayısı da eleme puanlamasının 0.70 güvenilirlik düzeyini sağlayan madde sayısı olarak belirlenmiştir.

Geleneksel ve eleme puanlamasının 0.70 güvenilirlik düzeyini sağlayan madde sayıları oranı olarak  $15 / 8 = 1.88$  değeri bulunmuştur. Bu durum, eleme puanlamasının geleneksel puanlamanın 1.88 katı daha etkili olabileceği olarak yorumlanabilir. Bu bulgu, eleme puanlamasının geleneksel puanlamaya göre daha az maddeyle aynı güvenilirlikte ölçümler yapabileceğini göstermektedir. Daha az maddeyle aynı güvenilirlikte ölçümler yapabilme, zaman ve maliyet yönünden ekonomi sağlayabilme özelliği kazandırabilir. Bu özellik geniş ölçekli test uygulaması gerektiren durumlar için önemli bir üstünlük olarak kabul edilebilir.

### **c) Test Güvenirlikleri Karşılaştırması**

Araştırmanın birinci alt probleminin c maddesinde, kullanılan sözel yetenek testinin geleneksel puanlama ve eleme puanlaması uygulamalarından iç tutarlılık güvenilirlik katsayıları kestirilmeye çalışılmıştır. Testin geleneksel uygulamasında, elde edilen verilerin doğasından dolayı güvenilirlik kestiriminde KR-20 güvenilirlik katsayısı, eleme uygulamasından ise yine verilerin doğasından dolayı Cronbach  $\alpha$  katsayısı kullanılmıştır.

Araştırmada kullanılan 30 çoktan seçmeli maddeden oluşan sözel yetenek testinin Klasik Test Kuramı'na göre, geleneksel uygulamasında KR-20 güvenilirlik katsayısı 0.737 olarak bulunmuştur. Kullanılan 30 maddelik sözel yetenek testinin eleme uygulamasında Cronbach  $\alpha$  güvenilirlik katsayısı 0,829 olarak bulunmuştur.

Sözel yetenek testinin geleneksel ve eleme uygulamalarından elde edilen güvenilirlik katsayılarının karşılaştırılmasında Fischer'in z istatistiği kullanılmıştır. Güvenirlik indeksinin tanımında paralel iki test arasındaki korelasyon olduğundan, güvenilirlik katsayılarını karşılaştırmak için Fischer'in z istatistiği kullanılmıştır. Geleneksel ve

eleme uygulamalarından kestirilen güvenilirlik katsayılarının karşılaştırılmasına ilişkin z testi tablo 5'te verilmiştir.

**Tablo 5**

**Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Kestirilen  
Güvenirlik Katsayılarının karşılaştırılması**

	<b>Geleneksel</b>	<b>Eleme</b>	<b>Z</b>	<b>p</b>
<b>Güvenirlik</b>	0.737	0.829	-3.025**	0.002
<b>Fischer z Karşılığı</b>	0.944	1.185		

\*\*  $p < 0.01$

Tablo 5'te görülebileceği gibi geleneksel ve eleme uygulamasının karşılaştırılması sonucunda -3.025 z değeri elde edilmiştir. Bu z değerinin iki yönlü p değeri 0.002 olarak belirlenmiştir. Kullanılan çoktan seçmeli sözel yetenek testinin, eleme puanlamasından kestirilen güvenilirlik katsayısı geleneksel puanlamadan kestirilen güvenilirlik katsayısından 0.01 düzeyinde daha yüksek olduğu söylenebilir.

Bu bulgu, eleme puanlamasından sağlanabileceği düşünülen fayda olarak, şans başarısının engellenebileceği şeklinde düşünülebilir. Eleme puanlamasının, geleneksel puanlamaya göre, daha yüksek güvenilirlik katsayısı vermesi, eleme puanlamasından elde edilen verilere daha az hata karışmasından kaynaklanabileceği düşüncesini akla getirmektedir. Bu hata azalmasının olası en belirgin iki sebebinden birisi şans başarısı, bir diğeri de madde ölçeğinde daha duyarlı ölçümler yapabilme özelliğinden kaynaklanabilir. Eleme puanlamasının iddia edildiği gibi şans başarısını engelleme ve kısmi bilgiyi ölçme potansiyeline sahip olabileceği sonucu çıkarılabilir.

**d) Toplam Puanlar Arasındaki İlişki Düzeyinin Belirlenmesi**

Araştırmanın birinci alt probleminin d maddesinde, araştırmanın örneklemini oluşturan bireylerin kullanılan sözel yetenek testinin geleneksel ve eleme uygulamalarından aldıkları toplam puanlar arasındaki ilişki belirlenmeye çalışılmıştır. Klasik Test Kuramı'na göre bireylerin yetenek ölçüleri madde puanları toplamı veya bunların

standartlaştırılmış hali olarak kestirilmektedir. İster toplam puan alınsın ister standart puanlar alınsın ilişki değişmeyeceği için, ham puanlar üzerinden ilişki hesaplanmıştır.

Bireylerin çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamalarından elde edilen toplam puanlara ilişkin betimsel istatistikler tablo 6’da verilmiştir.

**Tablo 6**  
**Kullanılan Sözel Yetenek Testinin Geleneksel ve Eleme Puanlaması Uygulaması Yapılan Gruptan Elde Edilen Yetenek Ölçülerine İlişkin Betimsel İstatistikler**

	<b>Geleneksel Puanlama</b>	<b>Eleme Puanlaması</b>
En Düşük Puan	8	10
En Yüksek Puan	29	116
Aritmetik Ortalama	21.97	68.71
Ortanca	23	72
Standart Sapma	4.09	25.08
Bağıl Değişim Katsayısı	18.62	36.50
Toplam Puanlar Arasındaki İlişki	0.565**	

\*\*  $p < 0.01$

Tablo 6’da verilen betimsel istatistikler incelendiğinde, sözel yetenek testindeki performans açısından en yüksek puanlar her iki puanlamada da 29 tam doğru cevaba denk gelmektedir. En düşük puanlarda ise geleneksel puanlamada 8 tam doğru cevaba denk gelirken, eleme puanlamasında 2.5 tam doğru cevaba denk gelmektedir. Aritmetik ortalamalar ve standart sapmalar dikkate alındığında veya bağıl değişim katsayıları incelendiğinde eleme puanlamasının geleneksel puanlamaya göre daha heterojen olduğu açıkça görülmektedir.

Örnekleme katılan bireylerin geleneksel puanlama ve eleme puanlamasındaki yetenek ölçüleri arasındaki ilişkinin belirlenmesi amacıyla, geleneksel ve eleme puanlamalarından aldıkları toplam puanlar arasındaki ilişki Pearson momentler çarpımı korelasyon katsayısı ile belirlenmiştir.

Bireylerin iki puanlama yöntemine göre kestirilen yetenek düzeyleri arasındaki ilişki 0.565 olarak hesaplanmıştır. Bu katsayı 0.01 anlamlılık düzeyinde sıfırdan farklıdır. Bu katsayı ilişki düzeyi olarak orta düzeydedir. Bu korelasyon katsayısının karesi alındığında,  $r^2=0.32$  belirtme katsayısı bulunacaktır. Bu katsayı iki puanlama türünün birbirlerindeki değişimin % 32'sini açıklayabildiğini göstermektedir. Birbirlerini açıklayamadıkları oran ise % 68'dir. Bu iki puanlama türünün oldukça farklı puanlar ürettiklerini söyleyebiliriz.

Ayrıca testte yer alan maddelerin geleneksel ve eleme puanları ortalamaları arasında 0.95 düzeyinde bir korelasyon hesaplanmıştır. Bu da bize maddelerin güçlük düzeylerinin birbirlerine çok benzediklerini göstermektedir. Güçlük düzeylerindeki sıralamanın korunmasına rağmen bireylerin puanlarında büyük değişimler gözlenmiştir.

## **ALT PROBLEM 2: ÇOKTAN SEÇMELİ TESTLERDE GELENEKSEL PUANLAMA VE ELEME PUANLAMASININ MADDE TEPKİ KURAMINA GÖRE KARŞILAŞTIRILMASI**

### **a) Belirlenen 7 Yetenek Düzeyinde Madde Bilgi Düzeylerinin Göreceli Etkililiğinin (Relative Efficiency) Belirlenmesi**

Araştırmanın ikinci alt probleminin a maddesinde belirlenen yedi yetenek düzeyinde geleneksel ve eleme puanlaması uygulamalarından elde edilen madde puanlarından madde ve test bilgi düzeyleri kestirilmiştir. Bu çalışmada kullanılan ve 30 maddeden oluşan sözel yetenek testinin geleneksel uygulama sonucunda elde edilen verilerine MULTILOG programında 2 parametrelili lojistik model kullanarak analiz yapılarak madde ve test bilgi düzeyleri kestirilmiştir. Aynı şekilde, eleme testi uygulaması sonucunda elde edilen madde puanlarına MULTILOG programında Genelleştirilmiş Kısmi Puan Modeli kullanarak analiz yapılarak madde ve test bilgi düzeyleri kestirilmiştir. Kullanılan sözel yetenek testinin geleneksel ve eleme puanlamalarından kestirilen madde ve test bilgi fonksiyonları tablo 7'de verilmiştir.

Tablo 7: Belirlenen Yetenek Düzeylerinde Geleneksel ve Eleme Puanlamalarından Keşirilen Madde ve Test Bilgi Düzeyleri

Yetenek	-3		-2		-1		0		1		2		3	
	Geleneksel	Eleme	Geleneksel	Eleme	Geleneksel	Eleme	Geleneksel	Eleme	Geleneksel	Eleme	Geleneksel	Eleme	Geleneksel	Eleme
m1	0.03	0.29	0.05	0.57	0.08	1.20	0.11	1.59	0.11	0.63	0.10	0.14	0.07	0.04
m2	0.02	0.29	0.01	0.47	0.01	0.79	0.01	1.62	0.01	0.47	0.01	0.06	0.00	0.01
m3	0.06	0.12	0.08	0.20	0.09	0.61	0.09	1.30	0.07	0.52	0.05	0.12	0.03	0.03
m4	0.22	0.29	0.23	0.98	0.15	1.94	0.08	1.62	0.03	0.31	0.01	0.05	0.01	0.01
m5	0.19	0.04	0.29	0.58	0.25	3.14	0.13	2.26	0.05	0.37	0.02	0.01	0.01	0.00
m6	0.14	0.19	0.25	0.53	0.28	1.38	0.18	1.68	0.08	0.37	0.03	0.09	0.01	0.03
m7	0.08	0.17	0.27	0.53	0.52	1.52	0.38	2.14	0.13	0.48	0.04	0.08	0.01	0.03
m8	0.06	0.06	0.13	0.32	0.25	2.08	0.28	1.20	0.19	0.21	0.09	0.07	0.03	0.03
m9	0.09	0.12	0.08	0.53	0.05	4.10	0.03	6.23	0.02	0.44	0.01	0.05	0.01	0.01
m10	0.18	0.06	0.44	0.43	0.46	2.00	0.08	2.82	0.06	0.18	0.02	0.02	0.00	0.00
m11	0.39	0.04	0.51	0.12	0.26	4.08	0.21	2.68	0.02	0.22	0.00	0.02	0.00	0.00
m12	0.17	0.04	0.30	0.26	0.29	2.43	0.16	2.75	0.07	0.44	0.02	0.06	0.01	0.02
m13	0.18	0.39	0.36	0.81	0.35	2.41	0.18	3.59	0.06	0.35	0.02	0.03	0.01	0.01
m14	0.21	0.52	0.32	2.54	0.26	3.34	0.13	3.48	0.05	0.31	0.02	0.04	0.01	0.01
m15	0.08	0.51	0.08	0.35	0.07	1.47	0.06	3.76	0.04	0.43	0.03	0.04	0.02	0.00
m16	0.51	0.18	0.94	0.26	0.33	1.51	0.06	2.59	0.01	0.20	0.00	0.03	0.00	0.01
m17	0.06	0.06	0.05	0.33	0.04	1.38	0.03	4.81	0.02	0.30	0.01	0.04	0.01	0.01
m18	0.19	0.00	0.15	0.18	0.09	8.85	0.04	1.02	0.02	0.12	0.01	0.02	0.00	0.01
m19	0.21	0.01	0.40	1.92	0.35	3.83	0.16	2.50	0.05	0.30	0.02	0.05	0.00	0.02
m20	0.36	0.00	0.20	2.15	0.07	3.99	0.02	0.30	0.01	0.03	0.00	0.02	0.00	0.02
m21	0.10	0.10	0.11	1.13	0.09	2.68	0.06	2.15	0.04	0.58	0.02	0.11	0.01	0.03
m22	0.06	0.20	0.11	0.35	0.16	0.66	0.17	0.81	0.13	0.38	0.08	0.12	0.04	0.05
m23	0.01	0.15	0.02	0.26	0.04	0.41	0.07	0.74	0.11	0.90	0.13	0.53	0.12	0.25
m24	0.06	0.09	0.07	0.65	0.06	1.94	0.05	1.14	0.04	0.53	0.03	0.18	0.02	0.06
m25	0.07	0.85	0.09	1.69	0.10	0.96	0.09	1.22	0.07	0.43	0.04	0.09	0.03	0.02
m26	0.09	0.12	0.15	0.31	0.19	1.00	0.17	1.39	0.11	0.47	0.05	0.13	0.03	0.05
m27	0.08	0.25	0.12	0.54	0.14	1.00	0.12	1.28	0.08	0.52	0.05	0.12	0.03	0.04
m28	0.01	0.19	0.02	0.51	0.02	1.29	0.03	1.29	0.03	0.52	0.03	0.16	0.04	0.06
m29	0.01	0.08	0.01	0.22	0.01	0.68	0.01	1.77	0.01	0.67	0.01	0.10	0.01	0.02
m30	0.08	0.29	0.08	0.79	0.07	2.03	0.06	0.67	0.04	0.15	0.03	0.04	0.02	0.02
Test	4.00	5.68	5.88	21.51	5.11	64.70	3.21	62.42	1.74	11.61	0.95	2.59	0.55	0.88

Tablo 7’de görüldüğü gibi, geleneksel puanlamada -3 yetenek düzeyinde madde bilgi fonksiyonu değerleri 0.01 (23, 28 ve 29. maddeler) ile 0.51 (madde 16) arasında, eleme puanlamasında ise, 0.00 (18 ve 20. maddeler) ile 0.85 (madde 25) arasında değişmektedir. Yetenek düzeyi -2’de geleneksel puanlamada 0.01 (madde 2 ve 29) ile 0.94 (madde 16) arasında, eleme puanlamasında 0.12 (madde 10) ile 2.15 (madde 20) arasında değişmektedir. Yetenek düzeyi -1’de geleneksel puanlamada 0.01 (madde 2 ve 29) ile 0.52 (madde 7) arasında, eleme puanlamasında 0.41 (madde 23) ile 4.10 (madde 9) arasında değişmektedir. Yetenek düzeyi 0’da (sıfır) geleneksel puanlamada bilgi düzeyi 0.01 (madde 2 ve 29) ile 0.38 (madde 7) arasında, eleme puanlamasında 0.30 (madde 20) ile 6.23 (madde 9) arasında değişmektedir. Yetenek düzeyi +1’de geleneksel puanlamada 0.01 (madde 2, 16, 20 ve 29) ile 0.19 (madde 8) arasında, eleme puanlamasında 0.12 (madde 18) ile 0.90 (madde 23) arasında değişmektedir. Yetenek düzeyi +2’de, geleneksel puanlamada bilgi 0.00 (madde 20) ile 0.13 (madde 23) arasında, eleme puanlamasında 0.01 (madde 5) ile 0.53 (madde 23) arasında değişmektedir. Yetenek düzeyi +3’de ise, geleneksel puanlamada 0.00 (madde 2, 10, 11, 16, 18, 19 ve 20) ile 0.12 (madde 23) arasında, eleme puanlamasında da 0.00 (madde 5, 10, 11 ve 15) ile 0.25 (madde 23) arasında değişmektedir.

Test bilgi düzeylerinde elde edilen değerlere baktığımızda, geleneksel puanlama en düşük bilgiyi +3 yetenek düzeyinde, en yüksek bilgiyi -2 yetenek düzeyinde vermiştir. Bu yetenek düzeylerine ait bilgi düzeyleri de 0.55 ve 5.88 olarak kestirilmiştir. Eleme puanlamasında, en düşük test bilgisini +3 yetenek düzeyinde ve en yüksek -1 yetenek düzeyinde vermiştir. Bu yetenek düzeylerine ait bilgi değerleri sırasıyla 0.88 ve 64.7 olarak kestirilmiştir.

Maddelere ve teste göre kestirilen geleneksel ve eleme puanlamalarına ait bilgi düzeylerinin oranlanması sonucu elde edilen göreceli etkililik değerlerine ilişkin bulgular tablo 8’de verilmiştir.

Tablo 8

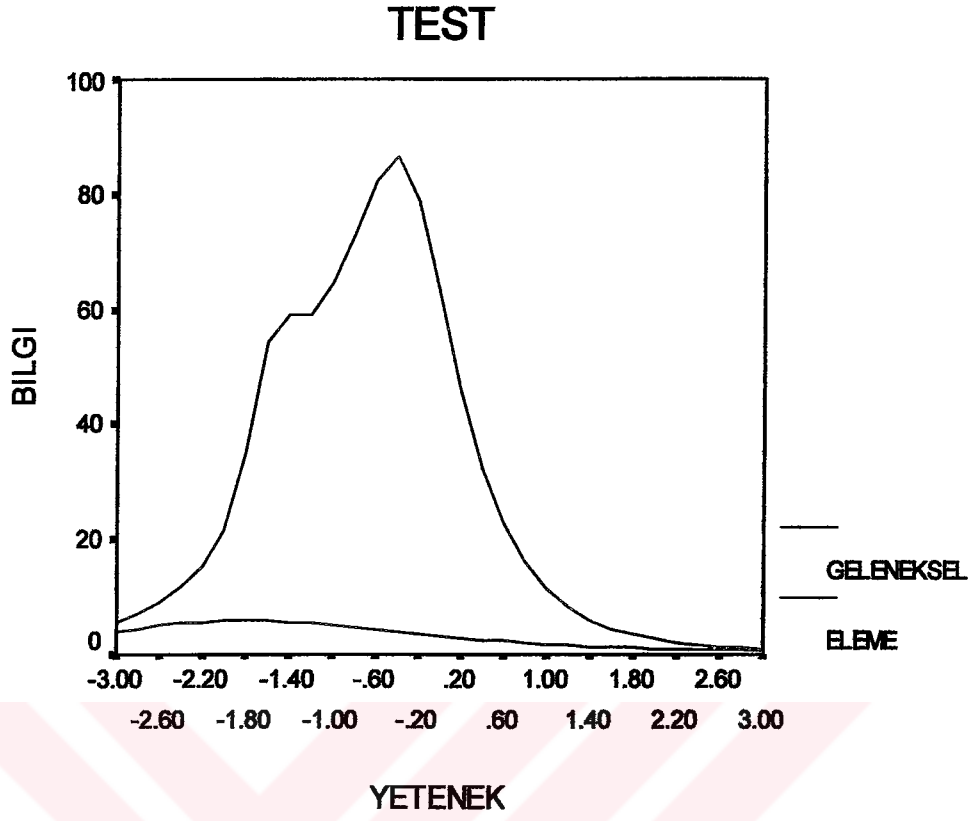
## Eleme Puanlaması'nın Geleneksel Puanlamaya göre Göreceli Etkiliği

Maddeler	Yetenek Düzeyleri						
	-3	-2	-1	0	1	2	3
1	9.57	11.22	15.14	15.14	5.49	1.42	0.54
2	19.20	36.23	72.09	179.44	66.86	9.50	1.75
3	2.12	2.56	6.78	14.97	7.37	2.32	1.03
4	1.35	4.36	12.73	21.59	9.72	3.46	1.80
5	0.19	1.98	12.54	16.98	3.13	0.68	0.43
6	1.40	2.14	5.03	9.34	4.52	2.69	2.82
7	2.09	1.98	2.95	5.59	3.58	2.40	3.50
8	1.09	2.39	8.44	4.30	1.15	0.80	1.03
9	1.30	20.27	77.30	183.21	21.75	3.83	0.86
10	0.31	0.28	4.32	13.76	3.10	1.27	0.75
11	0.11	0.85	15.88	35.32	11.42	4.75	3.00
12	0.22	0.87	8.39	16.99	6.80	2.39	1.88
13	2.13	2.29	6.85	20.29	5.71	1.53	1.00
14	2.47	8.07	12.81	26.76	6.27	2.18	0.83
15	6.58	4.31	20.40	67.13	11.10	1.64	0.27
16	0.35	0.28	4.57	47.15	25.25	31.00	Tanımsız
17	1.05	6.96	38.33	192.56	18.81	3.90	2.00
18	0.01	1.24	102.94	23.81	6.37	2.25	1.50
19	0.03	4.86	11.00	15.95	5.94	3.53	4.25
20	0.01	11.01	55.43	13.22	4.71	9.50	23.00
21	0.97	10.76	31.10	35.80	15.65	5.14	2.82
22	3.28	3.19	4.12	4.73	2.91	1.57	1.18
23	13.91	11.95	9.98	10.45	8.58	4.17	2.05
24	1.41	9.66	30.79	21.94	13.95	7.04	3.35
25	11.76	18.39	9.84	13.99	6.63	2.17	0.92
26	1.33	2.07	5.34	8.45	4.46	2.33	1.88
27	3.02	4.59	7.43	10.77	6.04	2.53	1.69
28	15.67	31.63	61.62	49.69	17.20	4.65	1.77
29	5.79	17.00	52.08	147.58	66.80	10.56	2.63
30	3.84	10.13	28.99	12.18	3.84	1.68	1.00
Test	1.42	3.66	12.65	19.44	6.69	2.72	1.59

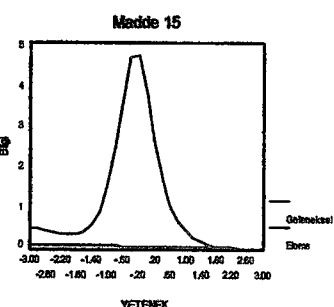
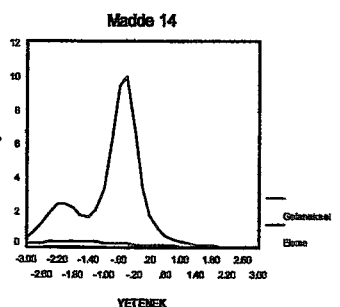
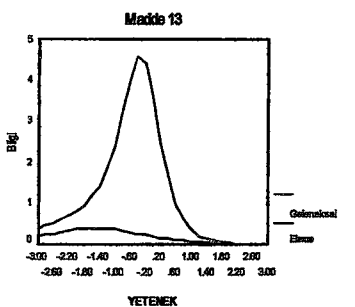
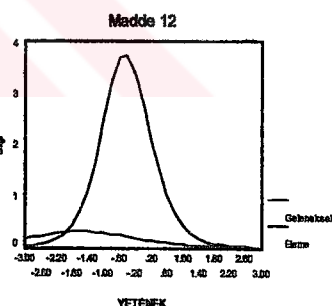
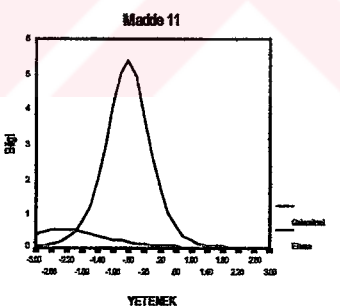
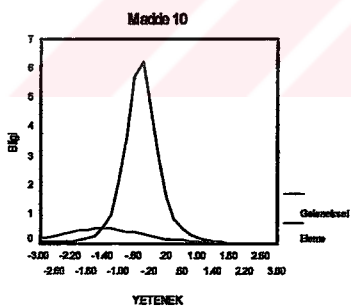
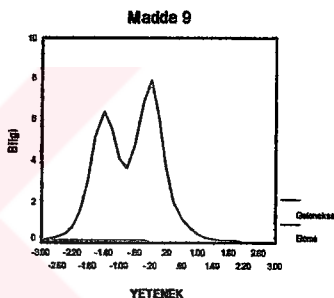
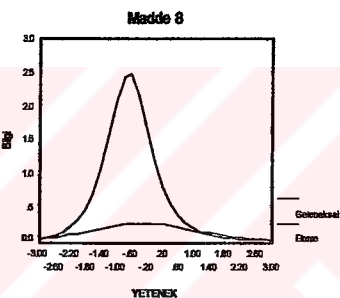
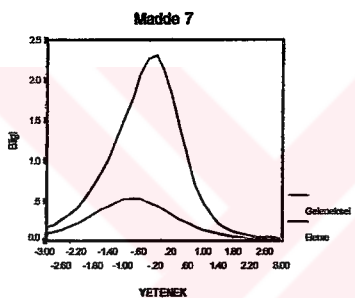
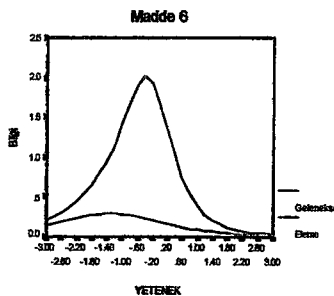
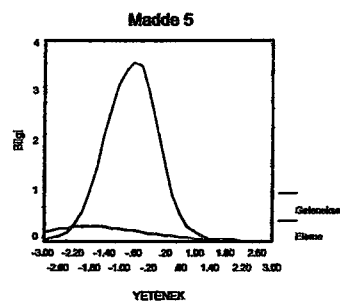
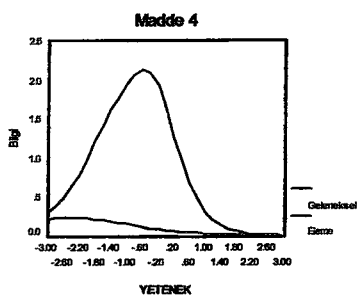
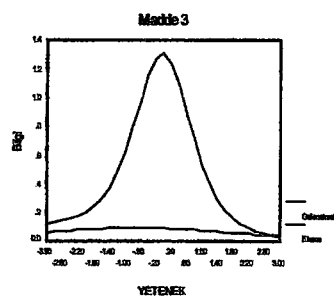
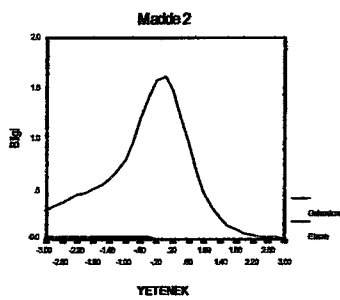
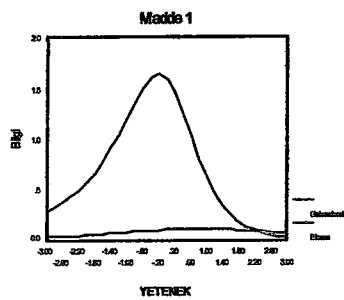
Tablo 8 incelendiğinde -3 yetenek düzeyinde 9 madde için göreceli etkililik değerleri 1'den küçük bulunmuştur. Bu da -3 yetenek düzeyinde 9 maddede geleneksel puanlamanın daha etkili olduğu anlamına gelmektedir. 21 maddede ise eleme puanlaması daha etkili ölçümler yapmaktadır. Bu etkililik 19 kata (madde 2) kadar çıkabilmektedir. Testin bütünü için ise -3 yetenek düzeyinde eleme puanlaması geleneksel puanlamanın 1.42 katı etkili görünmektedir. Yetenek düzeyi -2'de ise 4 madde için geleneksel puanlama daha etkili görünmektedir. Kalan 26 maddede eleme

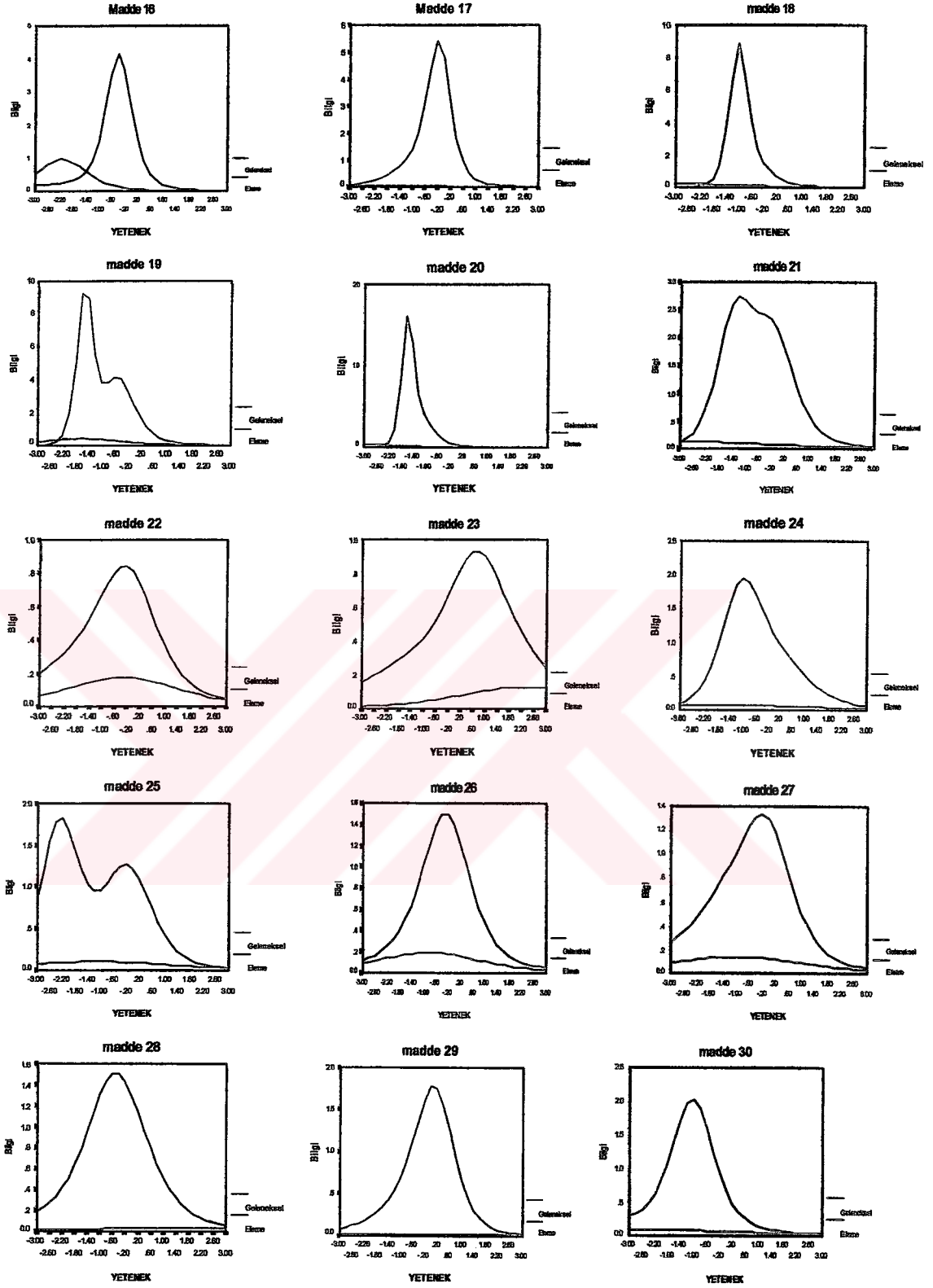
puanlaması daha etkili olarak belirlenmiştir. Testin bütününde ise, -2 yetenek düzeyinde, eleme puanlaması geleneksel puanlamanın 3.66 katı etkili görünmektedir. Yetenek düzeyi -1 için, göreceli etkililik değerlerinin tamamı eleme puanlamasının lehine görünmektedir. Göreceli etkililik yaklaşık 103 kata (madde 18) kadar çıkabilmektedir. Test düzeyinde ise -1 yetenek düzeyinde eleme puanlaması geleneksel puanlamaya göre, 12.65 kat daha etkili olarak görünmektedir. Yetenek düzeyi 0 (sıfır) için ise, yine tüm maddelerde eleme puanlamasının daha etkili olduğunu gösteren sonuçlar bulunmuştur. Etkililik 192.56 katına kadar (madde 17) çıkmıştır. Testin bütününe ait olarak ise eleme puanlaması 19 kat daha etkili olarak belirlenmiştir. Yetenek düzeyi +1 için ise, bütün maddelerde göreceli etkililik değerleri eleme puanlaması lehinedir. Testin bütünü için ise eleme puanlaması, geleneksel puanlamaya göre, 6.69 kat daha etkilidir. Yetenek düzeyi +2'de eleme puanlaması geleneksel puanlamaya göre 2 madde (madde 5 ve 8) dışında tüm maddelerde daha etkili görünmektedir. Test için ise, eleme puanlaması yaklaşık olarak 2.7 kat daha etkili görünmektedir. Yetenek düzeyi +3'de 7 madde geleneksel puanlamada, 11 madde eleme puanlamasında daha etkili, 2 madde ise eşit etkililikte görünmektedir. Testin bütünü için ise eleme puanlaması, geleneksel puanlamaya göre, yaklaşık olarak 1.6 kat daha etkili görünmektedir.

Geleneksel ve eleme puanlamalarından elde edilen bilgi düzeylerinin daha iyi ve karşılaştırmalı olarak görülebilmesi için Şekil 8 'de test bilgi grafikleri, şekil 9'da madde bilgi grafikleri verilmiştir.



**Şekil 8: Geleneksel ve Eleme Uygulamalarından Kestirilen Test Bilgi Düzeyleri Grafiği**





**Şekil 9: Geleneksel ve Eleme Uygulamalarına İlişkin Madde Bilgi Grafikleri**

Şekil 8 ve şekil 9’da açıkça görülebileceği gibi araştırmada kullanılan sözel yetenek testinin geleneksel ve eleme puanlamalarından Madde Tepki Kuramı’na göre kestirilen madde bilgi düzeyleri karşılaştırıldığında, açıkça eleme puanlamasında daha yüksek bilgi düzeyleri elde edildiği görülmektedir. Test bilgi düzeyi grafiğine bakıldığında tüm yetenek seviyelerinde eleme puanlamasının üstünlüğü göze çarpmaktadır. Özellikle yetenek düzeyinin 0 (sıfır) ile +1 düzeyinin arasında olduğu noktalarda eleme puanlaması geleneksel puanlamayla kıyaslanamayacak ölçüde fazla bilgi vermektedir.

Çoktan seçmeli testlerin bir sınırlılığı olarak kabul edilen kısmi bilgiye duyarsızlığın göstergesi olan düşük bilgi düzeyleri geleneksel puanlamadan kestirilen düzeylerden açıkça görülmektedir. Oysa, eleme puanlamasından kestirilen bilgi düzeylerinin çok daha yüksek olması, eleme puanlamasında bu sınırlılığın biraz engellenebileceğini gösterdiği düşünülebilir.

Eleme puanlamasının geleneksel puanlamaya göre madde ve test ölçeğinde çok daha etkili sonuçlar vermesi, eleme puanlamasının kısmi bilgiyi ölçme yoluyla daha duyarlı ölçümler yapabilmesinden kaynaklanabilir. Eleme puanlaması sadece test ölçeğinde değil, madde ölçeğinde de duyarlı ölçme yaparak hatayı düşürebilmektedir. Bu özellik eleme puanlamasından sağlanabilecek faydanın bir kısmıdır.

#### **b) Marjinal Güvenirlik Karşılaştırmaları**

Araştırmanın ikinci alt probleminin b maddesinde, araştırma için hazırlanan sözel yetenek testinin geleneksel puanlama ve eleme puanlaması uygulamalarından MTK’ya dayalı marjinal güvenirlik katsayıları kestirilmeye çalışılmıştır.

Kullanılan 30 çoktan seçmeli maddeden oluşan sözel yetenek testinin geleneksel uygulamasından elde edilen verilerin 2 parametrelili Madde Tepki Kuramı’na göre, madde bilgi düzeylerine dayalı olarak kestirilen marjinal güvenirlik katsayısı 0.738 olarak bulunmuştur. Kullanılan 30 maddelik sözel yetenek testinin eleme uygulamasından elde edilen verilerin genelleştirilmiş kısmi bilgi modeline göre, madde bilgi düzeylerine dayalı olarak kestirilen marjinal güvenirlik katsayısı 0,951 olarak bulunmuştur.

Kullanılan çoktan seçmeli maddelerden oluşan sözel yetenek testinin geleneksel ve eleme uygulamalarından elde edilen güvenilirlik katsayılarının karşılaştırılmasında Fischer'in z istatistiği kullanılmıştır. Geleneksel ve eleme uygulamalarından kestirilen güvenilirlik katsayılarının karşılaştırılmasına ilişkin z testi tablo 9'da verilmiştir.

**Tablo 9**

**Sözel Yetenek Testinin Madde Tepki Kuramına Göre Geleneksel ve Eleme Uygulamalarından Kestirilen Marjinal Güvenirlik Katsayılarının Karşılaştırılması**

	Geleneksel	Eleme	z	p
<b>Marjinal Güvenirlik</b>	0.738	0.951	-11.257**	0.000
<b>Fischer z Karşılığı</b>	0.946	1.843		

\*\* p<0.01

Tablo 9'da görülebileceği gibi geleneksel ve eleme uygulamasının karşılaştırılması sonucunda -11.257 z değeri elde edilmiştir. Bu z değerinin iki yönlü p değeri karşılığı 0.000 olarak belirlenmiştir. Kullanılan çoktan seçmeli sözel yetenek testinin, eleme puanlamasından kestirilen marjinal güvenilirlik katsayısı geleneksel puanlamadan kestirilen marjinal güvenilirlik katsayısından 0.01 anlamlılık düzeyinde daha yüksek olduğu söylenebilir.

Eleme puanlamasının daha yüksek güvenilirlik vermesi, eleme puanlamasından beklenen faydalar olan "şans başarısını engelleme" ve "kısmi bilgiyi ölçme" yoluyla duyarlı ölçüm yapabileceği iddiasını desteklemektedir. Şans başarısının engellenmesi yoluyla hata düşecek, kısmi bilginin ölçülmesi yoluyla da ölçmenin duyarlılığı artacaktır. Hatanın düşmesi ve duyarlılığın artması güvenilirlikle doğrudan ilişkili olduğu için güvenirlüğün yükselmesi beklenecektir.

**c) Geleneksel ve Eleme Puanlaması Uygulamalarından Madde Tepki Kuramına Dayalı Olarak Kestirilen Yetenekler Arasındaki İlişki Düzeyinin Belirlenmesi**

Araştırmanın ikinci alt probleminin c maddesinde, araştırmanın örneklemini oluşturan bireylerin kullanılan sözel yetenek testinin geleneksel ve eleme uygulamalarından MTK'na dayalı olarak kestirilen yetenek düzeyleri arasındaki ilişki belirlenmeye

çalışılmıştır. Örneklemi oluşturan bireylerin MTK'na dayalı olarak yeteneklerinin kestirilmesinde, MULTİLOG 7 programı kullanılmıştır.

Bireylerin çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamalarından kestirilen yetenek düzeylerine ilişkin betimleyici istatistikler tablo 10'da verilmiştir.

**Tablo 10**  
**Kullanılan Sözel Yetenek Testinin MTK'na Dayalı Olarak Geleneksel ve Eleme Puanlaması Uygulamalarından Kestirilen Yetenek Ölçülerine İlişkin Betimsel İstatistikler**

Betimsel İstatistikler	Geleneksel Puanlama	Eleme Puanlaması
En Düşük Yetenek	-0.868	-0.073
En Yüksek Yetenek	2.145	1.818
Aritmetik Ortalama	0.917	0.480
Ortanca	1.01	0.434
Standart Sapma	0.577	0.315
Kestirilen Yetenek Düzeyleri Arasındaki İlişki	0.564**	

\*\*  $p < 0.01$

Tablo 10'da verilen betimsel istatistikler incelendiğinde, aynı sözel yetenek testindeki geleneksel puanlamanın yetenek ranjı, eleme puanlamasının yetenek ranjından daha büyüktür. Geleneksel puanlamanın yetenek aritmetik ortalaması ve ortancası eleme puanlamasından daha yüksektir.

Örnekleme katılan bireylerin geleneksel puanlama ve eleme puanlamasındaki MTK'na göre kestirilen yetenek ölçüleri arasındaki ilişkinin belirlenmesi amacıyla, geleneksel ve eleme puanlamalarından kestirilen yetenek düzeyleri arasındaki korelasyon Pearson momentler çarpımı korelasyon katsayısı ile belirlenmiştir.

Bireylerin yetenek düzeyleri arasındaki ilişki 0.564 olarak hesaplanmıştır. Bu katsayı 0.01 anlamlılık düzeyinde sıfırdan farklıdır. Bu katsayı ilişki düzeyi olarak orta düzeydedir. Aynı bireylerin aynı maddelerden oluşan aynı testin iki farklı uygulamasından kestirilen yetenekler olarak düşünüldüğünde, düzeylerdeki değişimin göstergesi olarak kabul edilebilir. Bu korelasyon katsayısının karesi alındığında,  $r^2=0.32$

belirtme katsayısı bulunacaktır. Bu katsayı iki puanlama türünün birbirlerindeki değişimin % 32'sini açıklayabildiğini göstermektedir. Birbirlerini açıklayamadıkları oran ise % 68'dir. Bu iki puanlama türünün oldukça farklı puanlar ürettiklerini söyleyebiliriz.

### **ALT PROBLEM 3: ÇOKTAN SEÇMELİ TESTLERDE GELENEKSEL PUANLAMA VE ELEME PUANLAMASI UYGULAMASININ KARŞILAŞTIRILMASINA İLİŞKİN ÖĞRENCİ GÖRÜŞLERİ**

Araştırmanın üçüncü alt probleminde aynı sözel yetenek testinin geleneksel ve eleme uygulamalarını alan bireylerin uygulamaların karşılaştırılmasına ilişkin görüşleri anket yoluyla belirlenmeye çalışılmıştır. Testin iki uygulamasını alan bireylerden, geleneksel ve eleme uygulamalarını bazı yönlerden karşılaştırmaları istenmiştir. Bireylerin seçtikleri uygulama türüne göre frekans ve yüzdeleri belirlenerek yorumlanmıştır.

Bireylere araştırmanın tablo 11'de görülen anket soruları sorulmuş ve verilen cevapların frekans ve yüzde dağılımları aşağıda verilmiştir.

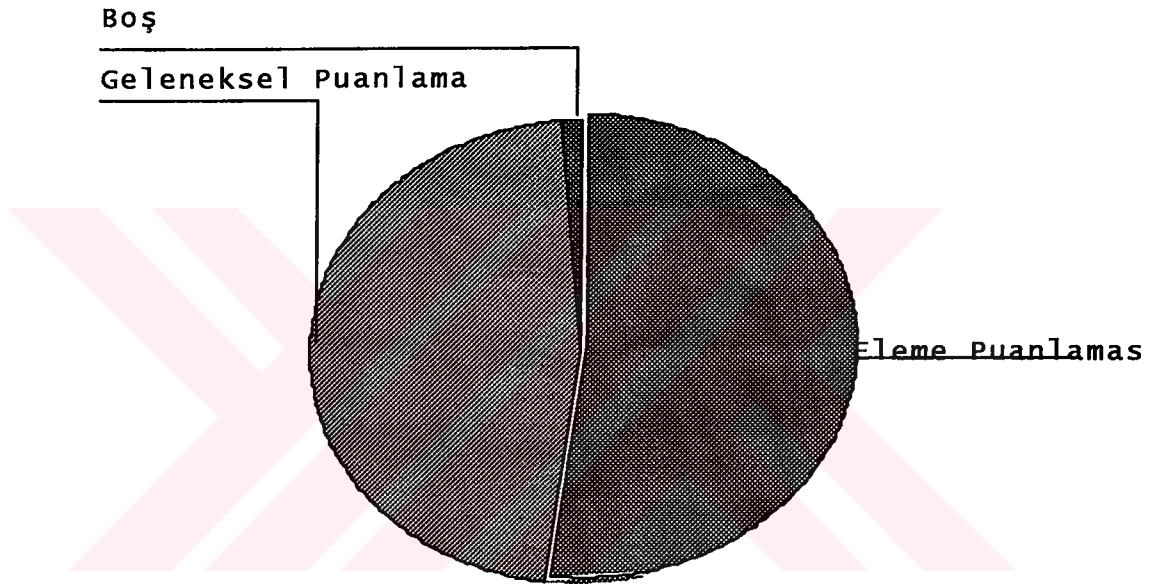
**Tablo 11**

#### **Test Alan Bireylerin Geleneksel ve Eleme Uygulamalarını Karşılaştırmalarına İlişkin Görüşlerinin Frekans ve Yüzde Dağılımları**

<i><b>İfadeler</b></i>	<b>Geleneksel</b>		<b>Eleme</b>	
	Frekans	%	Frekans	%
1. Cevaplaması daha kolay.	77	47.5	85	52.5
2. Daha fazla zaman alıyor.	64	39.3	99	60.7
3. Başarıyı belirlemede daha uygun olduğunu düşünüyorum.	85	53.1	75	46.9
4. Verdiğim cevapların doğruluğundan daha çok eminim.	69	43.9	88	56.1
5. Sınav aldığımızda hangisini tercih edersiniz?	91	55.8	72	44.2

Tablo 11'de görüldüğü gibi test alan bireylere sorulan "Aldığınız çoktan seçmeli sözel yetenek testinin geleneksel ve seçenek eleme yöntemlerinden hangisinin cevaplanmasının daha kolay olduğunu düşünüyorsunuz?" sorusuna, bireylerin % 47.5'i eleme uygulamasının , %52.5'i geleneksel uygulamanın cevaplanmasının daha kolay olduğunu belirtmiştir. Anket sorusunu bireylerin cevaplamasına ilişkin grafik şekil 10'da verilmiştir. Eleme uygulamasına bireylerin yabancı olması cevaplama işlemini

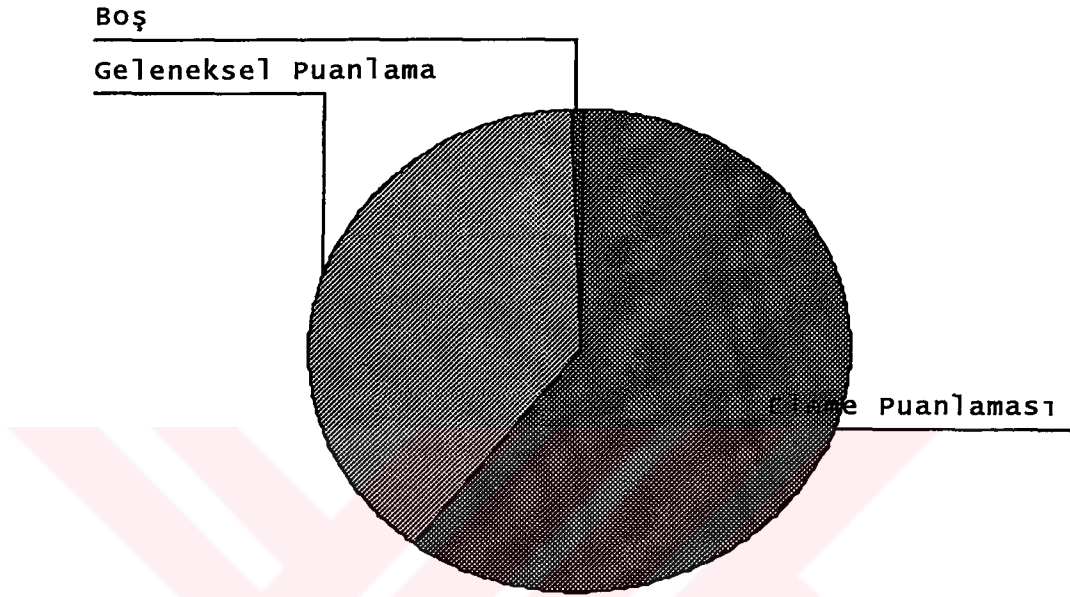
daha zor hale getirdiğinden, bireylere karmaşık bir işlem gibi gelebilmektedir. Bu sebeple bireylerden bazıları geleneksel uygulamada yaptığı gibi önce doğru cevabı belirleyip, daha sonra kalan seçenekleri işaretleme yoluna gidebilmektedir. Bu yolla cevaplama eleme uygulamasından beklenen faydanın elde edilmesini engellemektedir. Test cevaplama alışkanlıklarının da sadece bir yönerge ve açıklama ile değiştirilmesinin zor olması eleme uygulamasından beklenen faydanın tamamının ortaya çıkmasını engellediği düşünülmektedir.



**Şekil 10: Cevaplayıcıların Cevaplama Metotlarından Hangisinin Cevaplamasının Daha Kolay Olduğuna İlişkin Görüşlerinin Dağılımları Grafiği**

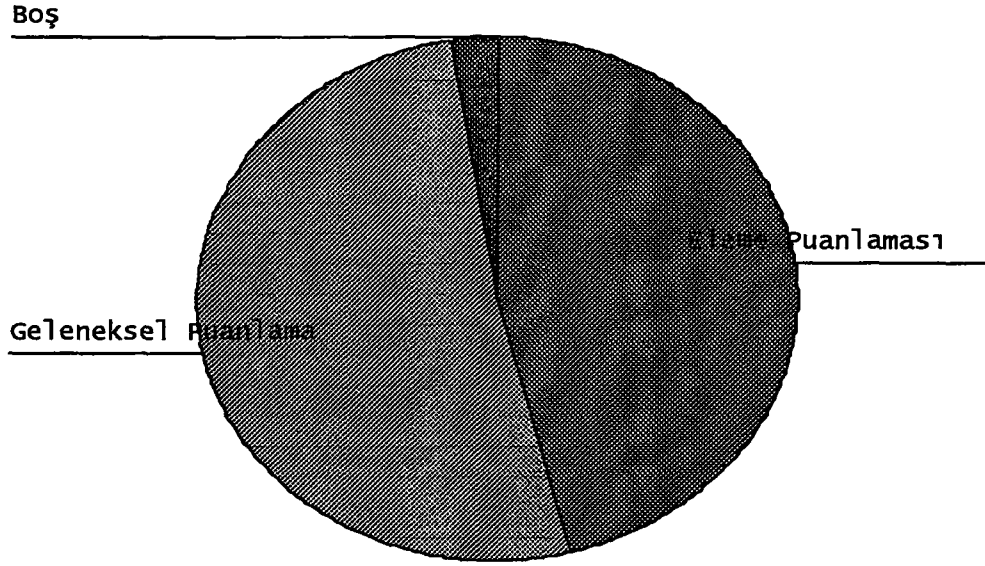
Cevaplayıcılara ikinci anket sorusunda çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamalarından hangisinin cevaplamasının daha fazla zaman aldığı sorulmuştur. Tablo 11 incelendiğinde cevaplayıcıların yaklaşık olarak % 61'i eleme uygulamasının daha fazla zaman aldığını, % 39'unun geleneksel uygulamanın daha fazla zaman aldığını belirtmiştir. Cevaplayıcıların 2. anket sorusuna verdikleri cevaplara ilişkin dağılımın grafiği şekil 11'de verilmiştir. Bu cevaplar beklenen şekildedir. Çünkü,

öğrencilerin alışık olmadığı bir uygulama türü olan seçenek eleme cevaplama öğrencilere karmaşık gelebilmektedir. Ayrıca işaretleme açısından tek seçenek işaretlemenin daha az zaman alması olağandır.



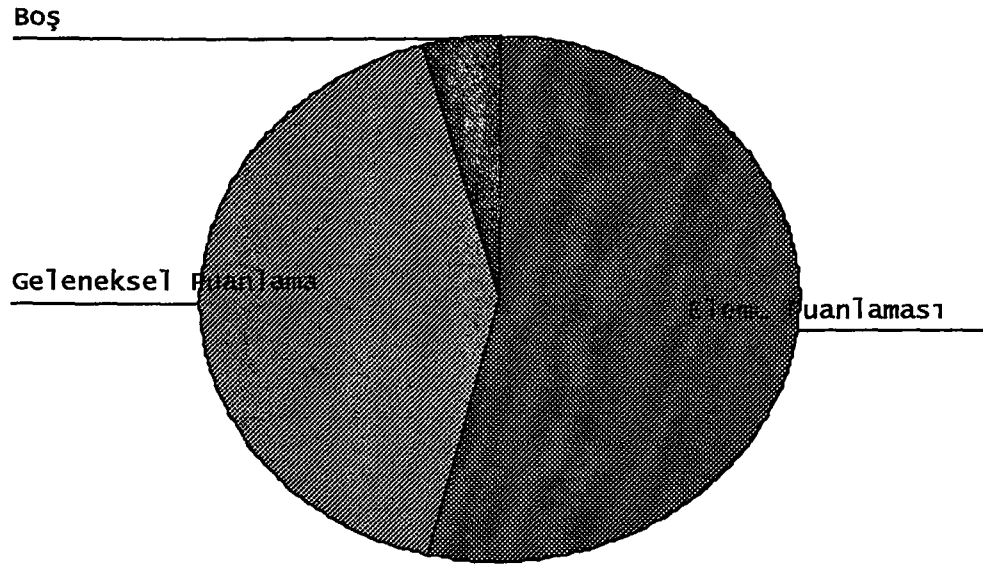
**Şekil 11: Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisinin Cevaplamasının Daha Fazla Zaman Aldığına İlişkin Öğrenci Görüşleri Dağılımı Grafiği**

Cevaplayıcılara üçüncü anket sorusunda çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamalarından hangisinin başarıyı belirlemede daha uygun olduğunu düşündükleri sorulmuştur. Tablo 11 incelendiğinde yaklaşık olarak cevaplayıcıların %53'ü geleneksel uygulamanın, % 47'si eleme uygulamasının başarıyı belirlemek için daha uygun olduğu görüşünde oldukları belirlenmiştir. öğrencilerin bu soruya ilişkin cevap dağılımlarını gösteren grafik şekil 12'de verilmiştir. Bu oranlar birbirine yakın olmakla birlikte, geleneksel uygulamanın daha uygun olduğu görüşünü daha fazla benimsedikleri görülmüştür. Geleneksel uygulamanın bireylere daha uygun gelmesi hem bireylerin bu cevaplama tarzına alışkın olmalarından, hem de şimdiye kadar aldıkları bütün önemli sınavların geleneksel uygulama şeklinde cevaplanmasından kaynaklanıyor olabilir.



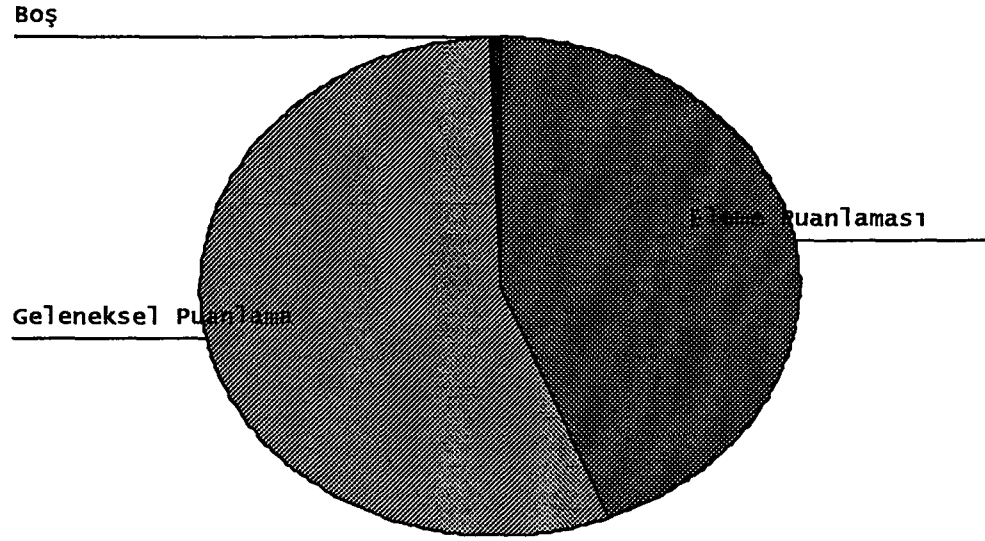
**Şekil 12: Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisinin Başarıyı Belirlemede Daha Uygun Olduğuna İlişkin Öğrenci Görüşleri Dağılımı Grafiği**

Cevaplayıcılara dördüncü anket sorusunda çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamalarından hangisinde verdikleri cevapların doğruluğundan daha fazla emin oldukları sorulmuştur. Tablo 11’de görüldüğü gibi cevaplayıcıların % 44’ü geleneksel, % 56’sı eleme uygulamasında verdikleri cevapların doğruluğundan daha çok emin olduklarını ifade etmişlerdir. Bu cevaplara ilişkin dağılımın grafiği şekil 13’te verilmiştir. Eleme uygulamasında, belirleyebildikleri kadar çeldiriciyi eleme olanağı olduğu için tahminle cevaplama gibi bir alternatifte ihtiyaç yoktur. Bu sebeple, bu yöntemi etkili olarak kullanan cevaplayıcıların verdikleri cevapların doğruluğundan emin olması gerekir. Bu sebeple, eleme uygulamasında daha büyük bir oranda verilen cevapların doğruluğundan daha emin olduğu ifade edilmiştir.



**Şekil 13: Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisinde Verilen Cevapların Doğruluğundan Emin Olunduğuna İlişkin Cevaplayıcı Görüşleri Dağılımı Grafiği**

Cevaplayıcılara beşinci ve son anket sorusunda çoktan seçmeli testin geleneksel ve eleme uygulamalarından hangisini önemli bir sınava girdiklerinde tercih edecekleri sorulmuştur. Tablo 11’de görüldüğü gibi cevaplayıcıların % 56’sı geleneksel, % 44’ü eleme uygulamasını tercih edeceklerini ifade etmişlerdir. Bu cevaplara ilişkin dağılımın grafiği şekil 14’te verilmiştir. Cevaplayıcıların daha çok geleneksel uygulamayı tercih etmesinin olası en önemli kaynağı bu sınav türüne alışık olmalarından kaynaklanıyor olabilir. Alışık olmadıkları eleme uygulamasını aldıklarında, işaretleme konusunda kafalarının karışacağını düşünmüş olmaları olasıdır. Geleneksel uygulamanın işaretlemesinin kolay olması bir diğer etken olarak görülebilir. Sadece doğru cevabı bulup işaretlemek tek tek yanlış seçenekleri bulup işaretlemekten daha basit bir işlem gerektirdiği için geleneksel uygulamayı tercih etmeleri olasıdır.



**Şekil 14: Çoktan Seçmeli Sözel Yetenek Testinin Geleneksel ve Eleme Uygulamalarından Hangisini Test Alırken Tercih Edeceklerine İlişkin Cevaplayıcı Görüşleri Dağılımı Grafiği**

## BÖLÜM IV

### SONUÇ VE ÖNERİLER

Bu bölümde araştırmanın bulgularından çıkarılan sonuçlara ve bu sonuçlara dayalı yapılacak önerilere yer verilmiştir. Ayrıca bu araştırmanın kapsamı dahilinde olmayan ancak, daha sonra yapılabilecek olan araştırma önerilerine yer verilmiştir.

#### SONUÇLAR

Bu araştırmada çoktan seçmeli bir sözel yetenek sınavının geleneksel ve eleme puanlaması uygulamaları yapılmış ve bu iki uygulamanın psikometrik özellikleri KTK ve MTK bağlamında karşılaştırılmıştır. Bu karşılaştırma sonucu elde edilen bulgulara dayalı olarak aşağıdaki sonuçlara ulaşılmıştır.

1. Çoktan seçmeli 30 maddelik bir sözel yetenek testinin geleneksel ve eleme puanlamalarından elde edilen madde ayıricılık güçleri karşılaştırılmış ve 15 maddenin ayıricılık gücünün 0.01 anlamlılık düzeyinde farklılaştıkları görülmüştür. Anlamlı fark bulunan 15 maddenin 11'inde geleneksel uygulamanın ayıricılık gücü daha yüksek, 4'ünün eleme uygulamasında daha yüksek olarak kestirildiği görülmüştür. Eleme uygulamasının önerilmesindeki etkenlerden biride geleneksel uygulamaya göre kısmi bilgiye daha duyarlı olabileceği ve bireyler arasındaki ayırımı daha iyi yapabileceği öngörüsüdür. Ancak, elde edilen bulgulara bakıldığında eleme puanlamasının KTK'na dayalı daha yüksek ayıricılık gücü vereceği öngörüsünü gerçekleştirmede yetersiz kaldığı görülmektedir. Eleme uygulamasının geleneksel puanlamadan daha düşük ayıricılık gücü verdiği sonucuna varılmıştır.

2. Klasik Test Kuramı'na dayalı olarak çoktan seçmeli testlerin geleneksel ve eleme puanlamalarının birbirine göre göreceliği etkililiğini belirlemek için 0.70 güvenirlik düzeyine ulaştıkları madde sayıları oranı belirlenmiştir. Çoktan seçmeli sözel yetenek testinin geleneksel puanlamada en yüksek ayıricılık veren 15 madde ile eleme puanlamasında 8 madde ile 0.70 güvenirlik düzeyine ulaştıkları belirlenmiştir. Elde edilen madde sayıları oranı 1.88 olarak belirlenmiştir. Bu bulguya dayalı olarak eleme

puanlamasının daha az sayıda soruyla geleneksel puanlama ile aynı güvenilirliğe ulaştığı sonucuna varılmıştır.

3. Çoktan seçmeli testlerin geleneksel ve eleme puanlamalarının güvenilirliklerinin karşılaştırılmasında, geleneksel puanlamada KR-20, eleme puanlamasında Cronbach  $\alpha$  güvenilirlikleri kestirilmiştir. Geleneksel puanlama için 0.737 ve eleme puanlaması için 0.829 güvenilirlik katsayıları elde edilmiştir. Elde edilen bu katsayıların karşılaştırılmasında Fischer'in z istatistiği kullanılmış ve aralarında 0.01 düzeyinde anlamlı fark bulunmuştur. Bu bulguya dayalı olarak eleme puanlamasının geleneksel puanlamaya göre daha yüksek güvenilirlik verdiği sonucuna varılmıştır.

4. Çoktan seçmeli sözel yetenek testinin geleneksel ve eleme puanlamalarından elde edilen toplam puanlar arasındaki ilişkinin incelenmesinde Pearson momentler çarpımı korelasyon katsayısı kullanılmıştır. Çoktan seçmeli sözel yetenek testinin geleneksel ve eleme puanlamalarında elde edilen toplam puanlar arasındaki korelasyon 0.565 olarak belirlenmiştir. Kestirilen bu ilişki katsayısı 0.01 düzeyinde anlamlıdır. Çoktan seçmeli sözel yetenek testinin aynı bireylerden elde edilen geleneksel ve eleme puanlamaları arasında orta düzeyde bir ilişki bulunmuştur. Elde edilen korelasyon katsayısı, puanlamaların birbirindeki değişkenliğin % 32'sini açıklayabildiğini göstermektedir. Bu bulgudan, puanlamaların birbiriyle ilişkili fakat, bir o kadarda farklı yetenek puanları ürettiği sonucuna varılmıştır.

5. MTK'ya dayalı olarak maddelerin ve testin geleneksel ve eleme puanlamalarının aynı maddelerden ve testten 7 yetenek düzeyinde kestirilen bilgi düzeyleri birbirlerine oranlanması yoluyla göreceli etkililikleri belirlenmiştir. Bu oranlama sonucunda bazı maddelerde uç yetenek düzeylerinde geleneksel uygulama daha etkili gözükmüş, fakat yetenek düzeylerinin bütünü göz önüne alındığında eleme puanlamasının çok daha etkili olduğu belirlenmiştir. Testin bütününe ait göreceli etkililik incelendiğinde bu sonuç açıkça görülmektedir. Testin bütünü için, eleme uygulamasının geleneksel uygulamaya göre -3, -2, -1, 0, 1, 2, 3 yetenek düzeylerinde göreceli etkililik değerleri sırasıyla 1.42, 3.66, 12.65, 19.44, 6.69, 2.72 ve 1.59 kat olarak belirlenmiştir. Belirlenen bu değerlerden, çoktan seçmeli sözel yetenek testinin eleme puanlamasının geleneksel puanlamaya göre daha etkili ölçümler yapabildiği ve her yetenek düzeyinde bireyler hakkında daha fazla bilgi verdiği sonucu çıkarılmıştır.

6. Çoktan seçmeli sözel yetenek testinin geleneksel ve eleme puanlamalarının MTK'ya dayalı olarak kestirilen marjinal güvenilirlik katsayıları karşılaştırılmıştır. Kestirilen marjinal güvenilirlik katsayıları geleneksel uygulama için 2 parametrelili lojistik model, eleme puanlaması için genelleştirilmiş kısmi puan modeli kullanılarak yapılmıştır. Yapılan analizler sonucunda marjinal güvenilirlik katsayısı geleneksel puanlama için 0.738, eleme puanlaması için 0.951 katsayıları elde edilmiştir. Bu iki katsayının farklı olup olmadığı Fischer'in z istatistiği kullanılarak test edilmiş ve 0.01 düzeyinde anlamlı olarak farklı oldukları belirlenmiştir. Bu bulgu sonucunda, eleme puanlamasının geleneksel puanlamadan daha yüksek güvenilirlik verdiği sonucu çıkarılmıştır.

7. Çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamalarına katılan bireylerin MTK'ya dayalı olarak yetenek düzeyleri kestirilmiştir. Kestirilen bu yetenek düzeyleri arasındaki ilişkinin incelenmesinde Pearson korelasyon katsayısı kullanılmıştır. Araştırmanın örneklemini oluşturan 315 bireyin geleneksel ve eleme uygulamalarından kestirilen yetenek düzeyleri arasındaki ilişki katsayısı 0,564 olarak belirlenmiştir. Kestirilen bu katsayı 0.01 düzeyinde anlamlı olarak sıfırdan farklıdır. Çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamasına katılan bireylerin iki puanlamadan kestirilen yetenek düzeyleri orta düzeyde ilişkili bulunmuştur. Kestirilen bu ilişki katsayısı puanlama türlerinin birbirlerindeki değişkenliğin % 32'sini açıklayabildiğini göstermektedir. Bu bulgudan, çoktan seçmeli sözel yetenek testinin geleneksel ve eleme puanlamasının MTK'ya dayalı olarak kestirilen yetenek düzeylerinin ilişkili olduğu, fakat bir o kadar da farklı oldukları sonucuna varılmıştır.

8. Çoktan seçmeli sözel yetenek testinin geleneksel ve eleme uygulamasını alan bireylere, iki uygulamanın karşılaştırılmasına ilişkin 5 maddeden oluşan bir anket cevaplatılmıştır. Ankette yer alan "Hangisinin cevaplama daha kolay?" sorusunda bireylerin % 47.5'i geleneksel, %52.5'i eleme uygulamasının cevaplama daha kolay olduğu görüşünü belirtmişlerdir. Bu bulgudan, eleme uygulamasının bireylerin iki veya daha fazla seçeneğin doğruluğundan şüphelendiklerinde karar vermelerini kolaylaştırabildiği sonucu çıkarılmıştır. Ankette yer alan "Hangisinin cevaplama daha fazla zaman alıyor?" sorusunda, bireylerin % 39'u geleneksel, % 61'i eleme

uygulamasının cevaplamasının daha fazla zaman aldığı görüşünü belirtmişlerdir. Bu bulgudan, eleme uygulamasının ilk defa alan bireylere daha karmaşık geldiği ve daha fazla işaretleme yükü getirdiğinden daha fazla zaman aldığı sonucu çıkarılmıştır. Ankette yer alan “Hangisi başarıyı belirlemede daha uygun?” sorusunda bireylerin % 53’ü geleneksel, % 47’si eleme uygulamasının başarıyı belirlemede daha uygun olduğu görüşünü belirtmişlerdir. Bu bulgudan, bireylerin daha çok geleneksel uygulamanın başarıyı belirlemede daha iyi bir yol olduğunu düşündükleri sonucuna varılmıştır. Bireylerin geleneksel uygulamayı eleme uygulamasından daha uygun görmelerinin olası kaynağı alışkanlıklardan kaynaklanabileceği düşünülmüş ve bu konunun araştırılması gerektiği sonucuna varılmıştır. Ankette yer alan “Hangisinde verdiğiniz cevapların doğruluğundan daha fazla eminsiniz?” sorusunda, bireylerin % 44’ü geleneksel, % 56’sı eleme uygulamasında verdikleri cevapların doğruluğundan daha fazla emin olduklarını belirtmişlerdir. Bu bulgudan, eleme uygulamasında bireylerin daha fazla emin olarak testi cevaplayabildikleri ve şans başarısını daha az kullandıkları sonucu çıkarılmıştır. Ankette yer alan “Bir test aldığınızda hangisini tercih edersiniz?” sorusunda, bireylerin % 56’sı geleneksel, % 44’ü eleme uygulamasını tercih edeceklerini belirtmişlerdir. Bu bulgudan, bireylerin geleneksel uygulamayı daha çok tercih ettikleri sonucuna varılmıştır. Ayrıca tercihin sebebinin de alışkanlıklardan kaynaklanıp kaynaklanmadığının araştırılması gerektiği sonucuna varılmıştır.

Sonuç olarak eleme puanlaması uygulaması ile hem KTK hem de MTK’ya dayalı olarak yapılan analizlerinde geleneksel puanlamadan daha güvenilir sonuçlar elde edilmiştir. KTK’ya göre yapılan analizlere göre, eleme puanlamasında madde ayırıcılık güçleri daha düşük bulunmasına rağmen MTK’ya dayalı analizlere göre her yetenek düzeyinde eleme puanlaması geleneksel puanlamadan daha fazla bilgi verici bulunmuştur. Eleme puanlaması ile yapılan uygulamanın yetenek düzeylerinde daha fazla bilgi verici olması, MTK’ya dayalı yöntemlerin puanlamaya daha duyarlı olmasından kaynaklanabilir. Puanlamadaki kategori sayısının artmasının her yetenek düzeyindeki bireyler hakkında daha fazla bilgi sağladığı görülmektedir. Eleme puanlamasına ilişkin daha önce yapılan çalışmalarda madde düzeyinde karşılaştırmalara rastlanmamıştır. Sadece Coombs, Milholland ve Womer (1956)’in yaptıkları çalışmada KTK’da madde parametrelerinin geleneksel ve eleme puanlamasında değişmediği belirtilmiştir.

Test güvenilirliđi, hem KTK hem de MTK'ya dayalı analizlerde eleme puanlaması lehine sonuçlar vermiştir. Bu durum, eleme puanlamasına daha az hatanın karıştığını ve bu yöntemle daha tutarlı ölçmeler yapılabileceğine işaret etmektedir. Coombs, Milholland ve Womer (1956), Collet (1971), Hakstian ve Kansup (1975), Jaradat ve Tollefson (1988) ve Ben-Simon, Budescu ve Nevo (1997) yaptıkları çalışmalarda eleme puanlamasının güvenilirliđi geleneksel puanlamadan çok küçük miktarda daha yüksek bulunmuştur, fakat bu fark istatistiksel olarak anlamlı değildir. Bradbard, Parker ve Stone (2004)'un yaptıkları çalışmada ise eleme puanlamasının güvenilirliđi çok az geleneksel puanlamadan daha düşük bulunmuştur, fakat bu fark da anlamlı değildir. Bu çalışmada yapılan tüm çalışmalardan farklı bir sonuca ulaşılmıştır.

Aynı testten iki yöntemle elde edilen puanlar ya da yetenek ölçüleri arasında orta düzeyde bir ilişkinin görülmesi, puanlama yöntemleriyle elde edilen sonuçlar arasında tam bir uyumun olmadığını göstermektedir. Puanlama yöntemleri öğrenci puanlarındaki değişimin yaklaşık olarak % 35'lik bir kısmını birlikte açıklayabilmektedir. Bu da iki puanlama yönteminin bireyler hakkında birbirinden oldukça farklı bilgiler verdiğini göstermektedir.

Eleme puanlaması öğrenciler için tamamen yeni bir uygulamadır. Öğrencilerin ankete verdikleri cevaplardan geleneksel puanlamayı daha çok tercih ettikleri görülmekle birlikte, eleme puanlamasının özelliklerini geleneksel puanlamadan daha üstün bulanlar da grubun azımsanamayacak bir kısmıdır. Öğrencilerin, geleneksel puanlamaya alışık oldukları için bu yöntemi tercih ettikleri ancak eleme puanlaması ile yapılan uygulamaların çoğalması ve bu uygulama hakkında daha fazla bilgilendirilmelerinin öğrenci görüşlerinde değişikliğe yol açabileceđi düşünülmektedir. Coombs, Milholland ve Womer (1956) yaptıkları çalışmada öğrencilerin daha çok eleme puanlamasını tercih ettiklerini belirlemişlerdir. Bu çalışmada Coombs, Milholland ve Womer (1956)'in bulgusunun aksine bir sonuç bulunmuştur.

## **ÖNERİLER**

1. Yapılan bu araştırmadan çıkarılan sonuçlardan, çoktan seçmeli testlerde eleme puanlaması geleneksel puanlamaya göre KTK'da güvenilirlik yönünden, MTK'da madde ve test bilgi düzeyleri, marjinal güvenilirlik yönünden daha iyi sonuçlar vermiştir. Bu

bağlamda, büyük ölçekli test uygulaması yapan bir kurumun desteklediği çok daha büyük örnekleme ve gerçeğe daha yakın sınav koşullarında deneme uygulaması yapılması ve bu uygulamanın sonuçlarına göre, büyük ölçekli sınavlarda eleme puanlamasının kullanılması önerilebilir.

2. Eleme puanlaması ile hem daha güvenilir ölçme sonuçları elde edebileceği hem de daha az soru kullanarak geleneksel puanlamaya yakın güvenilirliklerde ölçümler yapılabileceğinden ÖSS ve Milli Eğitim Bakanlığı gibi büyük ölçekli test uygulayan kurumlara eleme puanlaması konusunda araştırmalar yapmaları ve kullanmaları önerilebilir.

3. Sınava katılan bazı bireylerin eleme puanlamasını iyi anlayamamalarından, bu bireylerin önce bir doğru cevap tespit edip sonra kalan seçenekleri işaretledikleri cevap kağıtlarından ve sınav sonunda birkaç kişinin ifadelerinden anlaşılmıştır. Bu sınava katılan bireylerin eleme puanlamasına alışık olmamalarından kaynaklanmıştır. Bu sebeple, sınava katılacak bireyler eleme puanlaması yöntemi ve bunun üstünlükleri konusunda ayrıntılı olarak bilgilendirilebilir.

4. Yapılan bu çalışmada çoktan seçmeli testlerin geleneksel puanlaması ile eleme puanlaması karşılaştırılmıştır. Çoktan seçmeli testlerde kısmi bilgiye duyarlı ve şans başarısını engelleme iddiası taşıyan diğer alternatif metotlarında kullanıldığı daha geniş bir araştırma yapılabilir.

5. Bu çalışmada ölçüt alınabilecek bir veriye ulaşılamadığı için geçerlik çalışması kapsam dışında tutulmuştur. İleride geçerlik çalışmasını da içeren bir araştırma yapılması önerilebilir.

6. ÖSS veya OÖKSYS (Orta Öğretim Kurumları Seçme ve Yerleştirme Sınavı) gibi sınavlarda aynı bölümlere yerleştirilen öğrencilerin puan genişliğinin çok küçük olması geleneksel puanlamanın puanlamasından kaynaklanabileceğinden, eleme puanlamasının bu daralmaya çözüm olabileceği düşünülmüştür. Bu puan darlığı yapılacak geçerlik çalışmalarında gerçek sonuçlara ulaşılmasını engellemektedir. Eleme puanlaması kullanılması geçerlik çalışmalarında ortaya çıkan bu soruna çözüm olabilir.

## KAYNAKÇA

- AIKEN, L. R. (1988) *Psychological Testing and Assessment*. Boston: Allyn and Bacon.
- BAYKUL, Y. (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*. Ankara: ÖSYM yayınları.
- BEN – SIMON, A., Budescu D. V. and Nevo B. (1997) “A Comparative Study of Measures of Partial Knowledge in Multiple – Choice Tests.” *Applied Psychological Measurement*, Vol. 21, No. 1: 65-88.
- BOCK, R.D. (1972) “Estimating Item Parameters and Latent Ability When Response Are Scored in Two or More Nominal Categories.” *Psychometrika*, Vol. 37, No. 1: 29-51.
- BRADBARD, D. A., Parker D. and Stone G. L.(2004). “An Alternate Multiple-Choice Scoring Procedure in A Macroeconomics Course.” *Decision Sciences Journal of Innovative Education*, Vol. 2, No. 1: 11-26.
- COLLET, L. S. (1971) “Elimination Scoring: An Empirical Evaluation.” *Journal of Educational Measurement*, Vol. 8: 209-214.
- COOMBS, C.H., Milholland J.E. and Womer F. B. (1956). “The Assessment of Partial Knowledge.” *Educational and Psychological Measurement*, Vol. 16: 13-37.
- CROCKER, L. and Algina, J. (1986) *Introduction Classical And Modern Test Theory*, New York: CBS College Publishing.
- ECHTERNACHT, G. (1976). “Reliability and Validity of Item Option Weighting Schemes.” *Educational and Psychological Measurement*, Vol. 36, 301-309.
- EMBRETSON, S.E., and Reise, S.P. (2000). *Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum Associates.

- FRARY, R. (2000). "Testing Memo 1: Guessing on Multiple – Choice Tests." Test Scoring Services Pages, Blacksburg, 15 August 2000. <<http://www.testscoring.vt.edu/memo01.html>>.
- GIBBONS, J. D., Olkin I. and Sobel M. "A Subset Selection Tecnique for Scoring Items on A Multiple Choice Test." *Psychometrika*, Vol. 44, No. 3, 1979: 259-270.
- GULLIKSEN, H.O (1950). *Theory of Mental Tests*. New York: John Wiley and Sons, Inc.
- HAKSTIAN, A. R. and Kansup, W. (1975). "A Comparison of Several Methods of Assessing Partial Knowledge in Multiple Choice Tests: II. Testing Procedures." *Journal of Educational Measurement*, Vol.12, 219-230.
- HAMBLETON, R. K., and Swaminathan, H. (1985). *Item response theory: principles and Applications*, Boston: Kluwer-Nijhoff.
- JARADAT, D. and Tollefson N. (1988). "The Impact of Alternative Scoring Procedures for Multiple-Choice Items on Test Reliability, Validity and Grading. " *Educational and Psychological Measurement*, Vol. 48: 627-635.
- KURZ, T. Barber. (1999). "A Review of Scoring Algorithms for Multiple-Choice Tests." Paper Presented at The Annual Meeting of Southwest Educational Research Association, San Antonio.
- LORD, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, N. J.: Lawrence Erlbaum Associates.
- LORD, F.M., and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Reading, Mass.: Addison-Wesley.
- MAGNUSON, David. (1966). *Test Theory*. Stockholm: Addison-Wesley.
- MASTERS, G. N. and Wright, B. D. (1997). "The Partial Credit Model." W.J. VAN DER LINDEN ve R. K. HAMBLETON (Ed), *Handbook of Modern Item Response Theory*, Newyork; Springer, Chapter VI.

- POPLUN, M.; Omar MD H. (1997). "Multiple Mark Items: An Alternative Objective Item Format?." *Educational and Psychological Measurement, Vol. 57*: 949-962.
- STRASHNY, A. (2002). "A New Scoring Algorithm for Multiple-Choice Tests: Conditional Knowledge Model." Dept. Of Economics, UCI, 3151 Social Science Plaza A, Irvine, CA 92697-5100.
- TEKİN, H. (1993). *Eğitimde Ölçme ve Değerlendirme*, Ankara, Yargı Yayınları.
- THORNDIKE, R. L. (1971). *Educational Measurement*. Washington: American Council on Education.
- XIAOYING, M., Sireci S. G. (2004). "An Investigation of Polytomous Scoring of Multiple Response Items on A Certification Exam." American Educational Research Association Annual Meeting.
- VAN DER LINDEN, W. J., Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*, Springer, Newyork.
- WOOD, R. (1977). "Multiple Choice: A State of The Art Report." B. Choppin & T. N. Postlewaite (Ed). *Evaluation in Education: International Progress*. Oxford: Pergamon: 191-280.

## ÖZGEÇMİŞ

### Kişisel Bilgiler

Adı Soyadı : Bayram Çetin  
Doğum Yeri ve Tarihi : İslahiye 04.03.1972

### Eğitim Durumu

Lisans Öğrenimi : Çukurova Üniversitesi Eğitim Fakültesi Sınıf  
Öğretmenliği Bölümü  
Yüksek Lisans Öğrenimi : Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü  
Eğitim Bilimleri Anabilim Eğitimde Ölçme ve  
Değerlendirme Dalı Bilim Dalı  
Bildiği Yabancı Diller : İngilizce  
Bilimsel Faaliyetleri :

### İş Deneyimi

Stajlar :  
Projeler :  
Çalıştığı Kurumlar : Milli Eğitim Bakanlığı  
Sakarya Üniversitesi  
Hacettepe Üniversitesi

### İletişim

E-Posta Adresi : bayram@hacettepe.edu.tr  
Tarih : 10.06.2005