

ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİM BİLİMLERİ ANABİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI

KISA YANITLI MADDELERDE FARKLI AĞIRLIKL
PUANLANMA YÖNTEMLERİNİN TESTİN
PSİKOMETRİK ÖZELLİKLERİ AÇISINDAN
İNCELENMESİ

DOKTORA TEZİ

EMEL ILICAN

ANKARA
NİSAN, 2024



ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİM BİLİMLERİ ANABİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI

KISA YANITLI MADDELERDE FARKLI AĞIRLIKLI
PUANLANMA YÖNTEMLERİNİN TESTİN
PSİKOMETRİK ÖZELLİKLERİ AÇISINDAN
İNCELENMESİ

DOKTORA TEZİ

EMEL ILICAN

DANIŞMAN: PROF. DR. KAAAN ZÜLFİKAR DENİZ

ANKARA
NİSAN, 2024

Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Müdürlüğüne,

Emel Ilıcan adlı öğrencinin hazırladığı “Kısa Yanıtlı Maddelerde Farklı Ağırlıklı Puanlama Yöntemlerinin Testin Psikometrik Özellikleri Açısından İncelenmesi” başlıklı bu çalışma Eğitim Bilimleri Ana Bilim Dalı / Eğitimde Ölçme ve Değerlendirme Programı’nda jüri üyelerince oy birliği ile **Doktora Tezi** olarak kabul edilmiştir.

	Jüri Üyeleri	İmza
Başkan	Prof. Dr. Selahattin Gelbal	
Üye	Prof. Dr. Kaan Zülfikar Deniz (Danışman)	
Üye	Prof. Dr. Ömay Çokluk Bökeoğlu	
Üye	Doç. Dr. Celal Deha Doğan	
Üye	Doç. Dr. Ezgi Mor	

ONAY

Bu tez Ankara Üniversitesi Lisansüstü Eğitim Öğretim Yönetmeliği’nin ilgili maddeleri uyarınca jüri üyeleri tarafından 30/04/2024 tarihinde, Enstitü Yönetim Kurulu tarafından ise .../.../2024 tarihinde kabul edilmiştir.

.....

Prof. Dr. Mehmet İkbal YETİŞİR

Eğitim Bilimleri Enstitüsü Müdürü

ETİK İLKELERE UYGUNLUK BİLDİRİMİ

Tez içindeki bütün bilgileri akademik yazım kurallarına uygun biçimde raporlaştırdığımı ve bunları etik ilkelere (atıfta bulunulan tüm yapıtlara kaynaklarda yer verilmesi, tezde kullanılan bilgi ve belgelere resmi yollarla ulaşılması ve bunların aslı bozulmadan kullanılması vb.) uygun olarak elde ettiğimi ve sunduğumu bildiririm.

Emel ILICAN

ÖZET

KISA YANITLI MADDELERDE FARKLI AĞIRLIKLI PUANLANMA YÖNTEMLERİNİN TESTİN PSİKOMETRİK ÖZELLİKLERİ AÇISINDAN İNCELENMESİ

ILICAN, Emel

Doktora Tezi, Eğitim Bilimleri Anabilim Dalı

Tez Danışmanı: Prof. Dr. Kaan Zülfikar DENİZ

Nisan, 2024, xvi + 134 Sayfa

Ölçme ve değerlendirme uygulamalarındaki temel çalışma alanlarından test geliştirme ve puanlama çalışmaları, cevaplayıcılara ilişkin toplanan bilginin düzeyini ve özellikle elde edilen puanlar yoluyla verilen kararları etkileyebilmektedir. Bu araştırmada geliştirilen kısa yanıtli bir testin yanıtlarının; önce geleneksel puanlama (GP) sonra madde güçlük indekslerine dayalı ağırlıklı puanlama (MGİAP), olası cevapların ağırlıklandırılmasına dayalı puanlama (OCAP) ve çoklu ağırlıklandırmaya dayalı puanlama (ÇAP) yöntemleriyle puanlanması durumunda testin psikometrik özelliklerinin incelenmesi amaçlanmıştır. Araştırma verisi, ortaokul yedinci sınıf matematik dersine ilişkin 16 maddelik testin 2022-2023 öğretim yılında Ankara ilinde öğrenim gören 636 öğrenciye uygulanmasıyla elde edilmiştir. Testin puanlanmasında önce GP, ardından araştırmada incelenen OCAP yönteminin kullanılması durumunda klasik test ve madde tepki kuramlarına göre öncelikle test puanlarının geçerliğinin daha sonra test puanlarına ilişkin güvenirlik kestirimlerinin nasıl değişmekte olduğu araştırılmıştır. Ayrıca araştırma verisinden rastgele alt örnekleme yoluyla bu iki puanlama yöntemine ilişkin çok sayıda farklı veri seti oluşturularak elde edilen sonuçların genellenebilirliği incelenmiştir. Araştırmanın klasik test kuramına dayalı bulgularına göre, OCAP yönteminin kullanılmasının yapı geçerliği, uygunluk geçerliği ve güvenirlik bakımından genel olarak avantaj sağlayabileceği görülmüştür. Madde tepki kuramında OCAP yönteminin kullanıldığı durumda, cevaplayıcılar için daha fazla bilgi elde edildiği sonucuna ulaşılmıştır. Araştırma kapsamında üretilen veriyle yapılan geçerlik ve güvenirlik çalışmaları, madde faktör yükleri dışındaki tüm sonuçların

orijinal verinin kullanıldığı durumdaki sonuçlarla uyumlu olduğunu göstermiştir. Son olarak OCAP yönteminin yanında diğer iki ağırlıklı puanlama yönteminin kullanılması durumunda testin güvenilirliğinin ve geçerliğinin nasıl etkilendiği araştırılmıştır. Genel olarak OCAP yönteminin geçerlik ve güvenirlik bakımından diğer ağırlıklı puanlama yöntemlerine kıyasla daha büyük bir avantaj sağlayabileceği belirlenmiştir.

Anahtar kelimeler: kısa yanıtlı maddeler, puanlama yöntemleri, geçerlik, güvenirlik



SUMMARY

INVESTIGATION OF DIFFERENT WEIGHTED SCORING METHODS IN SHORT-ANSWERED ITEMS IN TERMS OF PSYCHOMETRIC PROPERTIES OF THE TEST

ILICAN, Emel

PhD., Department of Measurement and Evaluation

Thesis Advisor: Prof. Dr. Kaan Zülfikar DENİZ

April, 2024, xvi + 134 Pages

Test development and scoring studies, which are among the main areas of work in measurement and evaluation practices, can affect the level of information collected about the respondents and especially the decisions made through the scores obtained. In this research, the answers to a short-answer test developed within the scope of the research; It is aimed to examine the psychometric properties of the test if it is scored first with number-correct scoring (NCS) and then with weighted scoring based on item difficulty indices (IDIS), scoring based on weighting of possible answers (PAS) and scoring based on polyweighting (PS). The research data was obtained by administering the 16-item test for the seventh grade mathematics course in secondary school to 636 students studying in Ankara in the 2022-2023 academic year. If NCS was used in scoring the test and then the PAS examined in the study was used, it was first investigated how the validity of the test scores and then the reliability estimates of the test scores changed according to classical test and item response theories. In addition, the generalizability of the results obtained by generating many different data sets for these two scoring methods through random subsampling from the research data was examined. According to the findings of the study based on classical test theory, it was observed that the use of the PAS could provide general advantages in terms of construct validity, concurrent validity and reliability. It was concluded that when the PAS was used in item response theory, more information was obtained for the respondents. Validity and reliability studies conducted with the data generated within the scope of the research showed that all results except factor loadings were compatible with the results

obtained when original data was used. Finally, it was investigated how the reliability and validity of the test were affected when two other weighted scoring methods were used in addition to the PAS. In general, it has been determined that the PAS can provide a greater advantage in terms of validity and reliability compared to other weighted scoring methods.

Keywords: short-answered items, scoring methods, validity, reliability



ÖNSÖZ

Genel olarak bireyler hakkında bilgi toplama işi olarak ifade edilebilen, eğitim ve psikoloji alanlarına ilişkin ölçme ve değerlendirme uygulamalarında, bireylerin doğrudan gözlenemeyen psikolojik özelliklerinin ölçme araçları aracılığıyla ölçülerek nicelleştirilmesi esastır. Burada belirtilen nicelleştirme işlemi, yani bireyin ölçülen psikolojik özelliklere ilişkin durumunun nicel olarak ifade edilmesi ise puanlanma çalışmalarıyla yapılmaktadır. Puanlama sürecinin tüm ayrıntıları, öncelikle cevaplayıcılar hakkında toplanabilen bilginin düzeyini ve özellikle elde edilen puanlar yoluyla bireyler hakkında verilen kararları etkileyebilmektedir. Puanlama çalışmalarının yanında, madde formatı da cevaplayıcıların ölçülen özelliğe sahip olma düzeyine ilişkin bilgi düzeyinde önemli görülmektedir. Bu doğrultuda araştırmada kısa yanıtlı bir testin yanıtlarının; geleneksel puanlama ve çeşitli ağırlıklı puanlama yöntemleriyle puanlanması durumunda testin psikometrik özelliklerinin incelenmesi amaçlanmıştır.

Öncelikle bu çalışmanın her aşamasında değerli bilgilerini ve deneyimlerini benimle paylaşan, emeğini ve desteğini hiçbir zaman eksik etmeyen, danışmanım olduğu için kendimi hep şanslı hissettiğim değerli hocam Prof. Dr. Kaan Zülfikar Deniz'e saygılarımı ve teşekkürlerimi sunarım.

Tez izleme komitesinde bulunarak kendi yoğun çalışmalarına rağmen değerli bilgileri ve önerileriyle araştırmama katkı sağlayan saygıdeğer hocalarım Prof. Dr. Selahattin Gelbal'a ve Doç. Dr. Celal Deha Doğan'a; tez jürisinde bulunarak çok değerli katkılarıyla çalışmamı geliştiren Prof. Dr. Ömay Çokluk Bökeoğlu'na ve Doç. Dr. Ezgi Mor'a ve doktora eğitimim boyunca değerli bilgi ve tecrübelerinden yararlanma fırsatı bulduğum tüm hocalarıma teşekkür ederim.

Tanıştığımız günden beri ne zaman ihtiyaç duysam desteğini hissettiren, tez süreci boyunca da bilgi ve yardımlarıyla yanımda olan mesai arkadaşlarım Amine Canıdemir ve Onur Özyol'a; doktora ve tez süreci boyunca akademik bilgi ve deneyimleriyle bana destek olan arkadaşlarım Muammer Yüksel'e ve Arş. Gör. Dr. Tuba Gündüz'e teşekkürlerimi sunarım.

Araştırma verisinin elde edilmesinde yardımcı olan, uygulama yaptığım okullarda bulunan değerli yönetici ve öğretmenler ile sevgili öğrencilere teşekkür ederim.

Son olarak her zaman bana inanan ve sabırla yardımcı olan eşim Ömer İlcan ile varlıklarıyla bana güç veren sevgili çocuklarım Kerem ve Ece'ye sevgilerimi ve teşekkürlerimi sunarım.





Kerem ve Ece'ye...

İÇİNDEKİLER

ETİK İLKELERE UYGUNLUK BİLDİRİMİ.....	iii
ÖZET	iv
SUMMARY	vi
ÖNSÖZ.....	viii
İÇİNDEKİLER.....	xi
TABLolar DİZİNİ.....	xiii
ŞEKİLLER DİZİNİ	xv
KISALTMALAR/SİMGELER	xv
BÖLÜM 1.....	1
GİRİŞ.....	1
Problem	1
Amaç	6
Önem.....	7
Sayıtlar.....	8
Sınırlılıklar	9
Tanımlar.....	9
BÖLÜM 2.....	10
KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR.....	10
Puanlama Yöntemleri	10
Geleneksel Puanlama Yöntemi	10
Çok Kategorili/Ağırlıklı Puanlama Yöntemleri.....	13
Farklı Madde Türlerinde Puanlama	32
Çoktan Seçmeli Maddelerin Puanlanması.....	33
Kısa Yanıtlı Maddelerin Puanlanması.....	35
BÖLÜM 3.....	41
YÖNTEM	41
Araştırmanın Modeli	41
Çalışma Grubu	41
Verilerin Toplanması	42
Madde Güçlük İndekslerine Dayalı Ağırlıklı Puanların Elde Edilmesi	59

Olası Cevapların Deneyisel Olarak Ağırlıklandırılmasına Dayalı Ağırlıklı Puanların Elde Edilmesi	60
Çoklu Ağırlıklandırmaya Dayalı Ağırlıklı Puanların Elde Edilmesi	63
Verilerin Çözümlemesi	66
BÖLÜM 4.....	71
BULGULAR VE YORUMLAR	71
Birinci Alt Probleme İlişkin Bulgular ve Yorumlar	71
İkinci Alt Probleme İlişkin Bulgular ve Yorumlar	79
GP Yöntemiyle Elde Edilen Veri için MTK Kalibrasyonları	81
OCAP Yöntemiyle Elde Edilen Veri için MTK Kalibrasyonları.....	85
Üçüncü Alt Probleme İlişkin Bulgular ve Yorumlar	91
Dördüncü Alt Probleme İlişkin Bulgular ve Yorumlar.....	98
BÖLÜM 5.....	110
SONUÇLAR VE ÖNERİLER	110
Sonuçlar	110
Öneriler	113
Uygulayıcılar İçin Öneriler	113
Araştırmacılar İçin Öneriler	114
KAYNAKLAR.....	115
EKLER	124
EK 1. OCAP Yönteminin Kullanıldığı Duruma İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları	125
EK 2. OCAP Yönteminin Kullanıldığı Durumda Ölçme Aracının Tek Faktörlü Modeli için Yol Şeması	126
EK 3. MGİAP Yönteminin Kullanıldığı Duruma İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları	127
EK 4. MGİAP Yönteminin Kullanıldığı Durumda Ölçme Aracının Tek Faktörlü Modeli için Yol Şeması	128
EK 5. ÇAP Yönteminin Kullanıldığı Duruma İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları	129
EK 6. ÇAP Yönteminin Kullanıldığı Durumda Ölçme Aracının Tek Faktörlü Modeli için Yol Şeması.....	130
EK 7. Etik Kurul Onayı	131
EK 8. Millî Eğitim Bakanlığı Araştırma İzni	132
BENZERLİK BİLDİRİMİ	133
ÖZGEÇMİŞ.....	134

TABLolar DİZİNİ

Tablo	Sayfa
Tablo 1. Çalışma Grubunda Bulunan Okullar ile Öğrenci Sayıları.....	42
Tablo 2. Aday Maddelerin Davis Tekniğine Göre Hesaplanan Kapsam Geçerlik İndeksleri.....	43
Tablo 3. Deneme Formuna İlişkin Test İstatistikleri.....	45
Tablo 4. Deneme Formunun Madde İstatistikleri.....	46
Tablo 5. Paralel Analizden Elde Edilen Kriter Değerlerin Özdeğerlerle Karşılaştırılması.....	48
Tablo 6. Deneme Formundaki Maddelerin Faktör Yük Değerleri ve Ortak Faktör Varyansları.....	49
Tablo 7. Deneme Formuna İlişkin DFA Sonucunda Elde Edilen Uyum İndeksleri.....	50
Tablo 8. Nihai Uygulama Formuna İlişkin Test İstatistikleri.....	52
Tablo 9. Nihai Uygulama Formunun Madde İstatistikleri.....	53
Tablo 10. Nihai Uygulamaya İlişkin Paralel Analizden Elde Edilen Kriter Değerlerin Özdeğerlerle Karşılaştırılması.....	55
Tablo 11. Nihai Test Formundaki Maddelerin Faktör Yük Değerleri ve Ortak Faktör Varyansları.....	56
Tablo 12. Nihai Test Formuna İlişkin DFA Sonucunda Elde Edilen Uyum İndeksleri.....	57
Tablo 13. Nihai Uygulama Formundaki Maddelerin Madde Güçlük İndeksleri ve Ağırlıklı Puanlar.....	59
Tablo 14. Maddelerin Olası Cevaplarına İlişkin Örnek Bir Kodlama.....	61
Tablo 15. Maddelerin Olası Cevapları ile Toplam Puanlar Arasındaki Nokta Çift Serili Korelasyon Katsayıları.....	62
Tablo 16. Maddelerin Olası Cevapları için Hesaplanan Çoklu Ağırlıklar.....	65
Tablo 17. Puanlama Yöntemleriyle Elde Edilen Betimsel İstatistikler.....	71
Tablo 18. GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Maddelerin Faktör Yük Değerleri.....	73
Tablo 19. GP ve OCAP Yöntemlerinde AFA Sonucunda Elde Edilen Faktör Yük Değerlerinin Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları.....	74
Tablo 20. GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Uyum İndeksleri.....	74
Tablo 21. GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda DFA Sonucunda Görülen Hata Varyansları.....	75
Tablo 22. GP ve OCAP Yöntemlerinde DFA Sonucunda Elde Edilen Hata Varyanslarının Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları.....	76

Tablo 23. GP Yönteminin Kullanıldığı Durumda Maddeler Arasındaki Q3 Değerleri	80
Tablo 24. OCAP Yönteminin Kullanıldığı Durumda Maddeler Arasındaki Q3 Değerleri.....	81
Tablo 25. GP Yöntemiyle Elde Edilen Veri için MTK Model Karşılaştırma Sonuçları.....	82
Tablo 26. GP Yöntemiyle Elde Edilen Veri için Madde Uyum İndeksleri.....	82
Tablo 27. 2PLM'ye Göre Elde Edilen Madde Parametreleri	83
Tablo 28. OCAP Yöntemiyle Elde Edilen Veri için MTK Model Karşılaştırma Sonuçları	86
Tablo 29. OCAP Yöntemiyle Elde Edilen Veri için Madde Uyum İndeksleri	86
Tablo 30. KTM'ye Göre Elde Edilen Madde Parametreleri	87
Tablo 31. GP ve OCAP Yöntemlerinde AFA Sonucunda Elde Edilen Açıklanan Varyans Değerlerinin Karşılaştırılmasına İlişkin t Testi Sonuçları	92
Tablo 32. GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Maddelerin Ortalama Faktör Yük Değerleri	93
Tablo 33. GP ve OCAP Yöntemlerinde AFA Sonucunda Elde Edilen Ortalama Faktör Yük Değerlerinin Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları	93
Tablo 34. GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Ortalama Uyum İndeksleri.....	94
Tablo 35. GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda DFA Sonucunda Görülen Ortalama Hata Varyansları	95
Tablo 36. GP ve OCAP Yöntemlerinde DFA Sonucunda Elde Edilen Ortalama Hata Varyanslarının Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları.....	95
Tablo 37. GP ve OCAP Yöntemlerinde Güvenirlik Katsayılarının Karşılaştırılmasına İlişkin t Testi Sonuçları	97
Tablo 38. Ağırlıklı Puanlama Yöntemlerinin Kullanıldığı Durumlarda Maddelerin Faktör Yük Değerleri	100
Tablo 39. AFA Sonucunda Elde Edilen Faktör Yük Değerlerinin Ağırlıklı Puanlama Yöntemine Göre Karşılaştırılmasına İlişkin Kruskal Wallis H Testi Sonuçları	101
Tablo 40. OCAP, MGİAP ve ÇAP Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Uyum İndeksleri.....	102
Tablo 41. Ağırlıklı Puanlama Yöntemlerinin Kullanıldığı Durumlarda DFA Sonucunda Görülen Hata Varyansları	103
Tablo 42. DFA Sonucunda Elde Edilen Hata Varyanslarının Ağırlıklı Puanlama Yöntemine Göre Karşılaştırılmasına İlişkin Kruskal Wallis H Testi Sonuçları	104
Tablo 43. Ağırlıklı Puanlama Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Güvenirlik Katsayıları.....	106

ŞEKİLLER DİZİNİ

Şekil	Sayfa
Şekil 1. Geleneksel Puanlama Yöntemi	11
Şekil 2. Ağırlıklı Puanlama Yöntemleri	14
Şekil 3. Deneme Formuna İlişkin Özdeğer Grafiği	47
Şekil 4. Deneme Formuna İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları	48
Şekil 5. Deneme Formunun Tek Faktörlü Modeli için Yol Şeması	51
Şekil 6. Nihai Test Formuna İlişkin Özdeğer Grafiği	54
Şekil 7. Nihai Test Formuna İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları	55
Şekil 8. Nihai Test Formunun Tek Faktörlü Modeli İçin Yol Şeması	58
Şekil 9. GP Yönteminin Kullanıldığı Durumda Madde Bilgi Fonksiyonları	84
Şekil 10. GP Yönteminin Kullanıldığı Durumda Test Bilgi Fonksiyonu	85
Şekil 11. OCAP Yönteminin Kullanıldığı Durumda Madde Bilgi Fonksiyonları	88
Şekil 12. OCAP Yönteminin Kullanıldığı Durumda Test Bilgi Fonksiyonu	88
Şekil 13. GP ve OCAP Yönteminin Kullanıldığı Durumlara İlişkin Test Bilgi Fonksiyonları	89

KISALTMALAR/SİMGELER

GP	Geleneksel Puanlama
MGİAP	Madde Güçlük İndekslerine Dayalı Ağırlıklı Puanlama
OCAP	Olası Cevapların Ağırlıklandırılmasına Dayalı Puanlama
ÇAP	Çoklu Ağırlıklandırmaya Dayalı Puanlama
AFA	Açımlayıcı Faktör Analizi
DFA	Doğrulayıcı Faktör Analizi
KTK	Klasik Test Kuramı
MTK	Madde Tepki Kuramı
θ	Yetenek Düzeyi
p_j	Madde Güçlük İndeksi
r_{jx}	Madde Ayırt Edicilik İndeksi
ε	Hata Varyansı
OC	Olası Cevap
1PLM	Bir Parametrelili Lojistik Model
2PLM	İki Parametrelili Lojistik Model
3PLM	Üç Parametrelili Lojistik Model
GKPM	Genelleştirilmiş Kısmi Puan Modeli
KTM	Kademeli Tepki Modeli

BÖLÜM 1

GİRİŞ

Bu bölümde, araştırmanın problemi açıklanmıştır. Ayrıca araştırmanın amacı, önemi ve sınırlılıkları ile araştırmadaki önemli kavramların tanımları verilmiştir.

Problem

Eğitim ve psikoloji bakımından ele alındığında ölçme, bireyler hakkında bilgi toplama işidir (Thorndike ve Hagen, 1961). Söz konusu bilgi toplama işi kapsamında yapılan ölçme ve değerlendirme uygulamalarının özünde, bireylerin doğrudan gözlenemeyen psikolojik özelliklerinin geliştirilen ölçme araçları aracılığıyla ölçülerek nicelleştirilmesi yer almaktadır (Lord ve Novick, 2008). Bu nicelleştirme işlemi, yani bireyin ölçülen psikolojik özelliklere ilişkin durumunun nicel veriler olarak ifade edilmesi ise ölçme aracını oluşturan maddelerin puanlanması yoluyla yapılmaktadır. Maddelerin nasıl puanlandığı, bir ölçme aracında yer alan maddelerden elde edilecek bilgi miktarını artırmak için göz önünde tutulması gereken üç bileşenden biridir. Diğer iki bileşen ise cevaplayıcının maddeye nasıl yanıt vereceğini ifade eden madde formatı ve maddelerin bir toplam puan oluşturmak için nasıl bir araya getirildiğidir (Echternacht, 1976).

Puanlama işlemi, elde edilen puanların çeşitli psikometrik özelliklerinin yanı sıra test ve madde istatistiklerini de önemli ölçüde etkilemektedir (Kan, 2005). Puanlamaya ilişkin her bir ayrıntı, özellikle bu puanlar yoluyla bireyler hakkında verilen kararları etkilemektedir. Testlerle yapılan değerlendirmelerin niteliği, sadece maddelerle değil; o maddelerin puanlanmasıyla da ilgili olmaktadır. Puanlama işlemlerinde maddeden maddeye veya bir öğrenciden diğerine değişiklik yapılması ölçmeye hata karışmasına yol açmaktadır. Bu nedenle öncelikle puanlamanın niteliği ölçme ve değerlendirme sürecinde oldukça önemlidir. Bunun yanında puanlama işlemlerine ilişkin ayrıntıları içeren puanlama yöntemleri de değerlendirme sürecinde büyük bir öneme sahiptir. Çünkü bu yöntemler, ilgili ölçme aracının özellikle geçerlik gibi psikometrik özellikleri üzerinde çeşitli etkilere sahip olmaktadır (Burt, 1950).

Ölçme alanyazını incelendiğinde puanlama sürecinde kullanılan birçok farklı yaklaşım ve yöntemin ileri sürüldüğü görülmektedir. Puanlama yöntemleri temel olarak iki gruba ayrılmaktadır. Bunlar geleneksel puanlama yöntemleri ve ağırlıklı puanlama yöntemleridir. Geleneksel puanlama, “1-0” puanlama veya doğru yanıt sayısı metodu biçiminde isimlendirilen puanlama modellerinde adayların cevapları doğru veya yanlış olarak iki kategoride sınıflandırılmaktadır (Crocker ve Algina, 1986). Bu yöntemin kullanıldığı bir puanlama sürecinde, maddeyi cevap anahtarına tam uygun şekilde cevaplayan veya çoktan seçmeli testlerde doğru işaretleyen aday o maddeyi doğru cevaplamış sayılarak adaya “1” puan verilir. Bunun dışındaki tüm durumlarda, yani çoktan seçmeli testler için maddede işaretleme yapmayan, çeldiricilerden birini veya birden fazla seçeneği işaretleyen adayın ise maddeyi yanlış yanıtlamış olduğu kabul edilir. Benzer şekilde kısa yanıtlı testlerde de aday, tam doğru cevabın dışında herhangi bir cevap verdiğyse ilgili maddeden puan almamaktadır. Yani çoktan seçmeli maddelerin dışındaki maddelerle 1-0 puanlama yöntemi kullanıldığında maddeyi tam olarak puanlama anahtarında ifade edildiği şekilde cevaplayan adaylar tam puan/”1” puan, bu adayların dışındaki tüm adaylar yine “0” puan almaktadır. Böylece hiç bilmeyen veya tamamen yanlış bilgisi olan cevaplayıcıların yanında kısmi bilgiye sahip olan cevaplayıcıların da cevapları yanlış cevap kategorisinde sayılmaktadır. Bu nedenle yalnızca maddeye ilişkin tam bilgiye sahip olan ve seçme gerektiren maddeler için bir ölçüde tahminle doğru seçeneği işaretleyebilenler dışındaki adayların tamamı, madde için sıfır puan almaktadır (Jaradat & Tollefson, 1988).

Kullanılan madde türü hangisi olursa olsun geleneksel puanlama yönteminin kullanıldığı durumda toplam puan hesaplanırken maddelerin her biri eşit ağırlıklara sahip olmaktadır. Testte yer alan maddelerin tamamının ölçülen özelliği eşit bir biçimde yokladığı/sınadığı kabul edilmektedir. Ayrıca çoktan seçmeli madde türü özelinde bir maddedeki çeldiricilerin tamamının aynı düzeyde çekici oldukları varsayılmaktadır (Haladyna, 1990). Yöntemde kabul edilen bu durumlar, yaygın olarak kullanılmasına rağmen aynı zamanda geleneksel puanlama yönteminin eleştirilmesine sebep olmaktadır. Çünkü özellikle kısmi bilginin puanlamada dikkate alınmıyor olması, test puanlarının yeterince yorumlanamamasına neden olmaktadır (Lau, Lau, Hong & Usop, 2011). Oysaki ölçme uygulamalarının temel hedeflerinden biri, belirli bir madde kümesinden cevaplayıcılara ilişkin mümkün olduğu kadar fazla bilgi çıkarmaktır (Echternacht, 1976).

Geleneksel puanlama yöntemine ilişkin eleştirilerin bir diğer kaynağı da bu yöntemin çeldirici seçimlerini ayırt etmemesidir (Diedenhofen ve Musch, 2015). Çeldiriciler, maddenin mantıklı yanlış yanıtları olarak hazırlandığına göre (Haladyna, 1997), her bir çeldiricinin tercihi cevaplayıcılarla ilgili belirli düzeyde kısmi bilgiyi veya bilgi eksikliğini işaret ediyor olabilmektedir. Çünkü bir maddeye verilen cevaplar, farklı bilgi seviyelerini işaret edebilmektedir. Bu seviyeler; tam bilgi, kısmi bilgi ve yanlış bilgi ile cevaplayıcıda bilgi olmaması durumlarıdır (Garcia-Perez ve Frary, 1989). Tam bilgiye sahip olan cevaplayıcı, maddenin doğru cevabının yanında çeldiricilerin de farkında olup doğru seçeneği işaretlemektedir. Bir diğer bilgi seviyesi olan yanlış bilgi durumunda cevaplayıcı doğru yanıtın yanlış olduğuna düşünüp onu çeldirici kabul etmektedir. Bilgi yokluğu ise maddenin boş bırakılmasını veya birden fazla seçeneğin rastgele işaretlenmesini ifade etmektedir. Kısmi bilgiye sahip cevaplayıcı ise doğru olmayan bazı seçeneklerin farkındadır ancak bilgi eksikliği sebebiyle doğru seçeneği tam olarak belirleyememektedir (Ben-Simon, Budescu ve Nevo, 1997). Bu nedenle kısmi bilgisi olan yanıtlayıcı maddeyi doğru ya da yanlış yanıtlayabilmektedir (Wongwiwatthananukit, Popovich ve Bennett, 2000). Buradaki belirleyici özellik cevaplayıcının bilgi seviyesi bakımından orta düzeyde olması, dolayısıyla doğru ve yanlış cevap arasındaki bir yerde konumlanmasıdır (Abu-Sayf, 1979). Kısmi bilgi, uygun alanlar için, çözüme doğru başlamak fakat bir yanlışla devam etmek veya çözüm adımlarından birinin atlanması nedeniyle doğru cevabı tam olarak oluşturamamak biçiminde ifade edilebilmektedir (Grunert, Raker, Murphy ve Holme, 2013). Cevaplayıcıların yetenek kestirimlerinin daha doğru yapılabilmesi için kısmi bilgi ve yanlış bilginin birbirlerinden doğru şekilde ayırt edilmesi gerekmektedir. Çünkü bir testten aynı sayıda tam doğru cevaba sahip oldukları için aynı puanı almış olan bireyler, “0” puan ile değerlendirilen yanlış cevaplarda aslında farklı düzeylerde kısmi bilgiye sahip olabilmektedir. Böylece bir puanlama işlemine ilişkin kısmi bilginin dikkate alınıp alınmaması durumu test puanlarına ilişkin geçerliği ve güvenilirliği etkilemektedir (Kurz, 1999).

İki kategorili puanlama yöntemiyle ilgili olarak eleştirilen bu durumlarla başa çıkmak için çok kategorili/ağırlıklı puanlama yöntemleri önerilmiştir. Bu yöntemler, adayların yetenek kestirimlerinin daha az hatayla yapılmasını, kısmi bilgilerin belirlenebilmesini, geçerliğin ve güvenilirliğin yükselmesini, ayrıca kısmi bilgisi olan ancak cevaplama sırasında risk alma davranışını daha az sergileyen yanıtlayıcıların cezalandırılmamasını vaat etmektedir (Kurz, 1999). Çok kategorili puanlamada,

cevaplayıcının kendisinde mevcut olan farklı düzeylerdeki bilgisini çeldiricilerde kullanabildiği kabul edilmektedir (Crehan & Haladyna, 1994). Alanyazın incelendiğinde ağırlıklı puanlama yöntemleri için pek çok farklı sınıflamanın önerildiği görülmektedir.

Bir teste ilişkin kullanılacak puanlama yönteminin belirlenmesinde etkili olan pek çok faktör bulunmaktadır. Örneğin puanlamanın kolay ve objektif olması, sağlam bir istatistik temelinin olması nedeniyle geleneksel puanlama yöntemi yaygın biçimde kullanılmaktadır. Öte yandan kişilik, tutum ve ilgi envanterleri gibi doğru yanıt seçeneği olmayan testlerde cevap kategorilerini ağırlıklandırma için ağırlıklı puanlama yönteminin kullanımı önerilmektedir. Ayrıca bir başarı testinde çoktan seçmeli test maddelerini yazarken, çeldiricilerin eşit yüzdelere seçilmelerini sağlayacak şekilde düzenleme yapmak zor ise seçenek ağırlıklandırma yöntemlerinin kullanılması gerektiği belirtilmektedir (Wang ve Stanley, 1970). Kullanılacak puanlama yöntemi seçilirken testte kullanılan madde türünün de göz önünde bulundurulması önem arz etmektedir.

Test uygulamalarında en yaygın kullanılan iki madde türü olan çoktan seçmeli maddeler ile kısa yanıtlı maddelerin puanlama süreçlerinde de farklı birçok yöntem kullanılabilmektedir. Çoktan seçmeli maddeler yaygın olarak geleneksel puanlama yöntemiyle puanlanmasına rağmen bu maddelerin puanlanmasında ağırlıklı puanlama yöntemlerinden seçenek ağırlıklandırma veya madde ağırlıklandırmaya dayalı puanlama yöntemleri de kullanılabilmektedir. Ancak ağırlıklı puanlama yöntemleri kullanıldığında cevaplayıcı davranışları (Frery, 1982), bu davranışlardan özellikle şans başarısı her durumda elde edilen madde/test puanlarında ve bunlara dayalı olarak hesaplanan istatistiklerde veya yapılan ağırlıklandırmalarda hatalara sebep olmaktadır (Baykul, 2015).

Çoktan seçmeli testlerin puanlanmasında baş edilmesi gereken bir problem olarak karşımıza çıkan şans başarısını önlemek veya etkisini azaltmak için düzeltme formülleri geliştirilmiştir (Horst, 1966). Ancak en çok kullanılan düzeltme formülü olan “doğru yanıt sayısı – [yanlış yanıt sayısı/(seçenek sayısı – 1)]” formülünün kullanıldığı durumda, cevaplayıcının hiçbir seçeneği yanlış seçenek olarak eleyemediği, sunulan seçeneklerden hangisinin doğru olduğunu hiç bilmediği kabul edilmektedir. Bu durumda bazı çeldiriciyi/çeldiricileri eleyebilen ve hiçbir çeldiriciyi eleyemeyen öğrenciler birbirinden ayırt edilemediği için doğru bir düzeltme yapılamamaktadır. Örneğin; beş seçenekli bir maddede üç çeldiriciyi eleyebilen bir öğrenci ile çeldiricilerden hiçbirini eleyemeyen öğrenciye aynı formülün kullanılması, hiç

bilmeyenle kısmi bilgiye sahip cevaplayıcı arasındaki ayrımın yapılamamasına neden olmaktadır (Yaman, 2016). Düzeltme formülünün dışında, madde tepki kuramına dayalı üç parametrelili lojistik modelin kullanılması, böylece şansla doğru cevap verme ihtimali dışarıda tutularak gerçek puanların kestirilmesi önerilmiştir (Baker, 2001). Bu durumda da çoktan seçmeli testlerde şans başarısının etkisi yine de elde edilen sonuçları etkilediğinden (Özdoğan ve Doğan, 2018) çoktan seçmeli maddelerin yerine şans başarısından etkilenmeyen madde türlerinin kullanılması üzerinde durulmuştur.

Şans başarısı probleminden etkilenmeyen madde türlerinden biri olan kısa yanıtlı maddeler, çoktan seçmeli maddeler gibi “1-0” şeklinde (Umay, 1997; Kubiszyn & Borich, 2003) veya puanlama anahtarıyla (Doğan, 2020) puanlanabilmektedir. Geleneksel puanlama yönteminin kullanılmasına ilişkin söz edilen dezavantajlar kısa yanıtlı maddelerin kullanılması durumunda da benzer şekilde ortaya çıkmaktadır. Alternatif olarak kısa cevaplı testlerde puanlama anahtarı kullanılarak puanlayıcılar aracılığıyla puanlama yapılması durumunda ise puanlayıcıların yanlış davranarak öznel kararlar alması, puanlama anahtarının net olmaması gibi hata kaynaklarından ötürü puanlama güvenilirliği düşmektedir (Gültekin, 2017). Öte yandan puanlayıcıların görevlendirilmesi, eğitilmesi, puanlayıcılarla ilgili güvenliliğin kontrol altında tutulması ihtiyacı da puanlama anahtarıyla puanlama yapılırken ortaya çıkan unsurlardır. Aynı zamanda bahsedilen bu unsurların tamamı hem zaman hem de maliyet bakımından puanlama sürecine ciddi bir yük oluşturmaktadır. Başka bir alternatif olarak kısa yanıtlı testler ağırlıklı puanlama yöntemleriyle de puanlanabilmektedir. Ancak kısa yanıtlı maddelerin ağırlıklı puanlama yöntemlerinden sadece madde ağırlıklandırma yöntemleriyle puanlandığı az sayıda çalışmaya (Gözen, 2002, Phillips, 1943; Quist, 2004) rastlanmıştır. Öte yandan özellikle çoktan seçmeli maddeler çoğunlukta olmak üzere, farklı madde türleriyle birlikte ağırlıklı puanlama yöntemlerinin kullanıldığı çalışmalardan elde edilen sonuçların genel olarak tutarsız sayılabileceği görülmektedir. Bununla birlikte bu tutarsızlığın söz konusu çalışmalarda kullanılan testlerin ve çalışma evreninin özelliklerindeki büyük farklılıklardan ve kullanılan ölçme araçlarının niteliğinden kaynaklandığı da ifade edilmektedir (Frery, 1982). Ayrıca ağırlıklı puanlama yöntemlerinin daha uzun zaman aldığı ve sonuçta elde edilen puanların güvenilirlik ve geçerliklerinin harcanan çabaya değecek kadar yüksek olmadığını savunan çalışmalara (Akkuş, 2000; Guilford, Lovell ve Williams, 1942; Phillips, 1943) rastlanmış olsa da genellikle ağırlıklı puanlamanın birey hakkında daha fazla bilgi verdiği ve cevaplayıcıların kısmi bilgilerini de dikkate aldığı için kullanılmasının

faydalı olacağı belirtilmiştir. Günümüzde bilgisayar teknolojisinin oldukça gelişmiş olması sayesinde ağırlıklandırma işlemleri daha kolay/pratik şekilde yapılabilmektedir. Bu nedenle ağırlıklı puanlamanın testin psikometrik özelliklerine verebileceği küçük bir katkının bile cevaplayıcılar hakkında daha çok bilgi edinebilmek adına önemli olduğu düşünülmektedir. Özetle, kısmi bilgiyi dikkate alabilen, şans başarısından etkilenmeyen, maddelerin yoklanan bilgiyi ölçme bakımından farklı ağırlıkta/nitelikte olduğunu kabul eden ağırlıklı puanlama yöntemlerinin kısa yanıtli maddelerin puanlanmasında kullanılması durumunda elde edilecek sonuçların daha fazla araştırılmasına ihtiyaç duyulduğu düşünülmektedir. Ayrıca alanyazın taramasında kısa cevapli test maddelerinin ağırlıklı puanlama yöntemlerinden sadece madde ağırlıklandırma yöntemleriyle puanlandığı sınırlı sayıda çalışma olduğu görüldüğünden, çoktan seçmeli maddelerin seçeneklerinin ağırlıklandırılması gibi, kısa cevapli maddelerde de cevaplayıcıların kısmi bilgiyle oluşturduğu olası cevaplarından yola çıkılarak bir ağırlıklandırma işlemi yapılabileceği düşünülmektedir. Bu durumda öncelikle kısa cevapli bir testte böyle bir ağırlıklı puanlama yönteminin geleneksel puanlama yönteminin kullanıldığı duruma kıyasla testin psikometrik özelliklerinde bir iyileşme sağlayıp sağlamadığının araştırılmasına ihtiyaç duyulmaktadır. Ayrıca mevcut araştırmada önerilen bu puanlama yönteminin ve bilinen diğer ağırlıklı puanlama yöntemlerinin kullanıldığı durumlarda testin psikometrik özelliklerinin incelenmesi gerektiği düşünülmektedir. Bu doğrultuda kısa yanıtli bir testin yanıtlarının önce “1-0” puanlama daha sonra çeşitli ağırlıklı puanlama yöntemleriyle puanlanması durumlarında testin psikometrik özelliklerinin incelenmesi bu araştırmanın problemini oluşturmaktadır.

Amaç

Bu araştırmada ortaokul yedinci sınıf matematik dersine ilişkin geliştirilen kısa yanıtli bir testin yanıtlarının; önce “1-0” puanlama daha sonra maddelerin, olası cevapların hem maddelerin hem de olası cevapların ağırlıklandırılmasına dayalı ağırlıklı puanlama yöntemleriyle puanlanması durumunda testin psikometrik özelliklerinin incelenmesi amaçlanmıştır. Bu genel amaç doğrultusunda araştırma kapsamında aşağıda yer alan sorularla ilgili incelemeler yapılmıştır:

1. Kısa yanıtli bir testin puanlanmasında önce “1-0” puanlama, daha sonra olası cevapların ağırlıklandırılmasına dayalı ağırlıklı puanlama yönteminin kullanılması durumunda klasik test kuramına göre;

a) test puanlarının geçerliğı nasıl değışmektedir?

b) test puanlarına ilişkin güvenirlilik kestirimleri nasıl değışmektedir?

2. Kısa yanıtli bir testin puanlanmasında önce “1-0” puanlama, daha sonra olası cevapların ağırlıklandırılmasına dayalı ağırlıklı puanlama yönteminin kullanılması durumunda madde tepki kuramına göre;

a) test puanlarının geçerliğı nasıl değışmektedir?

b) test puanlarına ilişkin güvenirlilik kestirimleri nasıl değışmektedir?

3. Kısa yanıtli testin puanlanmasında önce “1-0” puanlama, daha sonra olası cevapların ağırlıklandırılmasına dayalı ağırlıklı puanlama yönteminin kullanılmasıyla elde edilen orijinal veriden rastgele 300 kişilik 100’er tane veri seti oluşturulması durumunda bunların her birine ilişkin geçerlik ve güvenirlilik analizleri tekrarlanırsa, kullanılan puanlama yöntemine göre;

a) test puanlarının geçerliğı nasıl değışmektedir?

b) test puanlarına ilişkin güvenirlilik kestirimleri nasıl değışmektedir?

4. Kısa yanıtli bir testin puanlanmasında maddelerin, olası cevapların, hem maddelerin hem de olası cevapların ağırlıklandırılmasına dayalı ağırlıklı puanlama yöntemlerinin kullanılması durumunda;

a) test puanlarının geçerliğı nasıl değışmektedir?

b) test puanlarına ilişkin güvenirlilik kestirimleri nasıl değışmektedir?

Önem

Kısa yanıtli bir testin yanıtlarının önce “1-0” puanlama yöntemiyle, daha sonra çeşitli ağırlıklı puanlama yöntemleriyle puanlanması durumlarında testin psikometrik özelliklerinin incelenmesini amaçlayan bu araştırmanın öncelikle cevaplayıcıların kısmi bilgisini de göz önünde bulunduran ve olası cevapların ağırlıklandırılmasına dayalı bir puanlama yöntemi önererek bu puanlamanın avantaj ve dezavantajlarını ortaya koyması bakımından alanyazına katkı sağlayacağı öngörülmektedir. Ölçme ve değerlendirme uygulamalarında geçmişten günümüze puanlama kolaylığı sağlaması nedeniyle çoktan seçmeli testler geleneksel puanlama yöntemiyle birlikte sıklıkla kullanılmıştır. Ancak

özellikle 1950’li yıllardan itibaren “1-0” puanlama yöntemine olan eleştirilerin çoğaldığı ve alternatif puanlama yaklaşımlarına ilişkin araştırmaların da oldukça arttığı görülmektedir (Coombs, Milholland ve Womer, 1956; Frary, 1982; Stanley ve Wang, 1970). Bu doğrultuda farklı ağırlıklı puanlama yöntemleri, incelenen araştırmaların çoğunda (Akkuş, 2000; Crehan ve Haladyna, 1994; Davis ve Fifer, 1959; Diedenhofen ve Musch, 2015; Gözen Çıtak, 2007; Haladyna ve Sympson, 1988; Haladyna, 1990; Özdemir, 2002; Özdemir, 2003; Salehi, Sadighi ve Bagheri, 2015; Saygı, 2004) yine çoktan seçmeli testlerin puanlanması sürecinde kullanılmıştır. Bunun yanında alanyazın taramasında araştırmacıların ağırlıklı puanlama yöntemlerini daha çok seçenek ağırlıklandırma yoluyla çoktan seçmeli maddelerin puanlanmasında kullandığı ve kısa cevaplı test gibi farklı türdeki madde türleriyle birlikte maddelere ilişkin ağırlıklandırmaların yapıldığı çalışmaların (Gözen, 2002, Phillips, 1943; Quist, 2004) daha az sayıda olduğu belirlenmiştir. Ayrıca kısa cevaplı testlerde, puanlama anahtarı kullanılarak puanlayıcılar aracılığıyla puanlama yapılması dışında, kısmi doğru cevapların dikkate alındığı veya cevaplayıcıların oluşturabileceği olası cevaplara dayalı olarak ağırlıklı puanlama işleminin yapıldığı bir çalışmaya rastlanmamıştır. Bu nedenle öncelikle bu araştırmada önerilen olası cevapların ağırlıklandırılmasına dayalı puanlama yöntemlerinin testin psikometrik özellikleri bakımından ortaya çıkaracağı sonuçların incelenmesi önem arz etmektedir. Ayrıca kısa cevaplı bir testin halihazırda kullanılan ve farklı ağırlıklı puanlama durumlarından elde edilecek psikometrik özelliklerini ortaya koyacak kapsamlı bir çalışmanın alanyazına katkı sağlayacağı düşünülmektedir. Araştırma ile önerilen puanlamanın kısa cevaplı testlerin psikometrik özelliklerine ilişkin yüksek düzeyde kestirimler sağladığının bulunması durumunda, kısa cevaplı testler için de puanlama kolaylığı sağlayan aynı zamanda öğrencilerin kısmi bilgilerinin görüldüğü olası cevaplarını da dikkate alan puanlama uygulamalarının önerilebileceği öngörülmektedir.

Sayıtlar

Araştırma kapsamında kullanılan testi cevaplarken öğrencilerin gerçek performanslarını ortaya koyabilecekleri bir ortam düzenlenmiş ve öğrenciler yeterince güdülenmiştir.

Sınırlılıklar

Bu araştırma matematik dersi rasyonel sayılar konusundaki öğrenme düzeylerini ölçmeye yönelik yalnızca kısa cevaplı maddelerden oluşan bir test aracılığıyla elde edilen verilerle sınırlıdır.

Tanımlar

Ağırlıklı Puan: Araştırma kapsamında geliştirilen kısa cevaplı testin yanıtlarının ağırlıklı/çok kategorili puanlama yöntemleri kullanılarak puanlanmasıyla elde edilen puanlardır.

Kısmi Bilgi: Cevaplayıcının maddeyi çözümleme sürecine doğru başladıktan sonra çözüm adımlarından birinde hata yapması veya bu adımlardan birini atlaması nedeniyle doğru cevaba ulaşamaması durumudur.

BÖLÜM 2

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Bu bölümde eğitimde ölçme ve değerlendirme uygulamalarının önemli bir parçası olan puanlama sürecine ilişkin farklı yaklaşım ve yöntemlerle ilgili bilgiler sunulmuştur. Ayrıca söz konusu puanlama yöntemlerinin farklı madde türleriyle birlikte kullanıldığı çeşitli durumlardaki puanlama uygulamalarıyla ilgili bilgilere yer verilmiştir.

Puanlama Yöntemleri

Puanlama sürecinde kullanılan puanlama yöntemleri temel olarak iki gruba ayrılmaktadır. Bunlar cevaplayıcıların madde ile yoklanan davranışa ilişkin kısmi bilgilerinin dikkate alınmadığı geleneksel puanlama yöntemi ve kısmi bilgilere de puan verilen çok kategorili/ağırlıklı puanlama yöntemleridir. Ağırlıklı puanlama yöntemleri arasında ise testte yer alan maddelerin ya da çoktan seçmeli maddeler için seçeneklerin ağırlıklandırılmasına dayalı olan, ağırlıklandırma işleminin uzman görüşlerinden ya da cevaplardan yola çıkılarak yapıldığı ya da ağırlıklandırma yapılırken testin yapısına ilişkin iç ya da dış ölçütlere öncelik verilen çeşitli yöntemler bulunmaktadır (Haladyna, 1994; Özgüven, 1998).

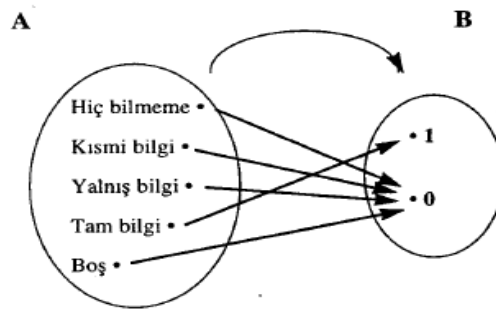
Geleneksel Puanlama Yöntemi

Geleneksel puanlama yöntemi, “1-0” yöntemiyle puanlama veya doğru yanıt sayısı metodu biçiminde isimlendirilen puanlama modelinde maddede cevap anahtarına uygun olarak doğru işaretleme yapan veya tam doğru cevabı oluşturan birey, o maddeyi doğru cevaplamış kabul edilerek madde için bireye bir puan verilmektedir. Maddeyi cevaplayamayan/boş bırakan, çeldiricilerden birini ya da birden fazla seçeneği işaretleyen bireyin ise cevabı yanlış sayılmakta böylece bu maddeden bireye puan verilmemektedir. Bu yöntemde esas olan sadece tam bilgiye sahip bireylerin doğru cevaplarına puan vermektir. Ancak kaçınılmaz bir biçimde tam bilgiye sahip

cevaplayıcılarla birlikte şansla/tahminle oluşturulan doğru yanıtlara da tam doğru yanıt puanı verilmektedir. Bilgisi tamamen yanlış olan, tahmin yoluyla cevaplama davranışında bulunup yanlış seçeneği işaretleyen ve aynı zamanda kısmi bilgisi olan bireylerin de tamamının yanıtları, yanlış yanıt olarak sayılmaktadır (Jaradat ve Tollefson, 1988). Burada kısmi bilgiye sahip olan cevaplayıcı ile tamamen yanlış bilgisi olan ya da maddede yoklanan bilgiye sahip olmayan cevaplayıcıların hepsinin sıfır puan olarak birbirinden ayırt edilemiyor olması puanlama işleminin adilliğine dair şüphe oluşturmaktadır.

Cevaplayıcının oluşturduğu cevapların yalnızca doğru veya yanlış biçiminde ayrıştırıldığı geleneksel puanlama modelinin puanlama işlemlerinde yaygın olarak kullanıldığı görülmektedir. Geleneksel puanlamanın yaygın olarak kullanılmasında, Binom teoremine dayanan güçlü bir istatistik temelini olmasının ayrıca puanlama işlemlerinde kolaylık ve pratiklik sağlamasının etkili olduğu düşünülmektedir. Ayrıca bu modeller test geliştirme ve madde analizi işlemlerinde, madde ve test istatistiklerinin hesaplanmasında kullanıcılarına birçok kolaylık sağlamaktadır (Lord ve Novick, 2008; Hambleton ve Swaminathan, 1989).

Yaygın olarak kullanılmasına rağmen geleneksel puanlama yönteminin çeşitli sebeplerle birçok eleştiriye maruz kaldığı da görülmektedir. En büyük eleştirilerden biri, bu yöntemin sadece maddeye doğru cevap verilip verilmemesine göre puan ataması, bireyin soru kökünde yoklanan davranışa ilişkin kısmi bilgisini dikkate almaması yönündedir. Kısmi bilgi, cevaplayıcının maddeyi çözümleme sürecine doğru başladıktan sonra çözüm adımlarından birinde hata yapması veya bu adımlardan birini atlaması nedeniyle doğru cevaba ulaşamaması durumunu belirtmektedir (Grunert, Raker, Murphy ve Holme, 2013). Geleneksel yöntemde cevaplayıcıların cevaplama durumlarına dayalı olarak verilen puanlar Şekil 1’de gösterilmiştir (Akkuş ve Baykul, 2001).



Şekil 1. Geleneksel Puanlama Yöntemi

Şekil 1’de görüldüğü gibi “1-0” puanlamada tüm cevaplayıcı-madde ilişkileri yalnızca iki kategorili modellerle incelenmektedir. Buna göre herhangi bir test maddesiyle ölçülen bilginin cevaplayıcıda sadece “var” ya da “yok” olduğu kabul edilmektedir. Bu durum hem test puanlarını yorumlamada yetersizliğe yol açmakta hem de yapılan ölçme işleminin kesinliğinin azalmasına sebep olmaktadır (Hutchinson, 1982). Hâlbuki bireylerin test maddesiyle ölçülmeye çalışılan bilgisi, tam bilme ile hiç bilmeme aralığında; tam, kısmi veya yanlış bilgi, ayrıca bilginin yokluğu gibi farklı seviyelerde olabilmektedir (García-Pérez ve Frary, 1989). Hatta cevaplayıcıların maddede yoklanan içeriğe ilişkin her şeyi bilmeleri ya da tam tersi olarak hiçbir şey bilmemeleri durumları ile nadiren karşılaşılmaktadır. Cevaplayıcılar sıklıkla bir kısmı doğru olan kısmi bilgilere sahip olduğundan bunların gösterilmesi gerekmektedir. Bu nedenle farklı bilgi seviyelerine sahip cevaplayıcıların gerçek bilgi düzeyini daha doğru belirleyebilmek amacıyla, cevaplayıcılara ait kısmi bilgilerin de puanlara yansıtılabildiği puanlama metodlarının geliştirilmesine ve kullanılmasına ihtiyaç duyulmuştur (Ben-Simon, Budescu ve Nevo, 1997).

Geleneksel puanlamaya ilişkin önemli başka bir sorun ise kısmi bilgi dikkate alınmadığında cevaplayıcının nerede öğrenme eksiği olduğunun bilinmemesidir (Lau, Lau, Hong ve Usop, 2011). Örneğin henüz öğrenme-öğretme süreci devam ederken yapılan bir değerlendirme türü olan izlemeye yönelik değerlendirmede öğrencinin öğrenme eksiklerinin tamamlanması, öğrenme sürecinin/gelişiminin izlenmesi amaçlanmaktadır (Andrade ve Cizek, 2010). Böyle bir değerlendirmede kullanılan ölçme araçlarındaki maddeler puanlanırken kısmi bilgilerin dikkate alınması değerlendirmenin amacına ulaşması bakımından özellikle önem taşımaktadır.

Geleneksel puanlama yöntemiyle ilgili eleştirilere sebep olan bir diğer durum, “1” puan verilen yani ilgili maddeyle ölçülen bilgiye tam olarak sahip olduğu düşünülen cevaplayıcının tahminle/şansla doğru seçeneği işaretleme durumunun kontrol edilememesidir. Buna göre doğru cevabı bilmedikleri ya da düşündükleri seçeneğin doğruluğundan emin olmadıkları için herhangi bir işaretleme yapmayan yani risk almaktan kaçınan cevaplayıcılar, tamamen şansla doğru seçeneği bulan cevaplayıcılara göre “0” puan verilerek cezalandırılmaktadır. Oysa cevaplayıcının sahip olduğu kısmi bilgi test puanlarının geçerliğini etkileyebilmektedir. Bunun nedeni çoktan seçmeli bir maddeye ilişkin eşit puan verilmiş olan cevaplayıcıların esasında eşit olmayan düzeylerde kısmi bilgisinin olabilmemesidir (Kurz, 1999). Özetle puanlamada geleneksel

yaklaşım, cevaplayıcıların çeldirici seçimlerini birbirinden ayırt etmediğinden kısmi bilgilerine ilişkin fikir vermemektedir (Diedenhofen ve Musch, 2015).

Cevaplayıcıların bir maddeye ilişkin farklı bilgi seviyeleri ile çeldirici seçimlerinin dışında, geleneksel puanlamada testte yer alan maddelerin nitelikleri de ayırt edilmemektedir. “1-0” puanlama yönteminde testte yer alan tüm maddelere eşit ağırlıklar verilerek maddelerin her birinin testle ölçülmek istenen özelliği eşit şekilde, eşit ağırlıkta sınıadığı düşünülmektedir. Oysa bazı maddeler diğerlerine göre psikometrik açıdan daha “güçlü” olabilmekte ya da bilenle bilmeyeni daha iyi ayırt edebilmektedir. Bu nedenle puanlamada bu maddelerin, ölçülen özellik bakımından bireyleri ayırt etmede daha az öneme sahip olduğuna karar verilen maddelere göre daha büyük bir katsayı ile çarpılması gerekmektedir (Cannon, 1994). Ancak geleneksel yaklaşımla tüm maddeler eşit şekilde puanlandığında yani maddelere eşit ağırlıklar verildiğinde her bir maddeye özgü bu nitelikler göz ardı edilmektedir.

Geleneksel puanlamada ayrıca, bir maddeyle ilgili çeldiricilerin hepsinin cevaplayıcılara aynı seviyede çekici geldiği kabul edilmektedir (Haladyna, 1990). Ancak tıpkı bireylerin performansları gibi seçenekler de doğruluk derecesine göre sıralanabilmektedir. Bazı seçenekler doğru cevaba çok yakın olan güçlü çeldiriciler olarak, bazıları ise doğru cevapla daha az ilişkili zayıf çeldiriciler olarak karşımıza çıkmaktadır. Bütün sınırlılıklardan hareketle, maddelerin her birinin cevaplayıcının performansına ilişkin farklı düzeyde bilgi içerdiğini ve maddelerdeki seçeneklerin farklı düzeylerde doğruluğa ya da çeldirici özelliğe sahip olduğunu varsayan, en önemlisi cevaplayıcının maddeyle yoklanan bilgisinin çeşitli seviyelerini dikkate alan çok kategorili puanlama veya ağırlıklı puanlama yöntemleri tanımlanmıştır (Frery, 1982; Haladyna, 1990; Stanley ve Wang, 1970).

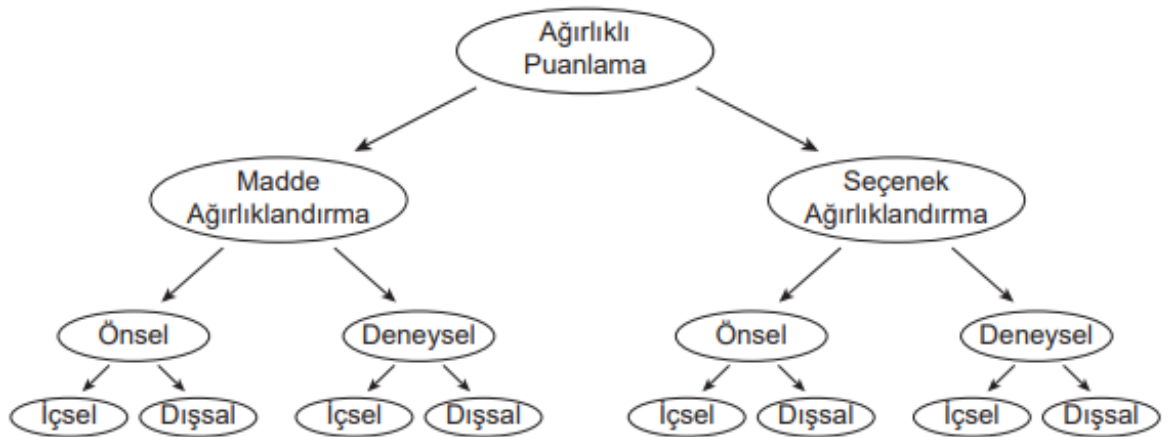
Çok Kategorili/Ağırlıklı Puanlama Yöntemleri

Bireyin bir maddeyle ölçülen davranışa ilişkin sahip olduğu tam bilginin yanında kısmi bilgisini de ölçmeyi amaçlayan, maddelerin ve seçeneklerin testi puanlamada farklı ağırlıklarda ele alınabildiğini iddia eden puanlama yöntemleri genel olarak çok kategorili/ağırlıklı puanlama yöntemleri olarak isimlendirilmektedir. Geleneksel yönteme göre bu yöntemlere ilişkin temel farklılık tam doğru olmayan yanıtlarda bulunan çeşitli seviyelerdeki kısmi bilginin dikkate alınarak bunlara kısmi puan

verilmesidir (Crehan ve Haladyna, 1994). Buradaki çıkış noktası, cevaplayıcının performansının ya da bir maddeye ilişkin bilgi düzeyinin orta düzeyde bulunabilir olması, başka bir deyişle doğru ile yanlış yanıtın arasında konumlanmasıdır (Abu-Sayf, 1979). Ağırlıklı puanlamada esas olan cevaplayıcıların performanslarını birbirlerinden farklı seviyelerde, daha hassas bir biçimde değerlendirmek/puanlamaktır.

Ağırlıklı puanlama yöntemleriyle puanlama yapılırsa, cevaplayıcıların maddelerden gerçek bilgi düzeylerine daha yakın puanlar alacağı böylece istenen test güvenilirliğinin daha az sayıda madde ile elde edilebileceği savunulmaktadır. Ayrıca kısmi bilgiye sahip öğrencilerin de ödüllendirilmesi, yani tam puan olmasa da bir miktar puan almaları gerektiği ifade edilmektedir (Wongwiwatthanakut, Popovich ve Bennett, 2000). Ancak burada teorik olarak cevaplayıcının tam doğru yanıtı bilmediği ve maddeyle ölçülmesi hedeflenen bilgiye kısmen sahip olması durumunda, çeldiricilerden birini sahip olduğu kısmi bilgi doğrultusunda mantıklı bir yolla seçtiği kabul edilmektedir. Yani tamamen şansla/tahminle cevaplama davranışı ayırt edilmemektedir. Bu durumda eğer çoktan seçmeli bir test için ağırlıklı puanlama yöntemleri kullanılıyorsa, ölçme sonuçlarının geçerliğinin ve güvenilirliğinin geleneksel puanlama yönteminin kullanıldığı duruma kıyasla artması için bir çeldiricinin gerçekten kısmi bilgiyle işaretlenmiş olması gerekmektedir (Waters, 1976).

Ağırlıklı puanlama yöntemlerine ilişkin ölçme alanında farklı yaklaşım ve sınıflandırmalar bulunmaktadır. Bu yöntemler genel olarak madde ağırlıklandırma ve seçenek ağırlıklandırma yöntemleri olarak iki ayrı grupta ele alınmıştır. Stanley ve Wang'a (1970) göre ağırlıklı puanlama yöntemlerinin üç boyutu Şekil 2 ile sunulmuştur.



Şekil 2. Ağırlıklı Puanlama Yöntemleri

Şekil 2'ye göre ağırlıklı puanlama yöntemleri birbirlerinden üç temel boyutta farklılaşabilmektedir. Birinci boyuttaki madde ağırlıklandırma, her bir maddenin bireyin performansının ölçülmesinde farklı düzeyde katkısı olduğunu esas almaktadır. Seçenek ağırlıklandırmada ise çeldiricilerin cevaplayıcılara ilişkin farklı düzeyde bilgi verdiği varsayılmaktadır. Maddelerin ağırlıklandırılmasına dayanan yöntemlerde maddeleri faktör analizi yoluyla ağırlıklandırma (Burt, 1950), tek tek maddeler yerine madde kümelerinin ağırlıklandırılması (Gulliksen, 1950), maddelerin çoklu regresyon yardımıyla ağırlıklandırılması, kısmi regresyon katsayılarına dayalı olarak ağırlıklandırılması (Guilford, 1954), test varyansına ya da madde varyansına dayalı olarak ağırlıklandırma (Dick, 1965), maddeleri ilgili oldukları konunun önemine göre ağırlıklandırma (Ghiselli, 1964), madde ayırıcılık indeksleri ile ağırlıklandırma (Birnbaum, 1968) gibi yöntemlerin önerildiği görülmektedir.

İkinci boyutta hem madde ağırlıklandırma hem de seçenek ağırlıklandırma yöntemleri önsel ya da deneysel olarak iki yolla yapılabilmektedir. Önsel ağırlıklandırmada maddelerin ya da seçeneklerin ağırlıklandırılması konu alanı uzmanlarının görüşlerine göre puanlamadan önce yapılmaktadır. Uzmanların görüşleri birbirinden oldukça farklı olabileceğinden bu yöntem subjektif olarak değerlendirilebilmektedir. Deneysel ağırlıklandırmada ise testteki maddelere ilişkin cevaplara dayalı olarak toplam test puanı ya da madde puanı-toplam puan korelasyonu gibi bir ölçüt belirlenerek ağırlıklandırma yapılmaktadır.

Üçüncü boyutta ise ağırlıklandırma yapılırken testin yapısına ilişkin iç ya da dış ölçütlerle olan ilişkilere öncelik verilmesine göre içsel ya da dışsal ağırlıklandırma ifade edilmektedir (Stanley ve Wang, 1970). Örneğin matematikle ilgili bir birime personel seçmede kullanılacak bir sınavda matematik alt testindeki maddelere, diğer alt testlerdeki maddelere göre daha fazla ağırlık verilirse dışsal ölçüte dayalı olarak madde ağırlıklandırma yapılmış olmaktadır. Bir testteki maddelere ilişkin ayırt edicilik düzeylerine bakılarak daha ayırt edici olan bir maddeye daha fazla ağırlık verilmesi ise içsel ölçüte göre madde ağırlıklandırma olarak ifade edilebilmektedir.

Frery (1989) tarafından önerilen ağırlıklı puanlama yöntemlerine ilişkin bir diğer sınıflamaya göre cevaplama sürecinde, cevaplayıcı yargısına dayalı yöntemler ya da doğrudan cevaplama yöntemleri kullanılmaktadır. Cevaplayıcı yargısına dayalı yöntemler kullanıldığında, bir cevaplayıcının kendi cevabına ilişkin yargısına dayalı olarak yapılan ağırlıklandırma ile puanlama yürütülmektedir. Örneğin cevaplayıcıdan

bir maddeye ilişkin doğru olduğunu düşündüğü seçeneği işaretlemesi ve yanına da işaretlediği seçeneğin doğruluğundan hangi oranda emin olduğunu belirtmesi istenir. Puanlama yapılırken işaretlediği seçenekle birlikte cevaplayıcının cevabından ne kadar emin olduğu beyanı da dikkate alınmaktadır. Bu yöntemle göre cevaplayıcılar, cevaplarının doğruluklarıyla ilgili emin olma düzeyleri farklı olduğundan, aynı seçeneği işaretlemelerine rağmen farklı puanlar alabilmektedir. Çoğunlukla sağlık alanındaki sınavlarda söz konusu metodun olduğu çalışmalara rastlanmaktadır. Bu sınavlarda insan yaşamına dair bilgiler yoklandığından cevaplayıcıların gerçekten tam bilgiye ulaşımadağını belirlemek gerekmektedir (Salehi, Sadighi ve Bagheri, 2015).

Doğrudan cevaplama yöntemlerinde ise cevaplayıcıdan yalnızca doğru olduğunu düşündüğü bir seçeneği işaretlemesi istenmektedir. Doğrudan cevaplama yöntemleri; doğru yanıt bulana kadar cevaplama, çoklu doğru seçenek yöntemi, seçenek ağırlıklandırmaya dayalı yöntemler ile MTK'ya dayalı yöntemler biçiminde gruplandırılabilir (Frary, 1989).

Doğru yanıt bulana kadar cevaplama yöntemi: Bu yöntemde cevaplayıcı, doğru cevaba ulaşmaya kadar, doğru olduğunu düşündüğü seçeneği işaretlemeye devam etmektedir. Puanlama yapılırken maddedeki seçenek sayısından cevaplayıcının yanlış işaretlediği seçenek sayısı çıkarılır. Yorucu olan ve uzun zaman alan bu yöntem avantaj olarak cevaplayıcıya hemen dönüt verilmesini sağlamaktadır.

Çoklu doğru seçenek yöntemi: Bir test maddesinin birden fazla doğru cevaba sahip olduğu durumda kullanılan bir yöntemdir. Bu yöntemde cevaplayıcı her bir seçeneğin doğru olup olmadığına karar vermek durumundadır. Puanlama doğru olarak seçilen seçenek sayısı esas alınarak yapılır. Burada her bir seçenek doğru-yanlış olarak puanlandığından hepsi ayrı birer doğru-yanlış tipi maddeye dönüştürülmüş gibi görünmektedir (Frary, 1989).

Seçenek ağırlıklandırma yöntemleri: Bu yöntemlerde seçeneklere farklı ağırlıklar verilerek maddeye ilişkin kısmi bilginin de ölçülmesi ve aynı zamanda kısmi bilginin önemine vurgu yapılması esastır (Lord ve Novick, 2008). Seçenek ağırlıklandırma yöntemleri ağırlıklandırmanın dayanağına göre ikiye ayrılmaktadır:

- 1) *Önsel/Mantıksal seçenek ağırlıklandırma:* Alan uzmanlarının kanısına dayalı olarak seçeneklerin test uygulanmadan önce ağırlıklandırılmasıdır. Bu yöntemde konu alanı uzmanları, maddelerin her birine ilişkin seçenekleri doğru yanıtla yakınlıklarına göre puanlamaktadır. Daha sonra uzmanlar tarafından

seeneklerin her birine verilen bu puanların ortalaması alınarak her bir seeneęe ilişkin aęırlıklandırmalar yapılmaktadır (Cross, Ross ve Geller, 1980).

- 2) *Deneysel seenek aęırlıklandırma*: Bu yntemde seeneklerin aęırlıklarına eřitli yollarla karar verilerek kısmi bilgi dikkate alınmaktadır. Seeneklerin her birine ilişkin aęırlıklar hesaplanırken sıklıkla cevaplayıcıların toplam test puanları lt olarak kullanılmaktadır (Frary, 1989). Bu yntemlere gre seeneklerin aęırlıklandırılması maddeye ilişkin cevaplar alındıktan sonra bu cevaplara dayalı olarak yapılır. Doęrusal ve doęrusal olmayan deneysel seenek aęırlıklandırma yntemleri bulunmaktadır. Doęrusal yntemlerde maddeye ilişkin cevaplarla seeneklerin aęırlıkları arasında doęrusal bir iliřki kurulmaktadır. Doęrusal olmayan yntemlerde ise seenekler aęırlıklandırılırken madde tepki kuramına dayalı doęrusal olmayan dnřmler kullanılmaktadır (Thissen, 1976).

Doęrusal deneysel seenek aęırlıklandırma yntemlerinin *ilki iki ynl ortalamar yntemidir*. Bu yntemde her bir seenek iin cevaplara dayalı olarak aęırlık belirlenir, bu aęırlıklar cevaplayıcıların toplam puanının hesaplanmasında kullanılır ve gvenirlięe ilişkin Cronbach alfa katsayısı ykselinceye kadar her bir cevaplayıcı iin bu uygulama tekrarlanır. Bu yntem yeterince iřlevsel olmaması nedeniyle bu haliyle kullanılmamaktadır.

Bir dięer doęrusal deneysel seenek aęırlıklandırma ynteminde *seenek-toplam korelasyonu* dikkate alınır. Bu ynteme gre seeneklerin aęırlıkları, maddelerin her birindeki cevap seeneklerinin her biriyle cevaplayıcının toplam puanı arasındaki nokta ift serili korelasyon hesaplanarak belirlenir.

Oran aęırlıkları ynteminde ise her bir seenek, o seeneęi iřaretleyen alt ve st gruptaki bireylerin puanları arasındaki yzde farklarına dayalı olarak aęırlıklandırılır. Deneysel seenek aęırlıklandırma yntemlerine ilişkin alıřmaların son yıllarda yoęunluk kazandıęı grlmektedir. Bununla birlikte bu arařtırmalar, deneysel seenek aęırlıklandırmanın testin gvenirlięinde iyileřme saęladıęını gstermiřtir (Diedenhofen ve Musch, 2015).

Madde tepki kuramına dayalı yntemler: Bu yntemlerin temelinde de seeneklerin aęırlıklandırılması yaklařımı yer almaktadır. Burada bir maddeyi cevaplayan tm bireylerin cevap daęılımlarının yetenek dzeylerine gre farklılařtıęı esas alınmaktadır. Seeneklerin aęırlıklandırılmasını esas alan bu modellerin kullanımı

doğruluk derecesine göre sıralanabilen cevaplar için uygundur. Tek boyutlu çok kategorili madde tepki kuramı (MTK) modellerinden bazıları; kademeli tepki modeli (Samejima, 1969), kısmi puan modeli (Masters, 1982) ve genelleştirilmiş kısmi puan modelidir (Muraki, 1992).

1) Kısmi Puan Modeli (Partial Credit Model)

Kısmi puan modeli, çözümü birden çok adımı gerektiren maddelerin analizi için geliştirilmiştir. Kısmi doğru yanıtların mümkün olduğu başarı testlerinde madde yanıtlarını puanlamada kullanılır. Bu modele göre bir maddenin çözümlenmesi sürecinde farklı aşamaların tamamlanması durumunda cevaplayıcıya kısmi puan verilmesi gerekmektedir. Bu durumun en tipik örneklerinden birden fazla adımda çözülen matematik alanından maddelerdir (Masters, 1982).

Modelin sıkça karşılaşılan örneklerinden biri yine matematiğe ilişkin bir maddeyle açıklanmaktadır. “ $(6/3) + 2$ işleminin sonucu kaçtır?” biçiminde verilen problemi tam doğru şekilde çözmek iki aşamayı tamamlamak gerekmektedir. Buna göre ilk aşamada parantezin içindeki bölme işlemi yapılır. İkinci aşamada ise bölme işleminden çıkan sonuçla dışarıdaki 2 sayısı toplanır. Bu problemi doğru çözmek için cevaplayıcıdan her iki aşamayı da yapması beklenmektedir. Buna göre ilk aşamayı tamamlayabilen cevaplayıcıya 1 puan, yani kısmi puan verilirken iki aşamayı da yapabilen cevaplayıcıya tam puan, yani 2 puan verilmektedir. Bunun dışında aşamalardan hiçbirini tamamlayamayan cevaplayıcıya puan verilmez. Bu durum bu problemin 0, 1 ve 2 olarak üç cevap kategorisine sahip olduğu anlamına gelmektedir. Burada dikkat çeken başka bir konu ilk aşamanın daha sonraki aşamadan zor olmasıdır. Ancak modelde cevaba ilişkin aşamaların kolaydan zora doğru sıralanması zorunluluğu yoktur. Aşamaların zorlukları değişiklik gösterebilmektedir. Modelde yalnızca adımların birbiri ardından geldiği varsayılmaktadır (de Ayala, 1993).

Kısmi puan modeli, Rasch modelin çok kategorili maddelere yönelik uzantısı özelliğindedir. Çok kategorili puanlanan maddelere ilişkin birçok MTK modeli bulunmasına rağmen, kısmi puan modeli Rasch modellerinin özelliklerini taşıdığından en yaygın olarak kullanılan modellerden biridir. Rasch modeli ve kısmi puan modeli yeni ölçme araçları geliştirmek ya da ölçme araçlarının psikometrik özelliklerini değerlendirmek için sıkça kullanılmaktadır (Tennant, McKenna ve Hagell, 2004). Kısmi puan modeline göre θ yetenek düzeyindeki cevaplayıcının j maddesine x cevabını verme olasılığı matematiksel olarak eşitlik (1)’de gösterilmiştir.

b_{jh} : (h-1) adımından (seçeneğinden) h adımına geçişteki adım güçlüğü parametresi olmak üzere,

$$P(x_j|\theta, b_{jh}) = \frac{e^{\sum_{h=0}^{x_j}(\theta - b_{jh})}}{\sum_{k=0}^{m_j} e^{\sum_{h=0}^k(\theta - b_{jh})}} \quad x = 0,1,2 \dots m \quad (1)$$

Kısmi puan modelinde bireyin yetenek parametresi θ ve madde güçlüğü parametresi β bulunmaktadır. β parametreleri, Masters (1982) tarafından “adım güçlüğü” biçiminde tanımlanmaktadır. Bunun sebebi, cevaplayıcının bir adımı başarılı bir şekilde gerçekleştirip sıradaki adıma başlamasıdır. Adım güçlükleri, çeşitli kaynaklarda kategori kesişim parametresi veya geçiş konumu parametresi biçiminde verilmektedir. Kısmi puan modelinde adım güçlüğü parametresinin sayısı madde kategori sayısının bir eksiği kadardır. Örnek olarak dört kategorili puanlanan bir maddeye ilişkin üç tane adım güçlüğü parametresi bulunmaktadır (Garrett, 2009).

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

Kısmi puan modelinin temeli Rasch modeline dayandığı için, testte bulunan tüm maddelerin eşit ayırt edicilik parametresine sahip olduğunu varsaymaktadır. Muraki (1992), kısmi puan modelini genelleştirip maddelerin birbirlerine eşit olmayan ayırt edicilikte olabilmesine izin veren genelleştirilmiş kısmi puan modelini sunmuştur. Bu modele dayalı olarak cevaplayıcının j maddesinde x yanıtını verme olasılığı eşitlik (2)’de verilmiştir.

α_j : j maddesinin madde ayırt edicilik parametresi olmak üzere,

$$P(x_j|\theta, \alpha_j, b_{jh}) = \frac{e^{\sum_{h=0}^{x_j}(\theta - b_{jh})}}{\sum_{k=0}^{m_j} e^{\sum_{h=0}^k(\theta - b_{jh})}} \quad x = 0,1,2 \dots m \quad (2)$$

Kısmi Puan Modeli’nde olduğu gibi, genelleştirilmiş kısmi puan modelinde seçenek karakteristik eğrilerinin kesişim noktaları adım güçlüğü parametrelerini oluşturmaktadır. Ayrıca yine benzer şekilde seçenek sayısının bir eksiği kadar adım güçlüğü parametresi bulunmaktadır.

Genelleştirilmiş kısmi puan modelinde bir maddenin madde ayırt edicilik parametresi 1.00 olduğunda, seçenek karakteristik eğrileri kısmi puan modeliyle aynı olmaktadır. Yani madde ayırt edicilik parametresi 1.00 olduğunda; genelleştirilmiş kısmi puan modeli, kısmi puan modeli ile aynı sonucu vermektedir.

3) Derecelendirilmiş/Kademeli Cevap Modeli – Kademeli Tepki Modeli (Graded Response Model)

Samejima (1969) tarafından önerilen derecelendirilmiş cevap modelinde seçenekler arasında birikimli karşılaştırmalar yapılmaktadır. Yani bu modelde diğer iki modelde olduğu gibi kategori 0 ile kategori 1 ya da kategori 1 ile kategori 2 karşılaştırılmaz. Bunun yerine kategori 0'dan 2'ye kadar olanlar birlikte dikkate alınmaktadır. Bu doğrultuda cevaplayıcının j maddesine x_j veya x_j 'den daha büyük bir kategoriyi tamamlama olasılığı eşitlik (3)'te ifade edilmektedir.

$p_{x_j}(\theta)$: j maddesine x_j veya x_j 'den daha büyük bir kategorinin seçilme olasılığı olmak üzere,

$$p_{x_j}(\theta) = \frac{e^{a_j(\theta - b_{x_j})}}{1 + e^{a_j(\theta - b_{x_j})}} \quad j = 1, 2, \dots, n \quad (3)$$

Bu eşitlikte b_{x_j} x_j seçeneği için seçenek eşik değerini ve a_j j maddesine ait ayırt edicilik parametresini ifade etmektedir. Derecelendirilmiş cevap modelinde cevaplayıcının bir maddeye x_j 'den düşük veya x_j ya da daha büyük bir yanıt verme olasılığı hesaplanmaktadır. Dolayısıyla derecelendirilmiş cevap modelinin uygulanacağı maddeye ait seçeneklerin, ölçülen özellik bakımından sıralı olmasına ihtiyaç duyulmaktadır. Bir maddede x_j ya da x_j 'den büyük olan kategorilerden birinin seçilmesinin olasılığı bilinirken, sadece x_j cevabının görülme olasılığını hesaplamak için, bireyin x_j+1 ve daha büyük bir yanıt verme olasılığından, x_j veya daha büyük bir yanıt verme olasılığının çıkarılması gerekmektedir.

Frery (1989) tarafından ileri sürülen ağırlıklı puanlama yöntemlerine ilişkin sınıflama ve puanlama yöntemleri dışında, alanyazın incelemesinde farklı ağırlıklı puanlama yöntemlerinin önerildiği çalışmalara da rastlanmıştır. Bu çalışmalardan birinde (Haladyna ve Sympson, 1988) “çoklu ağırlıklandırma” olarak isimlendirilen bir puanlama metodu önerilmiştir.

Çoklu ağırlıklandırma yöntemi: Haladyna ve Sympson (1988) tarafından önerilen bu yöntem, yetenek ve başarı testlerindeki çoktan seçmeli maddeler için zor soruların doğru yanıtları için kolay soruların doğru yanıtlarına göre daha fazla puan verilmesini ve aynı zamanda seçeneklerin de ağırlıklandırılmasını esas almaktadır. Test maddelerine ilişkin yanıtlar yalnızca "doğru" veya "yanlış" olarak sınıflandırıldığında cevaplayıcıların yetenek/beceri düzeyleriyle ilgili bilgi kaybı olmaktadır. Bu nedenle doğru cevaplanan soruların zorluk derecesi de dikkate alınarak cevaplayıcılara ilişkin ek bilgiler elde edilebileceği öne sürülmektedir. Bu yöntemde olduğu gibi seçenek puanı, o

seçeneği seçen adayların ortalama puanı ve cevaplayıcının toplam test puanı seçtiği seçenek puanlarının ortalaması olduğunda, madde-toplam korelasyonları ve alfa katsayısı maksimize edilmektedir (DeMars, 2008). Çoklu ağırlıklandırma, klasik optimal ölçeklendirme veya ikili ölçeklendirme (McDonald, 1983; Warrens, de Gruijter ve Heiser, 2007) ve çoklu puanlama (Haladyna, 1990) gibi isimlerle de karşımıza çıkabilmektedir.

Çoklu ağırlıklandırmada, madde-yanıt kategorilerine atanan puanlama ağırlıkları, çoklu ağırlıklar olarak tanımlanmaktadır. Bir puanlama sürecinde çoklu ağırlıkları üretmek için takip edilmesi gereken prosedür aşağıdaki gibidir:

1. Cevaplayıcıların her birine, testte yer alan maddelerdeki doğru oranına eşit bir geçici puan atanır.

2. Geçici puanların bir yüzdelik sıralama puanına dönüştürülmesi gerekir. Bunun için cevaplayıcının bir önceki adımda hesaplanan geçici puanına eşit ve daha küçük geçici puana sahip cevaplayıcıların sayısı, tüm cevaplayıcıların sayısına oranlanır ve 100 ile çarpılır.

3. Her bir madde için, her bir seçeneği/olası cevap kategorisini seçen cevaplayıcıların bir önceki adımda hesaplanan yüzdelik sıralama puanlarının ortalaması hesaplanır. Bu aşamada eğer doğru seçenek için hesaplanan yüzdelik sıralama puanlarının ortalaması, o madde için tüm cevaplayıcıların yüzdelik sıralama puanlarının ortalamasından küçükse ilgili maddenin analizden çıkarılması gerekir.

4. Tüm maddeler ve tüm seçenekler/yanıt kategorileri için çoklu ağırlıklar hesaplanır. Bunun için;

a) Tam doğru cevap kategorisi için, bu cevabı veren cevaplayıcıların yüzdelik sıralama puanlarının ortalaması en yakın tamsayıya yuvarlanır.

b) Yanlış cevap kategorisi 100 veya daha fazla cevaplayıcı tarafından işaretlendiyse, bu 100 veya daha fazla cevaplayıcının ilgili seçenek için hesaplanan yüzdelik sıralama puanlarının ortalaması en yakın tamsayıya yuvarlanır.

c) Yanlış cevap kategorisi 100'den az cevaplayıcı tarafından işaretlendiyse, çoklu ağırlık aşağıdaki eşitliğe göre hesaplanır. W_{ij} i. maddenin j. cevap kategorisinin çoklu ağırlığı olmak üzere,

$$W_{ij} = \overline{R_{i(w)}} + \left[\frac{N_{ij}}{100} \right]^{1/2} (\overline{R_{i(j)}} - \overline{R_{i(w)}})$$

eşitliğinden hesaplanan değer en yakın tamsayıya yuvarlanır. Eşitlikteki $\overline{R_{l(w)}}$ i maddesinde herhangi bir yanlış cevabı seçen cevaplayıcıların yüzdelik sıralama puanlarının ortalamasıdır. $\overline{R_{l(j)}}$ j cevap kategorisini seçen cevaplayıcıların yüzdelik sıralama puanlarının ortalamasıdır. N_{ij} ise j cevap kategorisini seçen cevaplayıcıların sayısıdır.

5. Çoklu ağırlıklar 0 ile 100 arasında herhangi bir tamsayı değeri alabilir. Çoklu ağırlıklar hesaplanırken bir yanlış yanıtın çoklu ağırlığı, doğru yanıtın çoklu ağırlığına eşit veya doğru yanıtın çoklu ağırlığından büyük çıkarsa, doğru yanıtın çoklu ağırlığının bir eksiği ilgili yanlış yanıtın çoklu ağırlığı olarak belirlenir. Böylece çoklu ağırlıklandırma yapılırken bir cevaplayıcı yanlış bir cevap için asla doğru bir cevaptan daha fazla kredi alamaz.

6. Çoklu ağırlıklara göre cevaplayıcıların ağırlıklandırılmış toplam puanları hesaplanır. Bunun için bir cevaplayıcının seçtiği cevap kategorilerinin çoklu ağırlıklarının ortalaması alınır.

Ölçme alanyazını incelendiğinde farklı puanlama yöntemlerinin kullanıldığı ve sonuçların çeşitli yönlerden incelendiği çalışmalarda genellikle testlerin psikometrik özelliklerinin karşılaştırıldığı, bunun yanında cevaplayıcıların puanlama yaklaşımına ilişkin görüşlerinin de incelendiği görülmüştür. Bu çalışmalardan birinde kısmi puan modeliyle yapılan bir ağırlıklı puanlama yönteminin tanıtılması, bu yöntemin bir lisedeki matematik sınavlarının puanlamasında kullanılmasının ardından öğrenci görüşlerinin alınması amaçlanmıştır. Çalışmada elde edilen sonuçlara göre cevaplayıcılar, ağırlıklı puanlamanın geleneksel yöntemle puanlamaya kıyasla daha doğru sonuçlar ortaya koyduğunu, puanlama işlemlerinde ağırlıklandırma yapıldığında kaygılarının azaldığını, ayrıca kısmi bilgiye puan verilmesinin hümanist bir yaklaşım olduğunu belirtmişlerdir. Ayrıca geleneksel puanlamanın cevaplayıcılarda gereksiz bir strese neden olduğu da bu araştırmada görüşü sorulan bireyler tarafından belirtilmiştir (Cherkas ve Roitberg, 1993).

Çeşitli puanlama yöntemlerinin kullanılması ya da önerilmesiyle birlikte testlerin psikometrik özelliklerinin incelendiği araştırmaların özellikle 1950'li yıllardan itibaren yoğunlaştığı görülmektedir. Araştırmacıların bu konuya dikkatini çeken ilk çalışmalardan biri ise Guilford, Lovell ve Williams (1942) tarafından gerçekleştirilmiştir. Bu araştırmada ağırlıklı puanlamanın ağırlıklandırılmamış

puanlamaya göre daha geçerli ve güvenilir sonuçlar verip vermediğini ortaya koymak amaçlanmıştır. Çalışmada İnsan Davranışı Sorunları konulu bir kursun final sınavından elde edilen cevaplar kullanılmıştır. Araştırma sonucunda ele alınan üç farklı test uzunluğunun (20 madde, 50 madde ve 100 madde) tamamında ağırlıklı puanlamayla elde edilen puanların geçerliği ve güvenilirliği, geleneksel yöntemle elde edilen sonuçlara göre yüksek çıkmıştır. Fakat yine de bu sonuçlar araştırmacılar için tatmin edici olmayan bir fark olarak görülmüştür.

Phillips (1943) ise, bir zekâ testini dördüncü, beşinci ve altıncı sınıf seviyelerinden toplamda 204 öğrenciye uygulayarak elde ettiği ağırlıklı ve ağırlıklandırılmamış puanları karşılaştırmıştır. Çalışmanın sonuçlarına göre ağırlıklı puanlamanın daha iyi kestirim yaptığını ve kestirim hatalarının %4-5 oranında azaldığını belirlemiştir. Ancak bu sonucun harcanan çabaya değer olmadığını, aynı katkının testte daha fazla sayıda madde kullanılarak ya da daha büyük bir madde havuzundan sadece en kaliteli maddelerin seçilmesiyle elde edilebileceğini belirtmiştir. Çalışmanın yapıldığı dönemin teknolojik sınırlılıklarından dolayı tüm hesaplamalar elle yapıldığı için, çalışmada ayrıca ağırlıklandırılmamış puanlamanın zamandan tasarruf sağladığı ve puanların hesaplanmasında daha az hata olasılığı olduğu belirtilmiştir.

Echternacht (1976) araştırmasında önsel ve deneysel seçenek ağırlıklandırma yöntemlerinin kullanıldığı durumlarda güvenilirlik ve geçerliğin nasıl değiştiğini incelemeyi amaçlamıştır. Bu amaç doğrultusunda çalışmada geleneksel puanlama yöntemleri, düzeltme formülüyle birlikte geleneksel puanlama yöntemi, önsel seçenek ağırlıklandırma ve içsel bir kriterle yapılan deneysel seçenek ağırlıklandırma yöntemi kullanılmıştır. Puanlama yöntemlerinin birbirlerine göre etkililiğini belirlemek için, Lisansüstü Kayıt Sınavları (Graduate Record Examinations, GRE) kapsamında Haziran 1972 döneminde yetenek testinin uygulandığı kişilerden rasgele gruplar oluşturulmuştur. Her bir gruptan elde edilen yanıtlar, ele alınan puanlama yöntemlerinin her biriyle sırayla puanlanarak güvenilirlik ve geçerliğe ilişkin bulgular elde edilmiştir. Çalışmada, deneysel seçenek ağırlıklandırma yönteminin hem güvenilirlik hem de geçerlik açısından diğer yöntemlere göre üstün olduğu sonucu elde edilmiştir.

Young (1991), MTK kapsamında kademeli tepki modeliyle KTK kapsamında klasik yığılmalı puanlama modelini kız ve erkek öğrencilerin yeteneklerini tahmin etmesi açısından karşılaştırmıştır. Bu araştırmanın sonuçlarında, kız ve erkek öğrenciler için MTK'ya dayanan puanlama sisteminin diğer puanlama yaklaşımına göre yetenek kestiriminde daha iyi yordayıcı olduğu görülmüştür.

Wongwiwatthanakul, Popovich ve Bennett (2000) arařtırmalarında deęiřen derecelerde doęrulukla (yani, en doęru, biraz doęru) yanıtlanabilen oktan semeli sorularda ğrencilerin kısmi bilgilerini de dllendirmeyi bylece ğrenci bilgisini daha doęru biimde lmeyi amalamıřlardır. Arařtırmanın alıřma grubunu Purdue niversitesi, Eczacılık ve Eczacılık Bilimleri Okulunda Reetesiz İlalar (PHPR 470) dersine kayıtlı yz elli ğrenci oluřturmaktadır. Arařtırmanın amacı doęrultusunda alıřmada ok kategorili ve iki kategorili puanlama yntemleri kullanarak madde glę, madde ayırt edicilięi, gvenirlik bakımından konu alanları (analjezik ve mřhil) ve biliřsel alanların her birinde karřılařtırmalar yapılmıřtır. Elde edilen bulgulara gre ok kategorili puanlama yaklařımı madde glęn arttırmaktadır. Ayrıca ok kategorili puanlamada geleneksel yaklařıma kıyasla cevaplayıcılara iliřkin daha ok bilgi edinilmektedir.

Kinsey (2003), arařtırmasında testteki madde sayısı ve puanlamanın eřidi (iki ve ok kategorili puanlama) deęiřkenlerinin yetenek kestirimi ile test bilgi dzeyine etkisini incelemeyi amalamıřtır. alıřmada 33 karma test iin 1000 kiřilik simlatif veri kullanılmıřtır. Elde edilen sonulara gre, testin uzunluęu ve testteki ok kategorili puanlama ynteminin kullanıldıęı maddelerin oranının test aracılıęıyla alınan bireylere iliřkin bilgi zerinde anlamlı bir etkisi olduęu grlmřtr. Testteki maddelerin hepsinin ok kategorili puanlama yntemiyle puanlanması durumunda testten elde edilen bilginin en yksek dzeye ulařtıęı belirlenmiřtir. Ayrıca bu bulguya zellikle dřk yetenek dzeyindeki cevaplayıcı gruplarında ulařılmıřtır.

DeMars (2008), arařtırmasında madde tepki kuramına dayalı olarak 3PLM ile sınıflamalı tepki modelini ve madde tepki kuramına dayalı olmayan puanlama yntemlerinden “1-0” puanlama ve oklu aęırlıklandırma yntemlerini kullanmıřtır. Bu arařtırmada puanlama yntemlerinin yanlılık, standart hata ve gvenilirlik aısından karřılařtırılması amalanmıřtır. 20 maddelik bir oktan semeli testten elde edilen parametrelere dayalı olarak simle edilen verinin kullanıldıęı arařtırmada, nceki arařtırmalarla tutarlı olarak ok kategorili puanlama yntemleri daha byk gvenirlik kestirimleri saęlamıřtır. oklu aęırlıklandırma yntemiyle elde edilen puanların en az sınıflamalı tepki modelinin tahminleri kadar gvenilir olduęu belirlenmiřtir. Ayrıca ok kategorili modellerin kullanılmasından kaynaklanan gvenirlikteki kk artıřların, adaylara iliřkin geti-kaldı gibi nemli kararların verildięi bazı baęlamlarda anlamlı olacak kadar byk olabileceęi vurgulanmıřtır.

Yurdugül (2010) tarafından yapılan araştırmada KTK kapsamında ağırlıksız, ağırlıklandırılmış madde puanlama yöntemi, MTK kapsamında 1PLM ve 2PLM'den kestirilen puanlama metodu ile konjenerik test kuramına dayalı puanlama yöntemlerinin karşılaştırılması amaçlanmıştır. Çalışmada 2001 senesinde liselere geçişi düzenlemek amacıyla gerçekleştirilen merkezi sınavdaki 25 maddeden oluşan fen bilgisi testine 10000 öğrenci tarafından verilen cevaplar kullanılmıştır. Puanlama yöntemlerinden elde edilen sonuçlar güvenirlik, tek boyutluluk ayrıca model-veri uyumu bakımından karşılaştırılmıştır. Çalışmada madde puanlama türleri arasında madde tepki kuramı ve konjenerik test kuramı kapsamındaki puanlama türleri en iyi sonuçları vermiştir. Bu araştırmada ayrıca düzeltme formülünün kullanıldığı ve ağırlıklandırma yapılmadığı durumda elde edilen puanlara göre oluşan başarı sıralamalarının diğer yöntemler kullanıldığında oluşan başarı sıralamaları ile en düşük uyumu gösterdiği görülmüştür.

Alanyazın incelendiğinde ağırlıklı puanlama yöntemlerinin kullanıldığı durumlarda sadece maddelerin, sadece seçeneklerin veya hem maddelerin hem de seçeneklerin ağırlıklandırıldığı araştırmaların gerçekleştirildiği görülmüştür. Yalnızca maddelerin ağırlıklandırılması yoluyla puanlama yapılan bir çalışmada Gözen (2002), kısa yanıtlı ve çoktan seçmeli test maddelerinin geleneksel yaklaşımla ve ağırlıklı puanlama yöntemleriyle puanlanması durumunda testin psikometrik özelliklerinin incelenmesini amaçlamıştır. Araştırmada kullanılan testler 2001-2002 öğretim yılı bahar döneminde Ankara ve Hacettepe Üniversitesi Sınıf Öğretmenliği ve Rehberlik ve Psikolojik Danışmanlık bölümlerinde öğrenime devam eden 316 öğrenciye uygulanmıştır. Çalışmadaki puanlama sürecinde geleneksel puanlama, madde güçlük indekslerine dayalı ağırlıklı puanlama ve hem madde güçlük indekslerine hem de madde ayıricılık gücü indekslerine bağlı ağırlıklı puanlama yöntemleri kullanılmıştır. Elde edilen sonuçlara göre, üç farklı puanlama yönteminin kullanıldığı durumların her birinde test güvenirliği ve geçerliğine ilişkin benzer sonuçlar elde edilmiştir.

Quist (2004), araştırmasında bir personel seçme ölçeğini alan 875 bireyden elde edilen verilerin yanında farklı koşullar altında ürettiği simülasyon verilerini kullanarak farklı puanlama yöntemlerinin kullanıldığı durumlarda testin geçerliğinin nasıl değiştiğini incelemiştir. Çalışmada geleneksel puanlama yönteminin yanında maddelerin ağırlıklandırılmasına dayalı ağırlıklı puanlama yöntemi kullanılmıştır. Bu çalışmanın sonuçlarına göre madde ağırlıklandırma yöntemiyle puanlama, puanların özellikle ölçüte dayalı geçerliğini ve yordama geçerliğini artırmada önemli bir katkı sağlamaktadır.

Kılıç ve Doğan (2019) tarafından yapılan araştırmada ise maddelerin ağırlıklandırılmasına ilişkin yeni bir yöntem önerilmiştir. Bu doğrultuda çalışmada araştırmacılar tarafından geliştirilen madde ağırlıklandırma yönteminin testin yapı geçerliğine ve güvenilirliğine etkisinin incelenmesi amaçlanmıştır. Simülasyon çalışması olarak yürütülen araştırmada örneklem büyüklüğü, madde sayısı ve ortalama faktör yükü koşulları ele alınarak toplamda 18 simülasyon koşulu incelenmiştir. Önerilen madde ağırlıklandırma yönteminin kullanılması durumunda geleneksel yöntemin kullanıldığı duruma kıyasla güvenirlik katsayısının daha yüksek olduğu gözlenmiştir. Çalışmada ayrıca kullanılan madde ağırlıklandırmanın testin yapı geçerliğine katkı sağladığı da belirlenmiştir.

Ağırlıklı puanlama yöntemlerinin kullanıldığı araştırmaların büyük bir bölümünde çoktan seçmeli testlerle birlikte seçenek ağırlıklandırma yöntemlerinin kullanıldığı görülmüştür. Davis ve Fifer (1959) çalışmalarında lise düzeyindeki öğrencilere uygulanan çoktan seçmeli bir aritmetik-muhakeme testinin seçeneklerine ağırlık verilerek elde edilen puanları ele almıştır. Bu çalışmada aritmetik-muhakeme testinde kullanılan ağırlıklı puanlama yönteminin elde edilen puanların güvenilirliği ve geçerliliği üzerindeki etkisini belirlemek amaçlanmıştır. Araştırmanın sonuçlarına göre ağırlıklı puanlama yöntemiyle puanlamanın test güvenirliğinde istatistiksel ve pratik öneme sahip olacak kadar büyük bir artışla sonuçlandığı belirlenmiştir.

Cross, Ross ve Geller (1980), üniversitedeki ders notların belirlenmesinde çoktan seçmeli bir testle seçenek ağırlıklandırma kullanılmasının uygulanabilirliğini araştırmıştır. Araştırmada seçenek ağırlıklandırma, hem deneysel hem de uzman yargısına dayalı önsel ağırlıklandırma prosedürleri dikkate alınarak gerçekleştirilmiştir. Bu çalışmada elde edilen güvenirlik ve geçerlik kanıtlarına göre, her iki seçenek ağırlıklandırma puanlama yönteminin de geleneksel puanlamaya göre testin psikometrik özellikleri açısından ciddi bir iyileşme sağlamadığı ifade edilmiştir. Bu çalışmada ayrıca öğrencilerin puanlama yöntemlerine ilişkin görüşleri anket yoluyla alınmış ve öğrencilerin büyük bir çoğunluğunun ağırlıklı puanlamayı tercih ettikleri belirtilmiştir.

Haladyna ve Simpson (1988) ise çalışmalarında yeni bir puanlama modeli önermişlerdir. Bu puanlama modeline göre her madde için farklı bir puan ağırlığı hesaplanmaktadır. Aynı zamanda seçeneklerin de ağırlıklandırıldığı çalışmada, çoklu ağırlıklandırma olarak isimlendirilen bu yöntemin test puanlarının güvenilirliğine ve kapsam geçerliğine etkisi incelenmiştir. Elde edilen sonuçlara göre önerilen puanlama

yönteminde geleneksel yöntemle göre puanların güvenilirliğinin ve kapsam geçerliğinin arttığı belirlenmiştir.

Ağırlıklı puanlamanın iki kategorili puanlamaya göre güvenilirlik ve geçti/kaldı kararlarının tutarlılığı açısından daha etkili olup olmadığının incelendiği başka bir çalışmada (Haladyna, 1990), çok kategorili puanlama yoluyla elde edilen puanlara ilişkin güvenilirliklerin diğerlerine kıyasla anlamlı bir biçimde daha yüksek olduğu belirlenmiştir. Bu çalışmanın en önemli bulgusunun test puanlarının dağılımına ilişkin olduğu düşünülmektedir. Buna göre geleneksel puanlamada test puanlarının dağılımı testte yer alan madde sayısı ile sınırlı kalmaktadır. Örneğin on maddelik bir test iki kategorili puanlama yöntemiyle puanlandığında sıfır ile on puan arasında, yani toplam 11 tane olası test puanı ortaya çıkabilir. Ancak aynı test seçeneklerin ağırlıklandırılmasına dayalı olarak ağırlıklı puanlama ile puanlandığında ise her bir seçeneğe ilişkin puan dikkate alınacağından 4^{10} tane farklı puan hesaplanabilir. Bu durumda ağırlıklı puanlamada testi alan bireyler arasında daha hassas bir ayırım yapılabilmektedir.

Crehan ve Haladyna (1994) 1-0 puanlama ile çok kategorili puanlama yöntemlerinin, geçti/kaldı kararlarının alındığı test programlarına ilişkin etkililiğini güvenilirlik-geçerlik bakımından incelemiştir. Araştırmanın verileri Ulusal Eczacılar Birliği Ruhsatlandırma Sınavının (NABPLEX) uygulandığı 3000 kişinin yanıtlarından elde edilmiştir. Çalışmada seçenek ağırlıklandırmanın puan dağılımının iç tutarlılığını artırma etkisine sahip olduğu görülmüştür. Ayrıca puan ortalamaları arasında farklar görülmüş ve Haladyna (1990)'nın çalışmasındaki sonuçlarla uyumlu bir şekilde çok kategorili puanlamada puan hassasiyetinde bir artış olduğu belirtilmiştir. Son olarak bu araştırmada çok kategorili puanlama için çeldiricilerin kalitesinin çok önemli olduğu vurgusu yapılmıştır.

Akkuş (2000) ise çalışmasında çoktan seçmeli test maddelerinin puanlanmasında geleneksel yöntem, uzman yargısına dayalı puanlama, Zinger Z_1 ve Z_2 puanlama yöntemlerini kullanmıştır. Araştırmada 1999-2000 yılında sekizinci, dokuzuncu ve onuncu sınıfta öğrenim gören 88 öğrenciye uygulanan bir yetenek testinden elde edilen cevaplar belirlenen yöntemlerle puanlanmıştır. Araştırma sonucunda bu puanlama yöntemlerinin her birine ilişkin hesaplanan madde ayırt edicilik indekslerinin benzer olduğu görülmüştür. Ayrıca her bir durumda kestirilen test güvenilirlik katsayılarının yeterince yüksek olduğu ifade edilmiştir. Fakat bu katsayıların kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaşmadığı belirtilmiştir. Benzer şekilde test

puanlarının geçerliğinin de çalışmada ele alınan puanlama durumlarında farklılaşmadığı görülmüştür.

Özdemir'in (2002) çalışmasında çoktan seçmeli bir testte "1-0" ve ağırlıklı puanlama yöntemlerinin kullanılmasının test puanlarının güvenilirlik ve geçerliğine etkisi KTK ve örtük özellikler teorisine dayalı olarak ayrı ayrı araştırılmıştır. Araştırmanın verisi, 2001-2002 yılında dördüncü, beşinci, altıncı ve yedinci sınıf düzeyinde öğrenim gören 1608 öğrenciye uygulanan bir çoktan seçmeli test yardımıyla toplanmıştır. Bu çalışmada, geleneksel puanlama yöntemiyle birlikte, önsel seçenek ağırlıklandırma yöntemi ayrıca örtük özellikler teorisine ilişkin derecelendirilmiş yanıt modeli kullanılmıştır. Klasik test teorisinde ağırlıklandırılmış puanlara dayalı olarak hesaplanan alfa güvenilirlik katsayısının, geleneksel puanlamayla hesaplanan KR-20 güvenilirlik katsayısından yüksek olduğu gözlenmiştir. Ayrıca iki kategorili puanlamada Lord'un katsayısının KR-20'den yüksek olduğu görülmüştür. Ağırlıklı puanlamayla ilgili olarak örtük özellikler kuramına dayalı olarak hesaplanmış olan marjinal güvenilirlik katsayısının KTK kapsamında hesaplanan Cronbach alfa güvenilirlik katsayısına göre daha büyük olduğu belirlenmiştir.

Saygı (2004) ise geleneksel yöntemle ve kısmi bilgiye de puan verilmesini amaçlayan bir puanlama yöntemi olarak önerilen bilgi referanslı puanlama yöntemini çalışmada kullanmıştır. Araştırmada bu yöntemin kullanılmasıyla puanlanan çoktan seçmeli bir testin madde ve test özellikleri arasındaki ilişkiler incelenmiştir. Araştırmada 2003-2004 öğretim yılında sekizinci sınıfta öğrenim gören 217 kişilik bir öğrenci grubuna çoktan seçmeli bir test uygulanarak elde edilen cevaplar puanlanmıştır. Bu çalışmanın sonuçlarına göre her iki puanlama yöntemiyle elde edilen puanların güvenilirlik ve geçerlikleri arasında anlamlı bir fark görülememiştir. Buna göre çoktan seçmeli bir testin bilgi referanslı puanlama yöntemiyle puanlanmasının madde ve test özelliklerine bir katkı sağlamadığı raporlanmıştır.

Gözen Çıtak (2007)'in çalışmasında çoktan seçmeli bir testteki maddelerin seçeneklerinin geleneksel puanlama, uzman yargısına dayalı ve deneysel seçenek ağırlıklandırma yöntemleriyle puanlandığında elde edilen güvenilirlik ve geçerliğin KTK ve MTK kapsamında hangi düzeyde etkilendiği incelenmiştir. Araştırmanın çalışma grubunu 2006-2007 yılı bahar döneminde Hacettepe ve Gazi Üniversitesi'nin çeşitli bölümlerinde öğrenimine devam eden 1593 öğrenci oluşturmaktadır. Çalışmaya ilişkin bulgulara göre MTK'da "1-0" puanlamada daha doğru yetenek parametresi kestirimi yapılabildiği ve geçerliği daha yüksek olan sonuçlar elde edilebildiği görülmüştür.

KTK'ya dayalı olan incelemelerde ise deneysel ağırlıklandırma kullanıldığında güvenilirlik bakımından daha iyi sonuçlar elde edilebildiği ancak geçerliğe ilişkin puanlama yöntemlerinin tamamının benzer sonuçlar ortaya çıkardığı görülmüştür.

Diedenhofen ve Musch (2015) ise çalışmalarında deneysel seçenek ağırlıklandırmanın iki kategorili puanlamaya göre geçerlik ve güvenilirliği nasıl değiştirdiğini belirlemeyi amaçlamışlardır. Araştırmada deneysel seçenek ağırlıkları belirlenirken tüm seçenekler ve maddelerin tamamı için seçenek-toplam korelasyonları elde edilmiştir. Bu sonuçlar geleneksel puanlamayla karşılaştırılarak incelendiğinde deneysel seçenek ağırlıklandırmanın test puanlarının geçerliğini ve güvenilirliğini yükselttiğini göstermiştir. Ayrıca sonuçlara göre cevaplayıcılara ilişkin doğru sınıflandırma oranı, deneysel seçenek ağırlıklandırma için daha yüksektir. Buna göre, deneysel seçenek ağırlıklandırma kullanıldığında bir maddeye ilişkin cevaplayıcıdaki bilginin derecesinin daha iyi yansıtılabildiği yorumu yapılmıştır.

Salehi, Sadighi ve Bagheri (2015) araştırmalarında güven ağırlıklandırması yöntemini kullanarak cevaplayıcılardan işaretledikleri seçeneğin yanına cevaplarından ne kadar emin olduklarını yazmalarını istemişlerdir. Bu çalışmada güven ağırlıklandırması yöntemiyle puanlamanın testin güvenilirliğini, yordama geçerliğini ve katılımcıların benlik saygısı düzeylerini ve maddelerin cinsiyet bakımından yanlılığını nasıl etkilediğinin incelenmesi amaçlanmıştır. Çalışma grubu ise Dilbilgisi dersini alan 49 İngilizce anadal öğrencisinden oluşturulmuştur. Öte yandan bu çalışmada “1-0” puanlama yöntemi de kullanılarak elde edilen sonuçlar karşılaştırılmıştır. Buna göre kullanılan iki puanlama yaklaşımının testin güvenilirliği bakımından benzer sonuçlar verdiği belirlenmiştir. Ancak güven ağırlıklandırmasına dayalı puanlama yaklaşımının geçerlik bakımından daha avantajlı bir durum ortaya koyduğu bulunmuştur.

Sočan (2015), bir simülasyon çalışmasıyla geleneksel puanlama, düzeltme formülünün kullanıldığı durumda geleneksel puanlama ve deneysel seçenek ağırlıklandırma yöntemleriyle elde edilen puanların geçerliğini ve iç tutarlılığını karşılaştırmıştır. Bu çalışmada deneysel seçenek ağırlıklandırmanın geleneksel puanlamaya kıyasla test geçerliğini ve güvenilirliğini artırabileceği görülmüştür. Ayrıca bu çalışmada puanlama yöntemlerinin göreceli performansının simülasyon koşullarından çeldiricilerin özelliklerine büyük ölçüde bağlı olduğu, örneklem büyüklüğü ve test uzunluğuna ise önemli ölçüde bağlı olmadığı ifade edilmiştir.

Diedenhofen ve Musch (2019), diğer bir çalışmalarında çoktan seçmeli bir testin puanlanmasında geleneksel yaklaşımla puanlama, uzman yargısına dayalı ve deneysel

seenek ağırlıklandırma yöntemleri kullanıldığında teste ilişkin güvenilirlik ve geçerlik kestirimlerini karşılaştırmıştır. Seenek ağırlıklarının uzman görüşüne mi yoksa deneysel verilere mi dayanması gerektiğini belirlemeyi amaçladıkları bu araştırmada genelleştirilebilir sonuçlar elde etmek için, tekrarlanan rastgele alt örnekleme doğrulamasını da kullanmışlardır. Araştırmada elde edilen sonuçlara göre deneysel seenek ağırlıklandırmanın, testin güvenilirliğini arttırdığı görölmüştür. Bunun yanında seenek ağırlıklandırma prosedürlerinden hiçbirinin test geçerliğini iyileştirmediği raporlanmıştır.

Bir diğerk alışmada ise yazılı bir yoklamanın ayrıntılı ve genel puanlama yöntemleriyle puanlanması durumunda sınav puanlarının güvenilirliği, aritmetik ortalaması, madde güçlükleri ve madde ayırıcılıklarının nasıl değıştiğinin gösterilmesi amaçlanmıştır (Acar, 2018). Lisans eğitiminde İstatistik dersi almış olan 79 öğrenci, alışmanın alışma grubunda yer almaktadır. Kullanılan puanlama yöntemine göre yazılı yoklamada yer alan maddelerin madde güçlük indekslerinde farklar ortaya çıktığı belirlenmiştir. Ayrıca ayrıntılı puanlama yapılan soruların madde ayırıcılıklarının genel puanlama yapılan madde ayırıcılıklarından daha yüksek olduğu belirlenmiştir. Güvenirlik bakımından bu iki yöntemin kullanılması durumunda benzer sonuçlar elde edildiği görölmüştür. Ancak sınav puanlarının bu iki puanlama durumuna ilişkin hesaplanan ortalamaları arasında manidar bir fark olduğu görölmüştür.

Çoktan seçmeli maddelerde seenek ağırlıklandırma yönteminin kullanıldığı başka bir alışmada, bu yöntemlerin testin faktör yapısına etkisi araştırılmıştır. Geleneksel yöntem, seenek-toplam korelasyonu ve karşılıklı ortalamalar yöntemleriyle puanlama yapıldıktan sonra her bir durumda AFA ve DFA gerçekleştirilerek testin faktör yapısına bakılmıştır. Faktör yapılarının karşılaştırılması amacıyla AFA sonucunda elde edilen faktör yükleri ve açıklanan varyansların değışimi incelenmiş, ayrıca farklı puanlama durumlarındaki DFA sonucunda elde edilen uyum istatistikleri karşılaştırılmıştır. AFA sonuçlarına göre deneysel puanlamanın özdeğer ve açıklanan varyans oranında düşmeye sebep olduğu gözlenmiştir. DFA sonuçları ise, deneysel puanlama durumunda model uyumunun iyileştğini göstermiştir. Ancak, indekslerin modelden modele istatistiksel olarak önemli düzeyde farklılaşmadığı görölmüştür (Erdem, Ertuna ve Doğan, 2016).

Ağırlıklı puanlama yöntemlerinin kullanılmasıyla ilgili bazı alışmalarda puanlama yöntemiyle maddelerde görülebilen değışen madde fonksiyonu arasındaki ilişkinin araştırıldığı görölmüştür. Özdemir (2003), oktan seçmeli bir testin

puanlanmasında “1-0” ve önsel ağırlıklı puanlama yöntemlerinin kullanılmasının değişen madde fonksiyonuna etkisini araştırmıştır. Söz konusu çoktan seçmeli test, çeşitli illerde yer alan 15 ilkokulda öğrenim gören öğrencilerden seçilmiş bir çalışma grubuna uygulanmıştır. Çalışmada geleneksel ve önsel ağırlıklı puanlama yöntemlerinin kullanılması durumunda hesaplanan madde puanlarının okul türü ve cinsiyet bakımından yanlılık gösterip göstermediği incelenmiştir. Çalışmada elde edilen sonuçlara göre, her iki puanlama yönteminin kullanıldığı durumda da test maddelerinin çoğunluğunun alınan değişkenlere göre önemli düzeyde yanlılık sergilemediği belirlenmiştir.

Selvi (2013)’ün çalışmasında KTK ve MTK kapsamındaki değişen madde fonksiyonu (DMF) belirleme yöntemleri, çeşitli puanlama durumlarında (1-0 puanlama, uzman yargısına dayalı önsel ağırlıklandırma (UYDA)) incelenmiştir. Çalışmada 2006-2007 yılı bahar döneminde, Hacettepe ve Gazi Üniversitelerinin çeşitli bölümlerine devam eden 1593 kişiye uygulanan sözel yetenek testine ilişkin veriler kullanılmıştır. Sonuçta UYDA puanlama durumunda, belirlenen DMF’li madde sayısının iki kategorili puanlama durumuna göre iki kattan fazla artışı görülmüştür. Yani UYDA puanlama durumunda daha fazla DMF varlığı belirlenmiştir. Elde edilen bu sonucun sebeplerinin ağırlıklandırma işlemiyle cevaplayıcıların kısmi bilgilerinin puanlamada daha çok yansıtılması, yetenek boyutuna ilişkin daha büyük bir ranj için bilgi sağlanması ve daha büyük güvenirlik kestirimleri elde edilmesi olduğu ifade edilmektedir.

Tunç (2016) da araştırmasında iki ve çok kategorili puanlanmış maddeler için belirlenen değişen madde fonksiyonlarının (DMF) karşılaştırılmasını amaçlamıştır. Bu çalışmada farklı örneklem büyüklüğü, DMF sergileyen maddelerin yüzdesi ve DMF büyüklüğü koşullarında bir simülasyon çalışması gerçekleştirilerek madde yanlılığına ilişkin I. tip hata ile istatistiksel güç oranlarını belirlenmiştir. Ağırlıklı puanlama yapıldığında madde yanlılığına ilişkin I. tip hata oranlarının genel olarak daha düşük, istatistiksel güç oranlarının daha yüksek olduğunu belirlemiştir. Araştırma sonuçları doğrultusunda DMF sonuçları üzerinde kullanılan puanlama yaklaşımının etkisinin olabileceği ve kısmi puanların göz önünde bulundurulması durumunda maddelerin sergilediği DMF sonuçlarının değişebileceği görülmüştür. Wetzel, Böhnke, Carstensen, Zeigler ve Ostendorf (2013) ise bir kişilik envanterindeki maddelere ilişkin kategorilerin çeşitli yöntemlerle puanlanmasıyla elde edilen veriye dayalı olarak cinsiyet bakımından MTK’ya dayalı yöntemlerle DMF analizlerini gerçekleştirmiş ve

maddelerin DMF sonuçlarında puanlanmadaki y nteme g re deęiřiklik olabileceęini ifade etmiřlerdir.

Maddelerde g r lebilen DMF sonu larında olduęu gibi, genel olarak kullanılan farklı puanlama modellerinin herhangi bir ama la yapılan istatistiksel analizlerin bulgularını deęiřtirebildięi s ylenebilmektedir.  rneęin  rt k sınıf analizinin    farklı puanlama durumundaki iřleyiřini incelemek amacıyla ger ekleřtirilen bir arařtırmada,  oktan se meli bir test ikili, uzman kanısına dayalı ve deneysel aęırlıklı puanlama y ntemleriyle puanlanmıřtır. Bu y ntemlerle alınan veriye  rt k sınıf analizi uygulanmıřtır (Selvi ve Derici Yıldırım, 2020).  alıřmadan elde edilen sonu lara g re, deneysel aęırlıklı puanlama y ntemiyle ulařılan modelin en az  rt k sınıf sayısı ve en az parametre kestirimiyle model veri uyumunu saęladığı, b ylece  rt k sınıf analizine en uygun model olduęu g r lm řt r.

Farklı Madde T rlerinde Puanlama

Psikolojik  zellikleri  l mek veya eęitim alanlarıyla ilgili olarak farklı d zeyde ger ekleřen  ęrenmeleri yoklamak  zere bir ok farklı madde t r nden yararlanılabilmektedir (G ltekin, 2017). Testlerde kullanılan madde t rleri, cevabın birey tarafından oluřturulup oluřturulmaması, puanlamanın tutarlılıęı ya da nesnellięi gibi  eřitli durumlar g z  n nde bulundurulurak farklı řekillerde sınıflandırılmaktadır. Madde t rleri i in  eřitli sınıflamalar yapılsa da bunlardan en  ok cevaplayıcının cevap verme řekline g re yapılan sınıflama kullanılmaktadır. Buna g re maddeler; cevaplayıcının yanıtı se erek verdięi ve yanıtını kendisinin oluřturduęu maddeler olarak iki gruba ayrılmaktadır (Crocker ve Algina, 1986, Turgut, 1997, Haladyna, 1994). Yanıtın cevaplayıcı tarafından se ildięi birinci gruptaki madde t rleri doęru-yanlıř t r ndeki maddeler, eřleřtirmeli ve  oktan se meli maddelerdir. Cevaplayıcıdan maddenin cevabını oluřturması ve yazması beklenen madde t rleri ise kısa yanıtlı ve uzun yanıtlı a ık u lu maddelerdir (Haladyna, 1997;  zg ven, 1998).

Se me gerektiren madde t rleri grubunda  oktan se meli maddeler, cevaplayıcının yanıtı oluřturduęu madde t rleri grubunda ise kısa yanıtlı maddeler olarak yanıtı sınırlı a ık u lu maddeler yaygın bi imde kullanılmaktadır (G ltekin, 2017). Alanyazın incelendięinde hem  oktan se meli hem de kısa yanıtlı maddelerin

puanlama sürecinde kullanılan pek çok farklı yaklaşım ve yöntemin bulunduğu görülmüştür.

Çoktan Seçmeli Maddelerin Puanlanması

Çoktan seçmeli maddeler yaygın kullanımda geleneksel puanlama yöntemiyle puanlanmaktadır. Bu yönüyle puanlamanın objektif, hızlı ve kolay bir biçimde yapılması sağlanmaktadır. Ancak bu durumda geleneksel puanlamaya ilişkin kısmi bilginin dikkate alınmaması, tüm maddelerin aynı nitelikte olduğunun kabul edilmesi gibi eleştirilen durumların tamamı gündeme gelmektedir. Ayrıca çoktan seçmeli maddelerin en büyük sınırlılıklarından biri şans başarısından etkilenmesidir. Şans başarısının etkisini azaltmak ve rastgele yanıt verme davranışının önüne geçmek için genel olarak düzeltme formülü kullanılması önerilse de (Sax, 1997) bu formülün kullanılması konusunda tartışmalar sürmektedir (Baykul, 2015). Düzeltme formülü maddeyi hiç okumadan tamamen rastgele yanıt veren öğrenciyle eksik/yanlış bilgiye sahip öğrenciyi ayıramadığından her ikisi de cezalandırılmaktadır (Gültekin, 2017). Ayrıca Yurdugül (2010)'un çalışmasında farklı ölçme kuramlarına dayalı olarak hesaplanan test puanlarına göre sıralama karşılaştırması yapıldığında, düzeltme formülüyle birlikte geleneksel puanlamayla elde edilen başarı sıralamalarının diğer test puanı türlerine ilişkin sıralamalarla en düşük uyuma sahip olduğu bulunmuştur.

Çoktan seçmeli maddelerin puanlanmasında ağırlıklı puanlama yöntemleri de kullanılabilmektedir. Yapılan araştırmalar incelendiğinde çoktan seçmeli madde türüne ilişkin genellikle seçenek ağırlıklandırmaya dayalı puanlama yönteminin kullanıldığı (Akkuş, 2000; Crehan ve Haladyna, 1994; Davis ve Fifer, 1959; Diedenhofen ve Musch, 2015; Gözen Çıtak, 2007; Haladyna ve Sympson, 1988; Haladyna, 1990; Özdemir, 2002; Özdemir, 2003; Salehi, Sadighi ve Bagheri, 2015) görülmüştür. Wang ve Stanley (1970)'e göre çoktan seçmeli maddelerde, doğru cevabı bilmeyen öğrencilerin mümkün olduğunca eşit yüzdelerle seçebilecekleri şekilde çeldiricileri düzenlemenin zor olduğu durumlarda seçenek ağırlıklandırmanın kullanımı uygundur. Çeldiriciler sorunun mantıklı yanlış yanıtları (Haladyna, 1997) olarak hazırlandığından bunların gerçekten kısmi bilgiye sahip cevaplayıcılar tarafından seçildiği varsayılmaktadır. Ancak pratikte seçeneklerin içinde kendi cevabını veya cevabının benzerini göremeyen cevaplayıcıların seçenekleri inceleyerek bunlardan eleme yapmak

yerine doğru seçeneği bulamadıklarını kabullenip seçeneklerden birini rastgele işaretledikleri görülebilmektedir (Umay, 1998). Waters (1976) eğer çoktan seçmeli bir testte seçenek ağırlıklandırma yapılıyorsa, bu çalışmaların doğru yürütülmesi için çeldiricilerin rastgele cevaplama davranışı yerine mantıkla, varsa gerçekten kısmi bilgiyle seçilmesi gerektiğini ifade etmiştir. Fakat çoktan seçmeli maddelerde cevaplayıcıların işaretledikleri seçenekleri seçme sebepleri ve bunların gerekçelerini tam olarak belirlemek mümkün değildir (Karataş, Köse ve Coştu, 2003). Bu durumda cevaplayıcı davranışları, çoktan seçmeli testlerde ağırlıklı puanlama yöntemleri kullanıldığında elde edilen sonuçların tutarsızlığına sebep olan etkenlerden biri olarak karşımıza çıkmaktadır (Frery, 1982).

Çoktan seçmeli maddeler puanlanırken seçenek ağırlıklandırma yöntemleriyle birlikte veya tek başına madde ağırlıklandırma yöntemleri de kullanılabilmektedir. Bu yöntemlerin kullanılabilmesi için maddelerin faktör analizi yoluyla ağırlıklandırılması (Burt, 1950), tek tek maddeler yerine madde kümelerinin ağırlıklandırılması (Gulliksen, 1950), maddelere çoklu regresyon ile değer atama, kısmi regresyon katsayılarıyla puanlama (Guilford, 1954), test varyansı veya madde varyansına dayalı ağırlıklandırmayla puanlama (Dick, 1965), ilgili oldukları konunun önemine göre ağırlıklandırma (Ghiselli, 1964), madde ayırıcılık indeksleri ile ağırlıklandırma (Birnbau, 1968) gibi çeşitli yöntemler önerilmektedir. Ancak bu ağırlıkların belirlenmesi için geleneksel puanlama yoluyla elde edilen madde puanlarına şans başarısı nedeniyle şansla alınmış puanların karışması durumu söz konusudur. Şans başarısı ve buna bağlı olarak şans puanı madde ve test puanlarında ve bunlara dayalı olarak hesaplanan istatistiklerde hatalara sebep olmaktadır (Baykul, 2015).

Çoktan seçmeli testlerin puanlanmasında kullanılan yöntemlerde, her durumda şans başarısı önemli bir sakınca olarak değerlendirilmiş, bu durumdan kurtulmak veya sebep olduğu problemleri en aza indirmek için çeşitli yollar aranmıştır. Bu doğrultuda testin puanlanması sürecinde şans başarısının etkisini azaltmaya yönelik bazı uygulamalar yapılmıştır. Gözlenen puanlara karışmış olan şans başarısını tahmin etmek amacıyla çeşitli düzeltme formülleri geliştirilmiştir (Horst, 1966). Bunun yanında farklı test türleri veya farklı uygulama koşulları da kullanılarak şans başarısının test puanlarına etkisi ortadan kaldırılmaya çalışılmıştır. Örneğin çoktan seçmeli bir sınavın düzeltme formülü kullanılarak puanlandığı bir çalışmada, bu düzeltilmiş puanların aynı bilgileri yoklayan kısa cevaplı bir sınava ilişkin puanlarla önemli ölçüde iyi uyum sağladığı görülmüştür. Ayrıca kısa yanıtlı maddelerin ve çoktan seçmeli maddelerin

düzeltilme formülüyle birlikte kullanılmasının ayrı ayrı öğrencilerin ders puanlarında önemli gelişmelerle sonuçlandığı belirtilmiştir. Düzeltme formülünün kullanılmasına ilişkin olası problemlerden dolayı kısa cevaplı maddelerin kullanılmasının avantajlı olabileceği ifade edilmiştir (Pinckard, McMahan, Prihoda, Littlefield & Jones, 2011). Ölçmeyi amaçladıkları davranışlar birbirinin aynısı olan biri çoktan seçmeli diğeri kısa yanıtı olan iki testin sonuçlarının karşılaştırıldığı bir çalışmada ise, başka bir sınav ölçüt olarak alındığında, bu sınavdaki başarıyı kısa yanıtı test ile hesaplanan puanların daha iyi yorduyor olduğu görülmüştür. Kısa yanıtı testte şans başarısının olmamasının bu duruma sebep olduğu ifade edilmiştir. Ayrıca çoktan seçmeli testin farklı türden testlere olan üstünlük ve avantajı korumak için şans faktörünü en aza indirmesi gerektiği ifade edilmiştir (Kan ve Kayapınar, 2006). Şans faktörünün en aza indirgenmesiyle ilgili önerilerden biri, madde tepki kuramına dayalı üç parametrelili lojistik modelin kullanılmasıdır. Bu modeldeki “c parametresi”, madde ayırt ediciliği ve madde güçlüğüne yanında cevaplayıcının maddeyi yanıtlarken sadece tahminle doğru seçeneği bulma olasılığını da dikkate alarak puanlama yapmaktadır. Her bir madde için şansa doğru cevap verme ihtimalini dışarıda tutarak gerçek puanları ortaya çıkarmaktadır (Baker, 2001; Crocker ve Algina, 1986; Hambleton, 1994). Bu alternatifili ilişkin güçlükler ise madde tepki kuramına dayalı üç parametrelili lojistik modelin kullanılabilmesi için oldukça büyük bir cevaplayıcı grubuna ihtiyaç duyulması ve modelin sağlanması zor olan birtakım varsayımlar gerektirmesidir (DeMars, 2016). Öte yandan c parametresinin hesaplanmasında da şans başarısı etkili olmaktadır. Özdoğan ve Doğan (2018), çoktan seçmeli bir testte düzeltme yapılmayan ve şans başarısından arıtılan veriye ilişkin c parametreleri arasında manidar bir fark bulunduğunu göstermişlerdir. Böylece çoktan seçmeli testlerde elimine edilmesi mümkün görünmeyen şans faktörünün en aza indirgenmesine yönelik çabaların yerine, özellikle çok büyük olmayan cevaplayıcı gruplarında farklı madde türlerinin kullanılması gerekliliği gündeme gelmektedir.

Kısa Yanıtlı Maddelerin Puanlanması

Cevaplayıcının yanıtını kendisinin yapılandığı madde türleri grubunda bulunan kısa yanıtı maddelerde ise cevaplayıcının seçim yapmak yerine cevabı sözcük, ifade, sembol veya sayılardan bir tanesiyle oluşturması, işlem basamaklarının

görülebileceği şekilde problemi çözmesi, iki ya da üç cümleyi geçmeyecek şekilde açıklama yapması istenmektedir (Tekin, 2000). Bu tür maddelerin puanlamasında da önerilen puanlama yöntemlerinden biri geleneksel puanlamadır (Tan ve Erdoğan, 2004). Örneğin matematiksel problem çözme becerisinin ölçülmesi amacıyla cevabı uzun olmayan ve fazla sayıdaki maddenin kullanılması durumunda, bu maddeler çoktan seçmeli testlerle aynı şekilde 0 ve 1 puanları atanarak puanlanabilmektedir (Umay, 1997). Böylece hem şans başarısından arınmış hem de büyük oranda objektif puanlama yapılabilmesi kısa yanıtli maddelerin kullanım alanını genişletmektedir (Nitko, 2004). Kısa yanıtli maddeler, daha çok soru sorulmasına imkân vermesi ve iki kategorili olarak puanlanması durumunda puanlamanın kolay olması sebebiyle avantajlı bir format olarak görülmektedir (Kubiszyn & Borich, 2003). Ancak geleneksel puanlama yöntemlerine ilişkin eleştirilerin tamamı (kısmi bilginin dikkate alınmaması, tüm maddelerin puanlamada aynı ağırlıklarda ele alınması gibi) yanıtın tamamen cevaplayıcı tarafından yapılandırıldığı bu madde türü için de geçerli olmaktadır.

Alanyazın incelendiğinde kısa cevaplı maddelerin geleneksel puanlama yöntemiyle puanlandığı ve farklı madde türlerinin kullanıldığı durumlarda elde edilen madde/test özelliklerinin karşılaştırıldığı çeşitli çalışmalara (Gözen, 2002; Gültekin ve Çıkrıkçı Demirtaşlı, 2012; Kan ve Kayapınar, 2006; Lancaster, 1987; Özdemir, 1997) rastlanmaktadır. Özdemir (1997), aynı davranışı ölçmek amacıyla hazırlanmış farklı türdeki maddelerden oluşan testlere ilişkin psikometrik özellikleri incelediği araştırmasında, çoktan seçmeli testin daha kolay; ancak daha ayırt edici olduğunu, ayrıca test güvenilirliği bakımından en güvenilir test olduğunu belirlemiştir. Lancaster (1987), çoktan seçmeli maddeler ile kısa yanıtli maddeler için hesaplanan madde güçlük ve madde ayırt edicilik indekslerini karşılaştırmıştır. Araştırma sonucunda madde güçlük ve ayırt edicilik indeksleri arasında manidar bir fark bulunmadığı görülmüştür. Aynı davranışları ölçmek amacıyla hazırlanmış olan çoktan seçmeli ve kısa yanıtli iki ayrı testin kullanıldığı bir çalışmada (Kan ve Kayapınar, 2006) ise madde ayırt edicilik indeksleri incelendiğinde, kısa yanıtli testin maddelerinin madde ayırt edicilik indekslerinin daha büyük olduğu belirlenmiştir. Gözen'in (2002) araştırmasında, kısa yanıtli ve çoktan seçmeli maddelerden oluşan iki ayrı testin farklı puanlama durumlarındaki geçerlik ve güvenilirlik kanıtları incelenmiştir. Sonuçta geleneksel yaklaşımla puanlama yapıldığında güvenilirlik bakımından kısa yanıtli testin daha iyi sonuçlar verdiği görülmüştür. Madde ayırt edicilik indeksine ilişkin sonuçlar incelendiğinde ise tüm puanlama durumlarında test türleri arasında manidar bir fark

görülmemiştir. Başka bir çalışmada (Gültekin ve Çıkrıkçı Demirtaşlı, 2012) ölçmeyi amaçladıkları yapıların benzer olduğu çoktan seçmeli ve kısa yanıtlı maddelerin iki farklı testte ve aynı testte bir arada kullanıldığı durumlarda elde edilen madde ve yetenek parametreleri; bilgi fonksiyonları ve göreceli etkinlik indeksleri karşılaştırılmıştır. Tüm koşullarda yalnızca 1-0 puanlama yöntemi kullanılmıştır. Araştırma sonuçlarına göre kısa yanıtlı maddelerle cevaplayıcılara ilişkin en çok bilgi elde edilebilmektedir. Ayrıca testte yer alan kısa yanıtlı madde sayısı arttıkça ölçülen özellik bakımından bireyler arası farklılıkların daha etkili ortaya konabildiği görülmüştür. Bu çalışmalarda kısa yanıtlı testlerle “1-0” puanlama yöntemi kullanılmış olmasına rağmen elde edilen bu sonuçlara göre, cevaplayıcıların performanslarının ağırlıklı puanlama yöntemleriyle daha hassas bir biçimde puanlanması durumunda psikometrik açıdan daha da nitelikli sonuçlar elde edilebileceği öngörülmektedir.

Kısa yanıtlı maddelerin puanlanmasıyla ilgili olarak diğer öneri ise yanıtlayıcıların uzun yanıt gerektiren açık uçlu maddelere göre az da olsa istedikleri yanıtı verme özgürlüklerinin olması nedeniyle puanlama anahtarının kullanılmasıdır (Doğan, 2020). Ancak bu durumda kısa yanıtlı maddeler puanlamanın öznelliğinden tam olarak arınmış olmaz; çünkü öğrencinin yazdığı yanıtın doğruluğuyla ilgili puanlayıcılar tarafından sıklıkla öznel kararlar alınabilmektedir (Gültekin, 2017). Sonuçta puanlayıcılardan veya puanlama anahtarından kaynaklanan yanlılık/sübjektiflik ve puanlama güvenilirliğinin tam olmaması problemleri ortaya çıkmaktadır. Puanlayıcıların görüş ayrılıkları, objektiflik eksikliği, açık/belirli olmayan puanlama anahtarı gibi çeşitli hata kaynaklarıyla karşılaşmaktadır. Puanlama işlemine bu hataların karışmasının öncelikli sebebi yapılan puanlamanın puanlayıcı kanısı gerektirmesidir. Oysaki ölçmelerin objektif olması, bütün puanlayıcıların aynı cevaba aynı puanı vermesinin sağlanmasına bağlıdır (Turgut ve Baykul, 2012). Dolayısıyla kısa cevaplı maddelerin puanlayıcılar aracılığıyla puanlanması durumunda puanlama işlemine puanlayıcılardan kaynaklı olan hataların karışması söz konusu olmaktadır.

Kısa yanıtlı maddelere ilişkin puanlama anahtarıyla puanlama sürecinde puanlayıcıların görev alması, maliyet ve zaman bakımından da önemli bir yük oluşturmaktadır (ETS, 2009). Özellikle puanlanacak madde ve cevaplayıcı sayısının çok olduğu durumlarda görevlendirilecek puanlayıcı sayısının da artırılmasına ihtiyaç duyulmaktadır. Ayrıca eğer puanlayıcılar yardımıyla puanlama yapılıyorsa, puanlayıcıların verdikleri puanların tutarlılığını ve doğruluğunu izlemek için puanlayıcı- içi ve puanlayıcılar-arası güvenilirliğin kestirilmesi gibi ekstra işlemlerin yapılması da

gerekmektedir (Cohen, Swerdlik ve Phillips, 1996). Çünkü puanlayıcılar tarafından verilen öznel puanların ve bu puanlar aracılığıyla alınan kararların hangi düzeyde doğru olduğunu araştırmak için puanların tutarlılığı ortaya konmalıdır (Atılğan, Kan ve Doğan, 2011). Dolayısıyla özellikle kalabalık cevaplayıcı gruplarının olduğu ve puanlayıcılar tarafından puanlamanın pratik bir alternatif olmadığı durumlarda kısa yanıtli soruların kullanımını sağlayacak farklı puanlama yöntemleri ihtiyaç haline gelmektedir.

Kısa yanıtli maddelerin ağırlıklı puanlama yöntemleriyle puanlanması da mümkündür ancak alanyazında incelenen araştırmaların büyük bir kısmında (Akkuş, 2000; Crehan ve Haladyna, 1994; Davis ve Fifer, 1959; Diedenhofen ve Musch, 2015; Gözen Çıtak, 2007; Haladyna ve Sympson, 1988; Haladyna, 1990; Özdemir, 2002; Özdemir, 2003; Salehi, Sadighi ve Bagheri, 2015; Saygı, 2004) ağırlıklı puanlama yöntemlerinin çoktan seçmeli testlerin puanlanmasında kullanıldığı ve farklı türdeki testlerle birlikte maddelere ilişkin ağırlıklandırmaların yapıldığı çalışmaların (Gözen, 2002, Phillips, 1943; Quist, 2004) oldukça az sayıda olduğu belirlenmiştir.

Farklı madde türleriyle birlikte geleneksel puanlama yöntemlerinin ve ağırlıklı puanlama yöntemlerinin kullanıldığı literatürde yer alan çalışmalarda elde edilen sonuçların genel olarak tutarsız sayılabileceği görülmektedir. Bununla birlikte bu tutarsızlığın söz konusu çalışmalarda kullanılan testlerin ve çalışma evreninin özelliklerindeki büyük farklılıklardan ve kullanılan ölçme araçlarının niteliğinden kaynaklandığı da ifade edilmektedir (Frery, 1982). Genellikle ağırlıklı puanlamanın birey hakkında daha fazla bilgi verdiği ve cevaplayıcıların kısmi bilgilerini de dikkate aldığı için kullanılmasının faydalı olacağı belirtilmiştir. Crehan ve Haladyna (1994) ise o döneme kadar yapılan çalışmaların özeti olarak doğrusal ağırlıklandırma yöntemlerinin iç tutarlılığı biraz artırdığını, ancak dış kriterlerle olan korelasyonların (geçerlik) genellikle iyileşmediği ifade etmiştir. Bununla birlikte incelenen çalışmaların bazılarında (Diedenhofen ve Musch, 2015; Guilford, Lovell ve Williams, 1942; Haladyna ve Sympson, 1988; Kılıç ve Doğan, 2019; Phillips, 1943; Selvi, 2013) çok kategorili puanlamanın testin güvenirliğini ve geçerliğini arttırdığı bulunmuş, bazı araştırmalarda (Bauer, Holzer, Kopp ve Fischer, 2011; Crehan ve Haladyna, 1994; Davis ve Fifer, 1959; Gözen Çıtak, 2007; Haladyna, 1990; Quist, 2004; Salehi, Sadighi ve Bagheri, 2015, Tunç, 2016) güvenirlik ve geçerlikten birini yükselttiği raporlanmış, bazı çalışmalarda ise (Akkuş, 2000; Gözen, 2002; Saygı, 2004) bu puanlama yaklaşımlarının güvenirlik ve geçerliğe katkı sağlamadığı yönünde sonuçlar elde

edilmiştir. Ayrıca farklı puanlama yöntemlerinin geleneksel puanlama yöntemine göre daha uzun zaman aldığı ve sonuçta elde edilen puanların güvenilirlik ve geçerliklerinin de en azından harcanan çabaya değecek kadar yüksek olmadığı sonucuna ulaşan çalışmalara (Akkuş, 2000; Guilford, Lovell ve Williams, 1942; Phillips, 1943) rastlanmıştır. Ancak günümüzde bilgisayar teknolojisinin oldukça gelişmiş olması sayesinde ağırlıklandırma işlemleri daha kolay/pratik şekilde yapılabilmektedir. Bu nedenle ağırlıklı puanlamanın geçerlik ve güvenilirliğe verebileceği kısmi katkıların bile, cevaplayıcıların ölçülen özellik bakımından düzeyine ilişkin daha fazla bilgi edinilmesi açısından önemli olduğu düşünülmektedir.

İncelenen bazı çalışmalarda ise ağırlıklı puanlamaya yönelik her durumda geçerli bir karar almak yerine, bu yöntemlerin kullanılabileceği ya da kullanılmaması gereken bazı özel durumların olduğu ileri sürülmüştür. Haladyna (1990), testte yer alan maddelere ilişkin çeldiricilerin farklı düzeyde bilgiler içermesi durumunda ağırlıklı puanlamanın kullanılması gerektiğini belirtmiştir. Başka bir çalışmada ise çok kategorili puanlamanın lisans/sertifika verme programlarına ilişkin testlerde uygun olabileceği ancak bir üniversite veya lisansüstü kabul testinde uygun olmayabileceği ifade edilmiştir (Crehan ve Haladyna, 1994).

Ağırlıklı puanlama yöntemlerinin de kullanıldığı durumlarda testlerin psikometrik özellikleri açısından incelendiği, Türkiye’de yürütülen araştırmaların sayısının ise oldukça az olduğu görülmüştür. Bu araştırmalarda (Akkuş, 2000; Gözen Çıtak, 2007; Özdemir, 2003; Saygı 2004) genelde çoktan seçmeli maddelerle birlikte seçenek ağırlıklandırmaya dayalı puanlama yaklaşımları kullanılmıştır. Ayrıca madde ağırlıklandırma yöntemlerinin de kullanıldığı oldukça sınırlı sayıda çalışmaya (Gözen 2002; Kılıç ve Doğan, 2019; Yurdugül, 2010) rastlanmıştır.

Ağırlıklı puanlama yaklaşımları işe koşularak yürütülen çalışmalarda ayrıca, daha tutarlı sonuçlar elde edilmesi adına dikkat edilmesi gereken durumlar da ifade edilmiştir. En dikkat çeken hususlardan biri, eğer çoktan seçmeli bir testte seçenek ağırlıklandırma yapılıyorsa, bu çalışmaların doğru yürütülmesi için çeldiricilerin rastgele cevaplama davranışı yerine mantıkla, varsa gerçekten kısmi bilgiyle seçilmesi gerektiğidir (Waters, 1976). Ancak incelenen çalışmalarda (Crehan ve Haladyna, 1994; Davis ve Fifer, 1959; Gözen, 2007; Haladyna ve Sympson, 1988; Haladyna, 1990; Saygı, 2004) genellikle cevaplayıcıların rastgele cevaplama davranışlarından kaynaklanabilecek sorunların etkisini azaltmaya yönelik herhangi bir işleme rastlanmamıştır. Ayrıca incelenen araştırmaların büyük bir kısmında (Akkuş, 2000;

Crehan ve Haladyna, 1994; Davis ve Fifer, 1959; Diedenhofen ve Musch, 2015; Gözen Çıtak, 2007; Haladyna ve Simpson, 1988; Haladyna, 1990; Özdemir, 2002; Özdemir, 2003; Salehi, Sadighi ve Bagheri, 2015; Saygı, 2004) ağırlıklı puanlama yöntemlerinin çoktan seçmeli testlerin puanlanmasında kullanıldığı ve farklı türdeki testlerle birlikte maddelere ilişkin ağırlıklandırmaların yapıldığı çalışmaların (Gözen, 2002, Phillips, 1943; Quist, 2004) oldukça az sayıda olduğu belirlenmiştir. Bütün bu durumlar göz önünde bulundurularak bu araştırmada, kısa yanıtlı bir testin yanıtlarının önce geleneksel puanlama daha sonra çeşitli ağırlıklı puanlama yöntemleriyle (maddelerin ağırlıklandırılması, olası cevapların ağırlıklandırılması, çoklu ağırlıklandırma) puanlanması durumlarında testin psikometrik özellikleri incelenmiştir.



BÖLÜM 3

YÖNTEM

Bu bölümde araştırmanın modeli, çalışma grubu, araştırmada kullanılan ölçme araçları, verilerin toplanması ve çözümlenmesine ilişkin bilgiler sunulmuştur.

Araştırmanın Modeli

Bu araştırmada farklı puanlama yöntemlerinin kullanıldığı durumlarda toplanan verinin bir olgunun tanımlaması, açıklanması, çıkarım yapılması için analiz edilmesi olarak ifade edilen nicel yaklaşım benimsenmiştir. Önce “1-0” puanlama daha sonra çeşitli ağırlıklı puanlama yöntemlerinin kullanıldığı durumlarda kısa yanıtlı bir testin psikometrik özelliklerinin incelenmesini amaçlayan bu çalışma, tarama modelinde ve ilişkisel araştırma türündedir. Tarama modeli, eskiden veya hala var olan durumları olduğu gibi tespit etmeye yönelik bir modeldir. İncelenen birey, vaka ya da nesne herhangi bir şekilde değiştirilmeden veya etkilenmeden kendi terimleriyle tanımlamaya çalışılır. İlişkisel araştırmalar ise değişkenler arasında karşılaştırma yolu ile elde edilen ilişkileri belirlemek için yapılan araştırmalardır (Karasar, 2016).

Çalışma Grubu

Araştırmanın verileri elde edilirken araştırma kapsamında geliştirilen kısa yanıtlı testin deneme formu 2021-2022 öğretim yılında Ankara ili Çankaya, Yenimahalle ve Mamak ilçelerinde bulunan resmî ortaokullarda yedinci sınıfta öğrenim gören öncelikle 59 öğrenciye ön deneme amacıyla uygulanmıştır. Daha sonra bu uygulamadan elde edilen sonuçlara göre düzenlenen deneme formu, deneme uygulaması kapsamında benzer özellikteki 218 öğrenciye uygulanmıştır. Deneme uygulamasına dayalı olarak düzenlenen kısa yanıtlı test formu doğrulayıcı faktör analizi için veri elde etmek amacıyla 2022-2023 öğretim yılında Ankara ili Çankaya, Yenimahalle ve Mamak ilçelerinde bulunan resmî ortaokullarda yedinci sınıfta öğrenim gören 200 öğrenciye uygulanmıştır. Son olarak testin hazırlanan nihai formu 2022-2023 öğretim yılında

yedinci sınıfta öğrenim gören 636 öğrenciye uygulanmıştır. Nihai test formunun uygulandığı çalışma grubu, Ankara ili Çankaya, Yenimahalle ve Mamak ilçelerinde bulunan resmî ortaokul ve imam hatip ortaokullarının yedinci sınıflarında öğrenim gören öğrencilerden belirlenmiştir. Ön deneme uygulamasından itibaren her aşamada çalışma grubunda yer alan öğrencilerin evreni temsil edecek özellikte, Ankara ilinde farklı ilçelerde bulunan farklı sosyoekonomik düzeydeki çeşitli resmî ortaokulların yedinci sınıflarında öğrenim görüyor olmasına dikkat edilmiştir. Uygulamanın gerçekleştirilebilmesi için gerekli izin alınmıştır (Ek 1). Nihai test formunun uygulandığı çalışma grubundaki öğrencilerin sayıları ve öğrenim gördükleri okullara ait bilgiler Tablo 1’de sunulmuştur.

Tablo 1

Çalışma Grubunda Bulunan Okullar ile Öğrenci Sayıları

Okul Adı	İlçe	Öğrenci Sayısı
Ortaokul-A	Çankaya	60
Ortaokul-B		76
İmam Hatip Ortaokulu-A		45
Ortaokul-C	Yenimahalle	118
Ortaokul-D		95
Ortaokul-E		50
İmam Hatip Ortaokulu-B		53
Ortaokul-F	Mamak	66
Ortaokul-G		73
Toplam		636

Tablo 1’de verilen öğrenci sayıları incelendiğinde, çalışma grubunda yer alan 636 öğrencinin 181’inin Çankaya, 316’sının Yenimahalle ve 139’unun Mamak ilçelerindeki okullara devam ettiği görülmektedir. Ayrıca çalışma grubundaki öğrencilerin 538’i resmî ortaokulların, 98’i ise resmî imam hatip ortaokullarının yedinci sınıfında öğrenim görmektedir.

Verilerin Toplanması

Bu araştırmada kullanılan verilerin elde edilmesi için öncelikle yedinci sınıf düzeyindeki öğrencilerin matematik dersinin rasyonel sayılar konusundaki öğrenme düzeylerini ölçmeye yönelik kısa cevaplı bir test geliştirilmiştir. Test geliştirme

sürecinde öncelikle testle ölçülecek davranışlar ve bunların her birine yönelik hangi bilişsel düzeyden kaç madde hazırlanacağı belirlenerek belirtke tablosu hazırlanmıştır. Daha sonra bu belirtke tablosuna uygun biçimde 32 tane kısa yanıtlı madde yazılmıştır. Yazılan maddelerin ilgili kazanıma, bilişsel düzeye, ölçme-değerlendirme ilkelerine uygunluğunun ve dil açısından anlaşılabilirliğinin incelenmesine yönelik matematik alanından beş, eğitimde ölçme ve değerlendirme alanından üç ve Türkçe alanından bir uzmandan görüş alınmıştır. Böylece belirtke tablosu ve uzman görüşleri yoluyla testin kapsam geçerliğinin sağlanması hedeflenmiştir.

Alan uzmanlarına ulaştırılan görüş formlarında, uzmanların hazırlanan her bir maddeye ilişkin görüşleri “Uygundur. (1)”, “Hafifçe gözden geçirilmelidir. (2)”, “Ciddi olarak gözden geçirilmelidir. (3)” ve “Uygun değildir. (4)” biçiminde dördü olarak derecelendirilmiştir. Uzman görüş formlarının tamamı geri alındıktan sonra her bir maddeye ilişkin uzman görüşleri bir formda birleştirilmiştir. Daha sonra her bir madde için kapsam geçerlik indeksi Davis tekniği (1992) ile hesaplanmıştır. Bu tekniğe göre aday maddelerin her biri için uzman görüş formlarındaki “Uygundur. (1)” ve “Hafifçe gözden geçirilmelidir. (2)” biçiminde ifade edilen görüşlerin sayısının uzman sayısına bölünmesiyle kapsam geçerlik indeksleri elde edilmektedir. Hesaplanan kapsam geçerlik indeksi 0.80’den büyük ise bu durum maddeye ilişkin kapsam geçerliğinin yeterli olduğunu ifade etmektedir. Bu değerden düşük kapsam geçerlik indeksine sahip olan maddelerin kullanılmaması önerilmektedir (Davis, 1992). Bu çalışma kapsamında hazırlanan aday maddelerin her biri için Davis tekniğine göre hesaplanan kapsam geçerlik indeksleri Tablo 2’de yer almaktadır.

Tablo 2

<i>Aday Maddelerin Davis Tekniğine Göre Hesaplanan Kapsam Geçerlik İndeksleri</i>										
Madde No	Mat 1	Mat 2	Mat 3	Mat 4	Mat 5	Ölçme 1	Ölçme 2	Ölçme 3	Türkçe	Kapsam Geçerlik İndeksi
1	1	1	2	1	1	2	1	2	1	1,00
2	2	1	1	1	2	2	2	2	2	1,00
3	2	1	1	1	1	1	2	2	1	1,00
4	2	1	4	1	1	4	2	4	2	0,67
5	1	1	2	1	1	1	1	1	1	1,00
6	1	1	1	1	3	1	1	1	1	0,89
7	1	1	1	1	1	4	1	1	1	0,89
8	2	2	1	1	1	1	2	2	1	1,00
9	2	1	1	1	1	2	2	2	1	1,00

(Devam ediyor)

Tablo 2 (Devam)

<i>Aday Maddelerin Davis Tekniğine Göre Hesaplanan Kapsam Geçerlik İndeksleri</i>										
Madde No	Mat 1	Mat 2	Mat 3	Mat 4	Mat 5	Ölçme 1	Ölçme 2	Ölçme 3	Türkçe	Kapsam Geçerlik İndeksi
10	1	1	2	1	1	2	1	1	1	1,00
11	1	1	2	1	1	2	1	1	1	1,00
12	1	1	2	1	1	2	1	1	1	1,00
13	1	1	1	1	1	4	1	1	1	0,89
14	2	1	2	1	1	1	2	2	1	1,00
15	2	1	2	1	1	1	2	2	1	1,00
16	1	1	1	1	1	2	1	1	1	1,00
17	1	1	1	1	1	2	1	1	1	1,00
18	1	1	1	1	1	2	1	1	1	1,00
19	1	2	1	1	1	1	1	1	1	1,00
20	2	1	2	1	1	3	2	2	1	0,89
21	1	1	2	1	1	3	1	1	1	0,89
22	1	1	2	1	1	3	1	1	2	0,89
23	1	1	1	1	1	2	1	2	1	1,00
24	1	1	1	1	1	2	1	1	1	1,00
25	1	1	1	1	1	2	1	1	1	1,00
26	1	2	1	1	1	2	1	1	1	1,00
27	1	1	1	1	1	4	1	1	1	0,89
28	1	1	3	1	1	2	1	2	1	0,89
29	1	1	1	1	1	1	1	1	1	1,00
30	1	1	1	1	1	1	1	1	1	1,00
31	1	1	1	1	2	1	1	1	1	1,00
32	1	1	1	1	1	1	1	1	1	1,00

Tablo 2'deki hazırlanan maddelerin her biri için Davis tekniğine göre hesaplanan kapsam geçerlik indeksleri incelendiğinde bir maddenin (madde 4) kapsam geçerlik indeksinin 0,80'den düşük olduğu görülmüştür. Buna göre bu madde daha sonraki aşamalarda kullanılmamıştır. Diğer maddeler ise uzmanların uzman görüş formlarında yer alan açıklamalardaki görüşlerine göre düzenlenmiştir.

Uzman görüşlerine göre yapılan redaksiyon sürecinin ardından maddelerin cevaplayıcılar tarafından anlaşılabilirliği, uygun cevaplama süresi gibi bilgilere ulaşmak için hazırlanan formlar ön deneme uygulaması kapsamında yedinci sınıf düzeyindeki 59 öğrenciye uygulanmıştır. Bu uygulamanın ardından alınan öğrenci cevapları incelenip maddeler düzenlenerek redaksiyon süreci tamamlanmıştır.

Redaksiyon sürecinin ardından 19 maddeden oluşan bir deneme formu hazırlanarak deneme uygulaması gerçekleştirilmiştir. Bu uygulama ile deneme formu

yedinci sınıf düzeyindeki 218 öğrenciye uygulanmıştır. Hazırlanan kısa cevaplı test formunun yönergesinde öğrencilerden maddenin altındaki boşluğa çözüm yaptıktan sonra ulaştıkları sonucu, gösterilen kutucuğa yazmaları istenmiştir. Ayrıca her bir maddede, madde kökünde istenen sonuca ulaşamayanların da çözümde geldikleri en son adımda buldukları sonucu yine bu kutucuğa yazmaları istenmiştir. Testler dağıtılmadan önce örnek durumlarla yapılacak puanlama işlemleri öğrencilere uygun biçimde anlatılmıştır. Ayrıca uygulamanın süresi ve uyulması gereken kurallara ilişkin açıklamalar yapılmıştır.

Deneme uygulamasının ardından 16 öğrenciye ait formlarda maddelerin tamamının boş bırakıldığı veya hepsine ilgisiz yanıtlar verildiği görüldüğünden bu formlar daha sonraki aşamalarda kullanılmamıştır. Kalan 202 öğrenciden elde edilen cevaplar geleneksel yöntemle puanlanarak madde ve test istatistikleri hesaplanmıştır. Deneme uygulaması kapsamında uygulanan 19 maddelik kısa yanıtı test formunun test istatistikleri Tablo 3'te verilmiştir.

Tablo 3

Deneme Formuna İlişkin Test İstatistikleri

N	202
Ortalama	8,15
ÖSH	1,84
Ortanca	7
Mod	7
Standart Sapma	4,75
Çarpıklık	-0,08
Basıklık	-0,33
Ranj	19
Minimum	0
Maksimum	19

Tablo 3'e göre testin deneme formuna ilişkin çarpıklık ve basıklık değerleri incelendiğinde test puanlarının normal dağılıma yakın bir dağılım sergilediği gözlenmiştir. Deneme uygulaması kapsamında uygulanan kısa yanıtı test formunun madde istatistikleri ise Tablo 4'te gösterilmiştir.

Tablo 4

Deneme Formunun Madde İstatistikleri

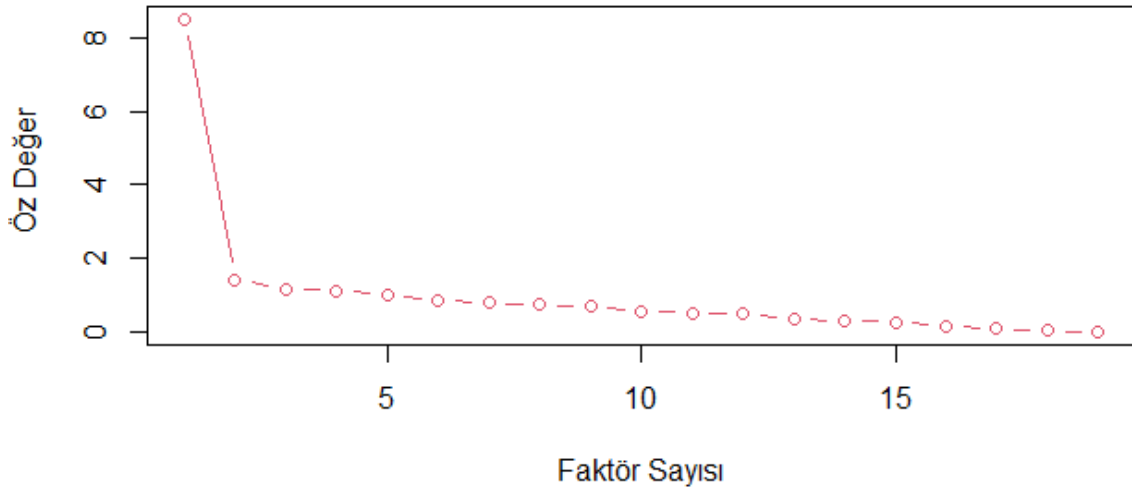
Madde No	p_i	r_{jx}	Madde No	p_i	r_{jx}
1	0,44	0,27	11	0,27	0,56
2	0,71	0,44	12	0,41	0,35
3	0,37	0,53	13	0,41	0,67
4	0,33	0,32	14	0,37	0,48
5	0,50	0,39	15	0,54	0,14
6	0,42	0,41	16	0,59	0,37
7	0,76	0,41	17	0,36	0,41
8	0,71	0,53	18	0,37	0,37
9	0,65	0,46	19	0,65	0,61
10	0,52	0,62			

Tablo 4'e göre testin deneme formunda yer alan maddelerin madde güçlük indeksleri $p_i=0.27$ ile $p_i=0.76$ aralığında değişmektedir. Buna göre test formunda yer alan maddelerden altısının kolay, yedi maddenin orta güçlükte ve altı maddenin zor olduğu belirlenmiştir. Deneme formundaki maddelerin ayırt edicilik indeksleri incelendiğinde ise maddelerin 17'si için bu değerlerin $r_{jx} = 0.30$ değerinden büyük dolayısıyla yeterli olduğu görülmüştür. Bu formdaki birinci maddenin ayırt edicilik indeksinin düşük olduğu için gözden geçirilmesi gerektiği ve 15 numaralı maddenin ise esas uygulamada kullanılmaması gerektiği görülmüştür.

Deneme uygulamasında toplanan veriler aracılığıyla madde ve test istatistikleri hesaplandıktan sonra testin yapı geçerliğini gösteren kanıtlar elde etmek için, alınan puanlara dayalı olarak AFA yapılmıştır. Burada faktör çıkarma (faktörleştirme) yaklaşımı olarak temel bileşen analizi yöntemi kullanılmıştır. AFA uygulanmadan önce kayıp değerler ve uç değerler incelenmiştir. Maddelerin tamamının boş bırakıldığı veya tüm maddelere ilgisiz yanıtlar verilen formlar çıkarıldıktan sonra kalan 202 kişilik veri setinde kayıp değer olmadığı görülmüştür. Uç değerleri incelemek amacıyla toplam puanlar standart z puanlarına dönüştürülmüştür. Elde edilen standart puanlarda (-3,3) aralığının dışında kalan değer olmadığı görülmüştür. Deneme uygulamasından elde edilen verilere ilişkin normal dağılım gereğinin sınanması amacıyla Kolmogorov Smirnov testi kullanılmıştır. Kolmogorov Smirnov Testi sonuçlarına göre puan dağılımının normal dağılımdan farklı olduğu görülmüştür ($p<0.05$). Fakat normal dağılım gereğinin sınanması amacıyla kullanılan testler, normal dağılım kararına ilişkin katı olabilmektedir. Bu sebeple bu test sonuçlarının yanında betimsel istatistik değerleri aracılığıyla normal dağılım gereği incelenebilmektedir. Bu çalışmada da deneme uygulamasında toplanan verilerin çarpıklık ve basıklık değerlerinin (-1,1) aralığında

olduğu görölerek toplam puanların dağılımının normal olduđu varsayılmıştır (Deniz, 2020). AFA öncesinde verinin faktör analizine uygun olup olmadığının kontrolü için örneklem büyüklüğünün yeterliliğini incelemek amacıyla Kaiser-Meyer-Olkin (KMO) değeri hesaplanmıştır. KMO değerinin 0.79 olduđu görölmüştür. Elde edilen bu değeri örneklem büyüklüğünün yeterli olduđunu göstermektedir. Bartlett küresellik testi sonuçları da incelenmiş ve bu test sonucunun ($p<0.05$) manidar olması da veri setinin faktörleşmeye uygun olduđuna bir kanıt olarak görölmüştür. Bu bulgular doğrultusunda örneklem büyüklüğünün yeterli olduđu ve analize ilişkin gereklerin sağlandığı sonucuna varılmıştır (Leech, Barrett ve Morgan, 2005; Çokluk, Şekercioğlu ve Büyüköztürk, 2012). Test maddelerinin iki kategorili olarak puanlanması nedeniyle tetrakorik korelasyon matrisi yardımıyla AFA yapılmıştır (Muthén, 1984).

Deneme formundan elde edilen verilere dayalı olarak 19 madde ile yapılan ilk analizde özdeğeri 1’den büyük olan beş faktör olduđu görölmüştür. AFA sonucuna göre elde edilen özdeğer grafiğı Şekil 3’te yer almaktadır.



Şekil 3. Deneme Formuna İlişkin Özdeğer Grafiğı

Şekil 3’teki özdeğer grafiğı incelendiğinde ikinci faktörden itibaren grafiğın yatay bir seyir izlediği görölmektedir. Söz konusu beş faktöre ilişkin özdeğerler ve özdeğer grafiğı incelenerek toplam varyansa yaptıkları katkının önemi çerçevesinde yalnızca ilk faktörün varyansa önemli bir katkı yaptığı, ikinci faktörden itibaren faktörlerin toplam varyansa katkısının hem küçük hem de yaklaşık olarak aynı olduđu görölmüştür. Öte yandan özdeğer grafiğı birinci faktörden sonra güçlü bir kırılma göstermektedir. Bu çerçevede analizin tek faktör için tekrarlanması düşünülmüştür. Bu

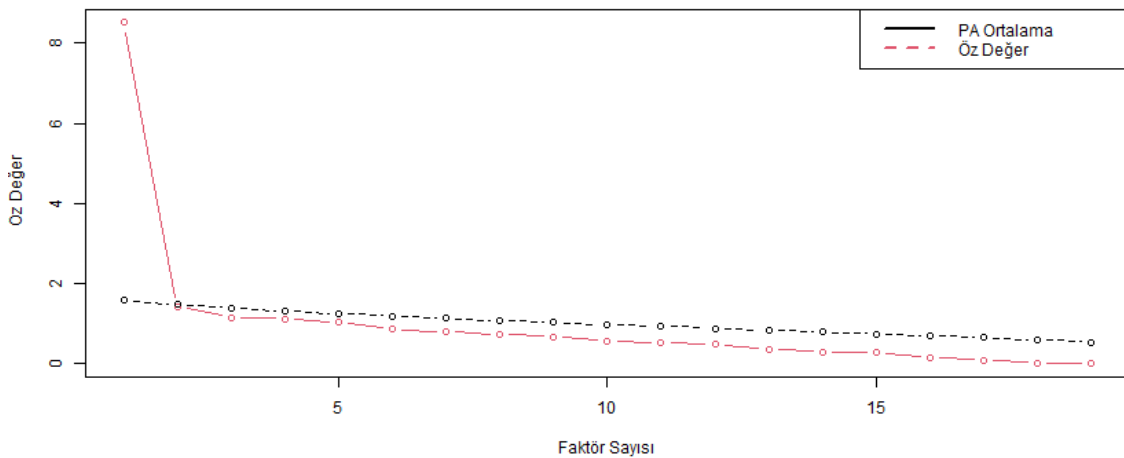
karara dayanak oluřturmak ve faktör sayısına karar vermek amacıyla Horn'un paralel analizi yöntemi (Horn, 1965) kullanılmıřtır. Bu dođrultuda (R core team, 2022)'da yer alan "hornpa" (Huang, 2015) paketindeki aynı isimli fonksiyondan faydalanılmıřtır. Faktör sayısını belirlemeden önce ilk analizde elde edilen özdeđerler, paralel analiz sırasında yansız üretilen örneklemler ile elde edilen özdeđer ortalamalarıyla karřılařtırılmıřtır. Paralel analizin sonuçları ile AFA'dan elde edilen özdeđerler ařađıdaki Tablo 5'te karřılařtırılmıřtır.

Tablo 5

Paralel Analizden Elde Edilen Kriter Deđerlerin Özdeđerlerle Karřılařtırılması

Bileřen numarası	AFA'dan Elde Edilen Özdeđer	Paralel Analiz Kriter Deđer	Karar
1	8,51	1,58	Kabul
2	1,42	1,47	Red
3	1,16	1,38	Red
4	1,12	1,31	Red
5	1,02	1,24	Red

Tablo 5'te görüldüğü gibi yalnızca ilk faktör için AFA'dan elde edilen özdeđerin paralel analiz aracılığıyla hesaplanan özdeđerlerin ortalamasından büyük olduđu görülmektedir. Bu durumda analizin devamında faktör sayısının bir olarak alınması kararı paralel analizin sonuçları ile desteklenmiřtir. Ayrıca hem AFA hem de paralel analiz sonucunda elde edilen özdeđerlerin bulunduđu özdeđer grafiđi řekil 4'te gösterilmiřtir.



řekil 4. Deneme Formuna İliřkin Özdeđer Grafiđi ve Paralel Analiz Sonuçları

Şekil 4'teki özdeğer grafiğinde kesikli siyah çizgi paralel analiz sonucunda elde edilen özdeğerlerin ortalamasını, düz kırmızı çizgi ise orijinal veriden elde edilen özdeğerleri göstermektedir. Sonuçlara bakıldığında ilk faktör için orijinal özdeğerlerin paralel analiz ile elde edilen özdeğerlerin ortalamasından büyük olduğu görülmektedir. Bu durumda analizin devamında faktör sayısının bir olarak alınmasına karar verilmiştir.

AFA sonucunda elde edilen faktör yük değerlerinin 0,32'den büyük olması gerekliliği (Field, 2009) dikkate alınmıştır. Analiz tek faktörlü olarak tekrarlandıktan sonra faktör yük değeri 0,32'den küçük olan üç madde (madde 1, madde 8 ve madde 15) olduğu görülmüş ve bu maddeler en düşük faktör yüküne sahip olandan başlanarak sırasıyla testten çıkarılmıştır. Son olarak 16 madde ve tek faktörle yapılan AFA'dan elde edilen maddelerin faktör yük değerleri ve ortak faktör varyansları Tablo 6'da yer almaktadır.

Tablo 6

<i>Deneme Formundaki Maddelerin Faktör Yük Değerleri ve Ortak Faktör Varyansları</i>		
Madde No	Faktör Yük Değeri	Ortak Faktör Varyansı
m2	0,58	0,34
m3	0,72	0,52
m4	0,70	0,49
m5	0,70	0,49
m6	0,72	0,52
m7	0,60	0,36
m9	0,77	0,59
m10	0,80	0,64
m11	0,74	0,55
m12	0,83	0,69
m13	0,80	0,64
m14	0,84	0,71
m16	0,74	0,55
m17	0,75	0,56
m18	0,75	0,56
m19	0,54	0,29

Tablo 6'da görüldüğü gibi madde 1, madde 8 ve madde 15 çıkarıldıktan sonra son durumda testte 16 madde bulunmaktadır. Maddelerin faktör yük değerleri 0,54 ile 0,84 arasında değişmektedir. Ayrıca faktör yük değeri 0,32'den küçük olan herhangi bir maddenin olmadığı görülmektedir. Maddelerin ortak faktör varyanslarına bakıldığında

ise ortak faktör varyansı 0,20'den düşük olan madde olmadığı için hesaplanan değerlerin yeterli olduğu söylenebilmektedir (Tabachnick ve Fidel, 2013). AFA sonuçları incelenerek toplam 16 maddenin tek faktör altında toplandığı ve bu faktöre ilişkin öz değerin 8,49 olduğu belirlenmiştir. Ayrıca 16 madde ile tekrarlanan analizdeki tek faktörün varyansın toplam %53,1'ini açıklamış olduğu görülmüştür. Son olarak deneme formunda yer alan maddelerin iç tutarlılığına ilişkin hesaplanan KR-20 güvenirlik katsayısı 0,85 bulunmuştur. Bu doğrultuda testin güvenirliğinin yüksek düzeyde olduğu söylenebilmektedir.

Araştırmada bir sonraki adımda AFA ile tanımlanan modelin doğrulanıp doğrulanmadığını kontrol etmek amacıyla 200 kişilik başka bir öğrenci grubundan toplanan veriye dayalı olarak DFA yapılmıştır. Ölçme aracındaki maddelere ilişkin ölçümlerin faktör yapısı, tek faktörlü ölçme modeli tanımlanarak DFA ile test edilmiştir.

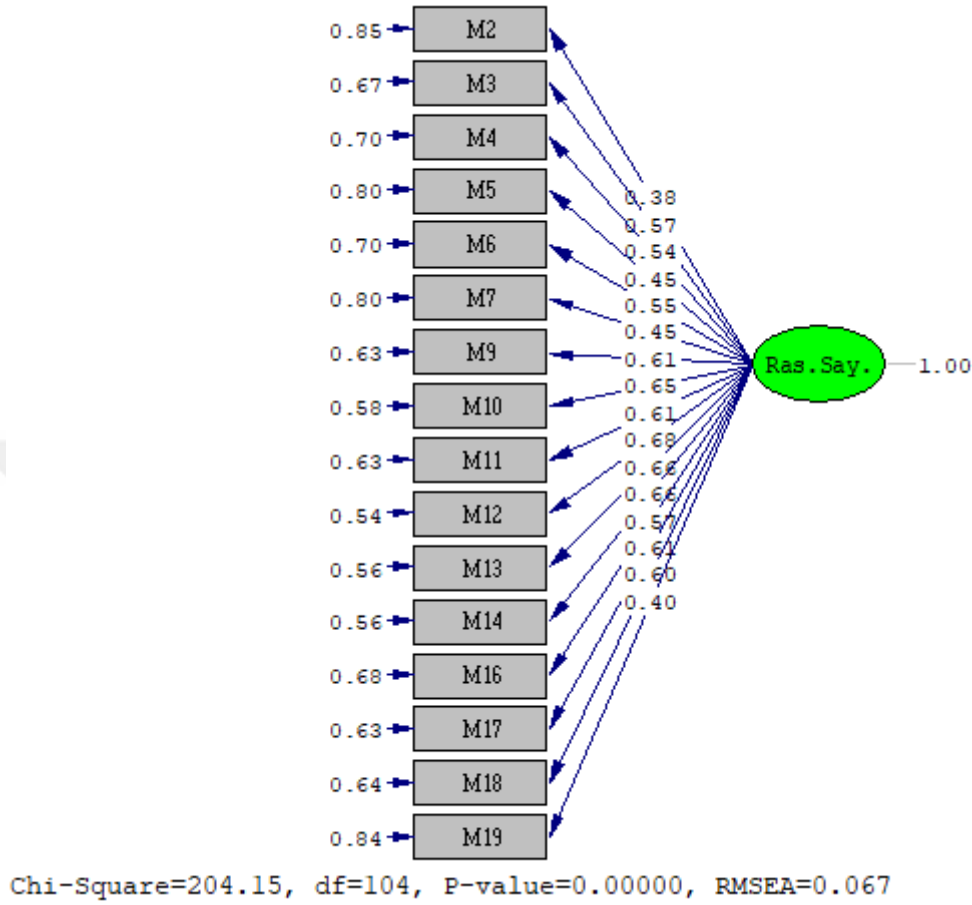
Verinin modele uyum sağlama düzeyini değerlendirirken uyum indekslerinden RMSEA (The Root Mean Square Error of Approximation), CFI (Comparative Fit Index), AGFI (Adjusted Goodness of Fit) ve NNFI (Non-normed Fit Index) dikkate alınmıştır. $RMSEA \leq 0,08$ (Jöreskog ve Sorbom, 1993) ve $CFI, AGFI, NFI \geq 0,90$ (Kline, 2011) olması modelin veriye yeterli uyum sağladığını göstermektedir. Deneme formundan elde edilen puanlara uygulanan DFA sonucunda görülen uyum indeksleri Tablo 7'de yer almaktadır.

Tablo 7

<i>Deneme Formuna İlişkin DFA Sonucunda Elde Edilen Uyum İndeksleri</i>				
RMSEA	CFI	AGFI	NNFI	χ^2/sd
0,067	0,96	0,96	0,95	1,96

Tablo 7'ye göre bu modelde RMSEA değeri 0,067 bulunmuştur. Bu değer model veri uyumunun kabul edilebilir düzeyde olduğunu ifade etmektedir. Bunun dışındaki uyum indekslerine bakıldığında CFI değerinin 0,96; AGFI değerinin 0,96; NNFI değerinin 0,95 olduğu görülmüştür. Bu uyum indekslerinin de iyi veya kabul edilebilir model veri uyumuna işaret ettiği söylenebilmektedir. Ayrıca modeldeki χ^2/sd oranının 1,96 olduğu belirlenmiştir. Bu değer 3'ün altında olması da model veri uyumunun istenen düzeyde sağlandığını göstermektedir (Kline, 2011). İncelenen

ölçme aracındaki maddeler için tanımlanan tek faktörlü modelin test edilmesiyle elde edilen yol şeması Şekil 5’te verilmiştir.



Şekil 5. Deneme Formunun Tek Faktörlü Modeli için Yol Şeması

Şekil 5’te verilen yol şemasındaki hata varyansları $\epsilon=0,54$ ve $\epsilon=0,85$ arasında değer almaktadır. Bu değerlerin de 0,90’dan küçük olması kabul edilebilir hata varyansı olarak değerlendirilmektedir (Tabachnick ve Fidell, 2013). İncelenen ölçme aracındaki verilere ilişkin yapılan DFA sonuçlarında elde edilen tüm uyum indeksleri ve hata varyansları birlikte ele alınırsa, tanımlanan modelin ilgili veri setine iyi uyum sağladığı ve göstergelerin ilgili değişkenlerin iyi birer temsilcisi olduğu sonucuna ulaşılmıştır.

Deneme uygulamasının ardından bu uygulamadan elde edilen sonuçlara göre testin 16 maddeden oluşan nihai uygulama formu düzenlenerek nihai uygulama gerçekleştirilmiştir. Bu uygulamada hazırlanan kısa cevaplı test formu yedinci sınıf düzeyindeki 636 öğrenciye uygulanmıştır. Uygulamanın gerçekleştirilebilmesi için yetkili kurumdan gerekli izin alınmıştır (Ek 1).

Kısa cevaplı test formunun yönergesinde öğrencilerden maddenin altındaki boşluğa çözüm yaptıktan sonra ulaştıkları sonucu, gösterilen cevap bölümüne yazmaları istenmiştir. Ayrıca her bir maddede, madde kökünde istenen sonuca ulaşamayanların da çözümde geldikleri en son adımda buldukları sonucu yine cevap bölümüne yazmaları istenmiştir. Testler dağıtılmadan önce örnek durumlarla yapılacak puanlama işlemleri öğrencilere uygun biçimde anlatılmıştır. Ayrıca uygulamanın süresi ve uyulması gereken kurallar açıklanmıştır.

Uygulamanın ardından 50 öğrenciye ait formlarda maddelerin tamamının boş bırakıldığı veya hepsine ilgisiz yanıtlar verildiği görüldüğünden bu formlar daha sonraki aşamalarda kullanılmamıştır. Kalan 586 öğrenciden elde edilen cevaplar ilk aşamada geleneksel yöntemle puanlanarak madde ve test istatistikleri hesaplanmıştır. Nihai uygulama kapsamında öğrencilere uygulanan 16 maddelik kısa yanıtlı test formunun test istatistikleri Tablo 8’de verilmiştir.

Tablo 8

Nihai Uygulama Formuna İlişkin Test İstatistikleri

N	586
Ortalama	6,12
ÖSH	1,65
Ortanca	5
Mod	3
Standart Sapma	4,42
Çarpıklık	0,39
Basıklık	-0,95
Ranj	16
Minimum	0
Maksimum	16

Tablo 8’e göre testin nihai formuna ilişkin çarpıklık ve basıklık değerleri incelendiğinde test puanlarının normal dağılıma oldukça yakın bir dağılım sergilediği gözlenmiştir. Nihai uygulama kapsamında kullanılan 16 maddelik kısa yanıtlı test formunun madde istatistikleri ise Tablo 9’da yer almaktadır.

Tablo 9

Nihai Uygulama Formunun Madde İstatistikleri

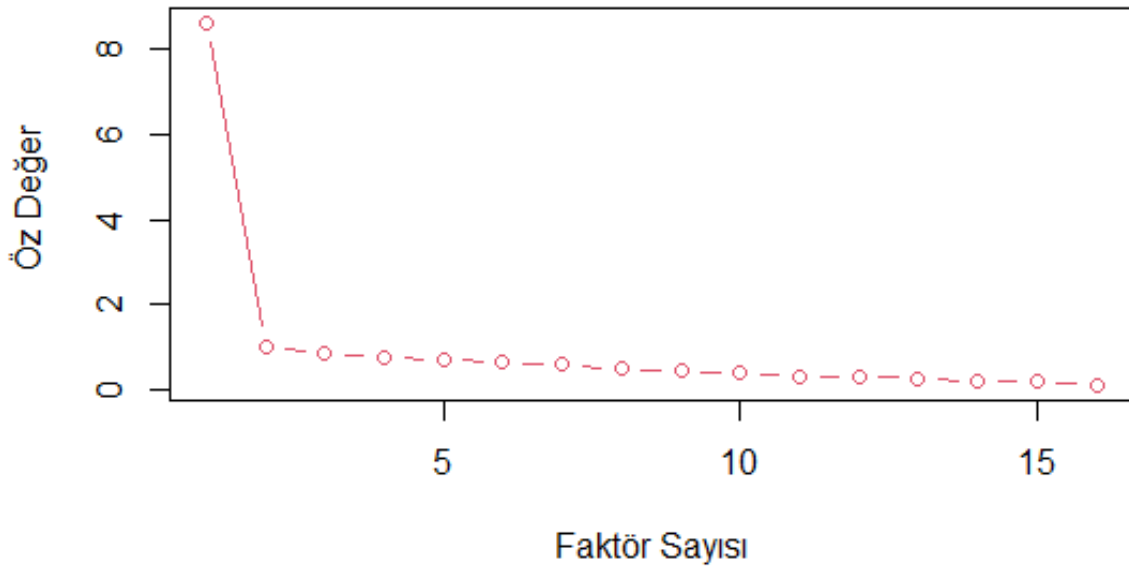
Madde No	p_j	r_{jx}	Madde No	p_j	r_{jx}
1	0,23	0,47	9	0,36	0,61
2	0,38	0,54	10	0,46	0,68
3	0,52	0,57	11	0,38	0,60
4	0,28	0,57	12	0,52	0,58
5	0,33	0,57	13	0,43	0,67
6	0,40	0,44	14	0,47	0,57
7	0,33	0,61	15	0,40	0,63
8	0,40	0,58	16	0,44	0,54

Tablo 9'a göre testin nihai formundaki maddelerin madde güçlük indeksleri $p_j=0,23$ ile $p_j=0,52$ arasında değişmektedir. Buna göre test formunda yer alan maddelerden sekiz maddenin orta güçlükte ve sekiz maddenin zor olduğu belirlenmiştir. Nihai uygulama formundaki maddelerin ayırt edicilik indeksleri için madde ve toplam puanlar arasındaki korelasyon nokta çift serili korelasyon katsayısı ile hesaplanmıştır (Baykul, 2015). Hesaplanan madde ayırt edicilik indeksleri incelendiğinde ise maddelerin tamamının ayırt edicilik indekslerinin oldukça yüksek olduğu görülmüştür.

Nihai uygulamadan elde edilen verilere dayalı olarak madde ve test istatistikleri hesaplandıktan sonra testin yapı geçerliğine kanıt elde etmek için AFA yapılmıştır. Bu aşamada AFA, nihai test formunun uygulandığı öğrenciler arasından tesadüfi olarak belirlenen 366 kişilik gruptan elde edilen verilere dayalı olarak yapılmıştır. Burada faktör çıkarma yaklaşımı olarak temel bileşen analizi yöntemi kullanılmıştır. AFA uygulanmadan önce kayıp değerler ve uç değerler incelenmiştir. Tesadüfi olarak belirlenen 366 kişilik veri setinde kayıp değer olmadığı görülmüştür. Uç değerleri incelemek amacıyla toplam puanlar standart z puanlarına dönüştürülmüştür. Elde edilen standart puanlarda (-3,3) aralığının dışında kalan değer olmadığı görülmüştür. Nihai uygulamadan elde edilen verilere ilişkin normal dağılım gereğinin sınanması için Kolmogorov Smirnov testi kullanılmıştır. Kolmogorov Smirnov Testi sonuçlarına göre puan dağılımının normal dağılımdan farklı olduğu söylenebilmektedir ($p<0.05$). Çarpıklık ve basıklık değerleri incelendiğinde ise bu değerlerin (-1,1) aralığında olduğu görülerek toplam puanların dağılımının normal olduğu varsayılmıştır (Deniz, 2020). AFA öncesinde veri setinin faktör analizine uygunluğunun kontrolü için örneklem büyüklüğünün yeterli olup olmadığını test etmek amacıyla KMO değeri hesaplanmıştır. Analiz sonucunda KMO değerinin 0,91 olduğu belirlenmiştir. Bu değer örneklem

büyükliğünün yeterli olduğunu göstermektedir. Bartlett küresellik testi sonuçları da incelenmiş ve bu test sonucunun ($p<0,05$) istatistiksel olarak anlamlı olması da veri setinin faktörleşme için uygun olduğuna bir kanıt olarak görülmüştür. Bu bulgular doğrultusunda örneklem büyüklüğünün yeterli olduğu ve analize ilişkin gereklerin sağlandığı sonucuna varılmıştır (Leech, Barrett ve Morgan, 2005; Çokluk, Şekercioğlu ve Büyüköztürk, 2012). Test maddelerinin iki kategorili olarak puanlanması nedeniyle tetrakorik korelasyon matrisi yardımıyla AFA yapılmıştır (Muthén, 1984).

Nihai test formundan elde edilen verilere dayalı olarak 16 madde ile yapılan ilk analizin sonucunda özdeğeri 1’den büyük olan iki faktör olduğu görülmüştür. Bu faktörlerin toplam varyansa yaptıkları katkının % 60,2 olduğu belirlenmiştir. Yapılan AFA sonucuna göre elde edilen özdeğer grafiği Şekil 6’da sunulmuştur.



Şekil 6. Nihai Test Formuna İlişkin Özdeğer Grafiği

Şekil 6’daki özdeğer grafiği incelendiğinde ikinci faktörden itibaren grafiğin yatay bir seyir izlediği görülmektedir. Söz konusu faktörlere ilişkin özdeğerler ve özdeğer grafiği incelenerek toplam varyansa yaptıkları katkının önemi çerçevesinde yalnızca ilk faktörün varyansa önemli bir katkı yaptığı, ikinci faktörden itibaren faktörlerin toplam varyansa katkısının hem küçük hem de yaklaşık olarak aynı olduğu görülmüştür. Öte yandan özdeğer grafiği birinci faktörden sonra güçlü bir kırılma göstermektedir. Bu çerçevede analizin tek faktör için tekrarlanması düşünülmüştür. Bu karara dayanak oluşturmak ve faktör sayısına karar vermek amacıyla Horn’un paralel analizi yöntemi kullanılmıştır. Bu doğrultuda (R core team, 2022)’da yer alan “hornpa” (Huang, 2015) paketindeki aynı isimli fonksiyondan faydalanılmıştır. Faktör sayısını

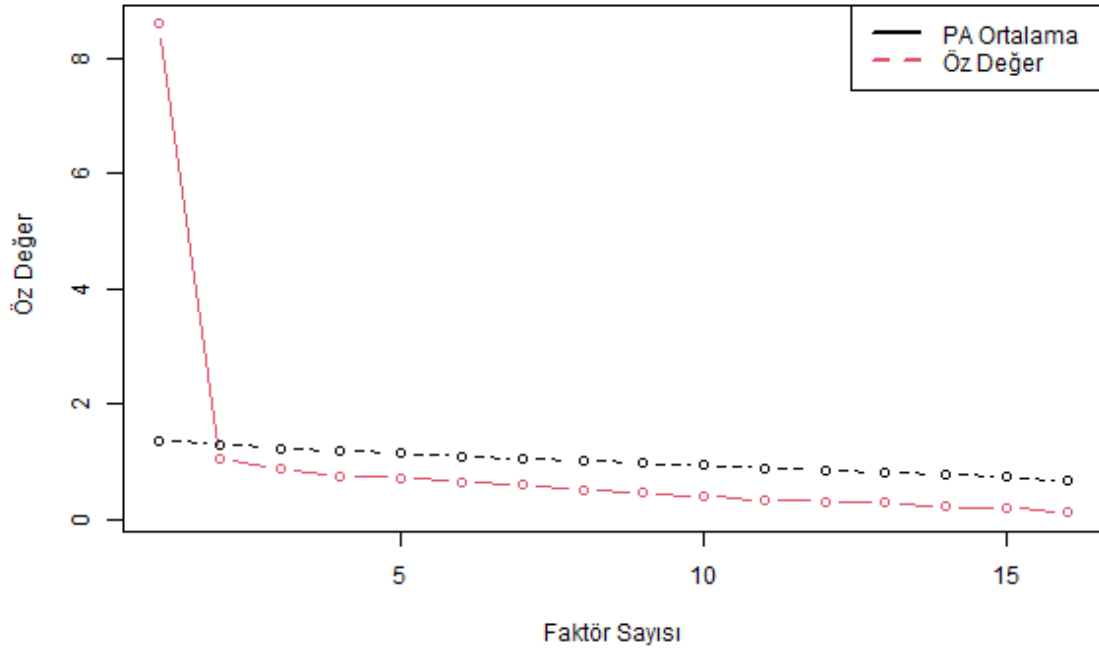
belirlemeden önce ilk analizde elde edilen özdeğerler, paralel analiz sırasında yansız üretilen örneklemeler ile elde edilen özdeğer ortalamalarıyla karşılaştırılmıştır. Paralel analizin sonuçları ile AFA'dan elde edilen özdeğerler Tablo 10'da karşılaştırılmıştır.

Tablo 10

Nihai Uygulamaya İlişkin Paralel Analizden Elde Edilen Kriter Değerlerin Özdeğerlerle Karşılaştırılması

Bileşen numarası	AFA'dan Elde Edilen Özdeğer	Paralel Analiz Kriter Değer	Karar
1	8,60	1,38	Kabul
2	1,03	1,29	Red
3	0,87	1,23	Red
4	0,75	1,18	Red
5	0,72	1,13	Red

Tablo 10'da görüldüğü gibi yalnızca ilk faktör için AFA sonuçlarında görülen özdeğerin paralel analizle elde edilen özdeğerlerin ortalamasından büyük olduğu görülmektedir. Bu durumda analizin devamında faktör sayısının bir olarak alınması kararı paralel analizin sonuçları ile desteklenmiştir. Ayrıca hem AFA hem de paralel analiz sonucunda elde edilen özdeğerlerin yer aldığı özdeğer grafiği Şekil 7'de sunulmuştur.



Şekil 7. Nihai Test Formuna İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları

Şekil 7’deki özdeğer grafiğinde kesikli siyah çizgi paralel analiz sonucunda elde edilen özdeğerlerin ortalamasını, düz kırmızı çizgi ise orijinal veriden elde edilen özdeğerleri göstermektedir. Sonuçlara bakıldığında ilk faktör için orijinal özdeğerlerin paralel analiz ile elde edilen özdeğerlerin ortalamasından büyük olduğu görülmektedir. Bu durumda analizin devamında faktör sayısının bir olarak alınmasına karar verilmiştir.

AFA sonucunda elde edilen faktör yük değerlerinin 0,32’den büyük olması gerekliliği (Field, 2009) dikkate alınmıştır. Analiz tek faktörlü olarak tekrarlandıktan sonra 16 madde arasında faktör yük değeri 0,32’den küçük olan madde olmadığı görülmüştür. Son olarak 16 madde ve tek faktörle yapılan AFA’da görülen maddelerin faktör yük değerleri Tablo 11’de yer almaktadır.

Tablo 11

<i>Nihai Test Formundaki Maddelerin Faktör Yük Değerleri ve Ortak Faktör Varyansları</i>		
Madde No	Faktör Yük Değeri	Ortak Faktör Varyansı
1	0,65	0,42
2	0,71	0,50
3	0,71	0,50
4	0,73	0,53
5	0,77	0,59
6	0,54	0,29
7	0,77	0,59
8	0,66	0,44
9	0,79	0,62
10	0,84	0,71
11	0,76	0,58
12	0,73	0,53
13	0,84	0,71
14	0,71	0,50
15	0,79	0,62
16	0,65	0,42

Tablo 11’de görüldüğü gibi nihai test formunda yer alan 16 maddeye ilişkin faktör yük değerleri 0,54 ile 0,84 arasında değişmektedir. Ayrıca faktör yük değeri 0,32’nin altında olan herhangi bir maddenin olmadığı görülmektedir. Maddelerin ortak faktör varyanslarına bakıldığında ise ortak faktör varyansı 0,20’den düşük olan madde olmadığı için hesaplanan değerlerin yeterli olduğu söylenebilmektedir (Tabachnick ve Fidel, 2013). AFA sonuçları incelendiğinde toplam 16 maddenin tek faktör altında

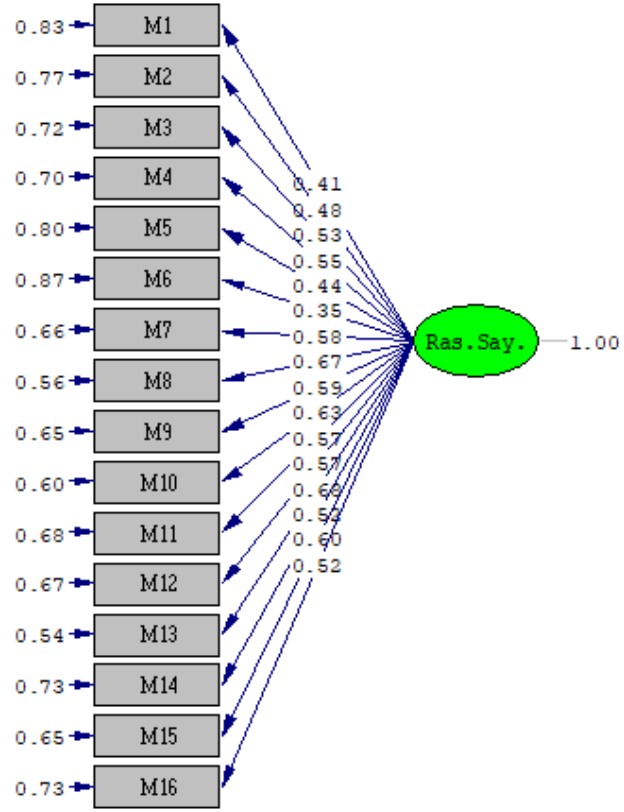
toplandığı ve tek faktöre ilişkin öz değerin 8,60 olduğu belirlenmiştir. Ayrıca 16 madde ile tekrarlanan analizdeki tek faktörün varyansın toplam %53,7'sini açıklamış olduğu görülmüştür. Son olarak bu formda yer alan maddelerin iç tutarlılığına ilişkin hesaplanan KR-20 güvenirlik katsayısı 0,86 bulunmuştur. Bu doğrultuda testin güvenirliğinin yüksek düzeyde olduğu söylenebilmektedir.

Araştırmada bir sonraki adımda veri setinde AFA sonrası kalan 220 kişilik başka bir öğrenci grubundan toplanan veriye dayalı olarak DFA yapılmıştır. Ölçme aracındaki maddelere ilişkin ölçümlerin faktör yapısı, tek faktörlü ölçme modeli tanımlanarak DFA ile test edilmiştir. Nihai test formundan elde edilen puanlara uygulanan DFA sonucunda görülen uyum indeksleri Tablo 12’de verilmiştir.

Tablo 12

<i>Nihai Test Formuna İlişkin DFA Sonucunda Elde Edilen Uyum İndeksleri</i>				
RMSEA	CFI	AGFI	NNFI	χ^2/sd
0,053	0,97	0,88	0,97	1,56

Tablo 12’ye göre bu modelde RMSEA değeri 0,053 bulunmuştur. Bu değer model veri uyumunun kabul edilebilir düzeyde olduğunu ifade etmektedir. Bunun dışındaki uyum indekslerine bakıldığında CFI değerinin 0,97; AGFI değerinin 0,88; NNFI değerinin 0,97 olduğu görülmüştür. Bu uyum indekslerinin de iyi veya kabul edilebilir model veri uyumuna işaret ettiği söylenebilmektedir. Ayrıca modeldeki χ^2/sd oranının 1,56 olduğu belirlenmiştir. Bu değer 3’ün altında olması da model veri uyumunun istenen düzeyde sağlandığını göstermektedir (Kline, 2011). İncelenen nihai test formundaki maddeler için tanımlanan tek faktörlü modelin test edilmesiyle elde edilen yol şeması Şekil 8’de verilmiştir.



Chi-Square=161.77, df=104, P-value=0.00025, RMSEA=0.053

Şekil 8. Nihai Test Formunun Tek Faktörlü Modeli İçin Yol Şeması

Şekil 8’de verilen yol şemasındaki hata varyansları $\varepsilon=0,54$ ve $\varepsilon=0,87$ arasında değer almaktadır. Bu değerlerin de 0,90’dan küçük olması kabul edilebilir hata varyansı olarak değerlendirilmektedir (Tabachnick ve Fidell, 2013). İncelenen nihai test formundaki verilere ilişkin yapılan DFA sonuçlarında elde edilen tüm uyum indeksleri ve hata varyansları birlikte ele alınırsa, tanımlanan modelin ilgili veri setine iyi uyum sağladığı ve göstergelerin ilgili değişkenlerin iyi birer temsilcisi olduğu sonucuna ulaşılmıştır.

Nihai test uygulamasının ardından yapı geçerliğine ilişkin elde edilen kanıtlara ek olarak uygunluk geçerliğine ilişkin inceleme yapılmıştır. Bu kapsamda öğrencilerin nihai test uygulamasından aldıkları toplam puanlar ile bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon hesaplanmış ve bu değer $r=0,65$ ($p<0,05$) bulunmuştur. Hesaplanan korelasyon değerinin anlamlı olması ve öğrencilerin nihai test uygulamasından aldıkları toplam puanlar ile bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasında orta düzeyde bir ilişkiye işaret etmesinden

dolayı elde edilen sonucun uygunluk geçerliğine ilişkin kanıt olabileceği düşünülmektedir.

Araştırmada verilerin çözümlenmesinde ihtiyaç duyulan ağırlıklı puanların hesaplanması için, kısa cevaplı testin nihai uygulanmasından sonra ilk aşamada “1-0” puanlama yöntemi kullanılarak her bir maddeye ilişkin tam doğru cevaba bir puan, diğer tüm cevaplara ise sıfır puan verilmiştir. Daha sonra ağırlıklı puanlama yöntemlerinin her biri için aynı öğrenci yanıtları bu kez her bir madde veya olası cevap için önceden belirlenmiş ağırlıklı puanlar ile puanlanmıştır. Bu araştırmada ele alınan üç ağırlıklı puanlama yönteminin her birine ilişkin ağırlıklı puanların elde edilmesi süreci aşağıda açıklanmıştır.

Madde Güçlük İndekslerine Dayalı Ağırlıklı Puanların Elde Edilmesi

Madde ağırlıklandırma yöntemi için, “1-0” puanlama yöntemi kullanılarak elde edilen puanlara göre hesaplanan madde güçlük indekslerine dayalı olarak maddelerin ağırlıklı puanları belirlenmiştir. Burada farklı güçlüğüne sahip maddelerin puanlama sürecinde aynı ağırlığa sahip olmaması düşüncesi esas alınmıştır. Yani zor olan bir maddenin doğru cevabına atanan puan değerinin, kolay bir maddeye göre daha yüksek olması gerektiği düşünülmektedir.

Kısa cevaplı testin nihai uygulanmasından sonra cevapların “1-0” puanlama yöntemi kullanılarak puanlanmasıyla elde edilen puanlara dayalı olarak hesaplanan madde güçlük indeksleri ve ağırlıklı puanlar Tablo 13’te verilmiştir.

Tablo 13

Nihai Uygulama Formundaki Maddelerin Madde Güçlük İndeksleri ve Ağırlıklı Puanlar

Madde No	p _j	Puan	Madde No	p _j	Puan
1	0,23	4	8	0,40	3
4	0,28	4	15	0,40	3
5	0,33	4	13	0,43	2
7	0,33	4	16	0,44	2
9	0,36	4	10	0,46	2
2	0,38	3	14	0,47	1
11	0,38	3	3	0,52	1
6	0,40	3	12	0,52	1

Tablo 13'e göre testin nihai formundaki maddelerin madde güçlük indeksleri $p_j=0,23$ ve $p_j=0,52$ arasında değişmektedir. Madde güçlüğünden yararlanılarak maddelerin ağırlıklı puanlanmasında; madde güçlük indeksi düşük olan bir maddenin daha zor olmasından dolayı bu maddeye, madde güçlük indeksi yüksek olan bir maddeye göre daha fazla ağırlık verilmesi gerekmektedir. Buna göre madde güçlük indeksleri küçükten büyüğe doğru sıralandığında, en düşük madde güçlük indekslerine sahip birinci, dördüncü, beşinci, yedinci ve dokuzuncu maddeye dört puan verilmiştir. Ayrıca ikinci, on birinci, altıncı, sekizinci ve on beşinci maddelere üç puan; on üçüncü, on altıncı ve onuncu maddelere iki puan verilerek ağırlıklandırma yapılmıştır. Son olarak en büyük madde güçlük indeksi değerlerine sahip on dördüncü, üçüncü ve on ikinci maddelere bir puan verilerek ağırlıklı puanlama uygulanmıştır. Buna göre madde ağırlıklandırma yönteminin kullanıldığı duruma ilişkin veri seti, burada belirlenen ağırlıklı puanlar dikkate alınarak oluşturulmuştur.

Olası Cevapların Deneysel Olarak Ağırlıklandırılmasına Dayalı Ağırlıklı Puanların Elde Edilmesi

Olası cevapların ağırlıklandırılmasıyla ilgili olarak öncelikle tüm maddelere ilişkin tüm cevaplar ayrıntılı biçimde incelenerek cevaplayıcıların kısmi bilgileriyle oluşturduğu ve maddelerin altındaki cevap bölümüne yazdığı olası cevaplar belirlenmiştir. Bu olası cevapların her birine 22, 33 gibi kodlar verilerek tüm öğrenci yanıtları tekrar kayıt edilmiştir. Tablo 14'te testte yer alan bir madde, maddenin tam doğru cevabı, cevaplayıcıların kısmi bilgileriyle oluşturduğu olası cevaplar, test formlarında görülen örnek çözümler ve tüm olası cevaplara verilen kodlar gösterilmiştir.

Tablo 14

Maddelerin Olası Cevaplarına İlişkin Örnek Bir Kodlama

- $\frac{27}{5}$ ile - $\frac{3}{2}$ rasyonel sayıları arasındaki tam sayılar nelerdir?	
Tam doğru cevap: -5, -4, -3, -2 Kod:11	
Kısmi doğru cevap	Örnek çözüm
-4, -3, -2, -1 Kod:22	
-5, -4, -3 Kod:33	
-5, -4, -3, -2, -1, 0, 1 Kod:44	
-5, -4, -3, -2, -1 Kod:55	
-4, -3, -2 Kod:66	
-4, -3 Kod:77	
5, 4, 3, 2 Kod:88	$\frac{27}{5} = 5, \quad \frac{3}{2} = 1,$

Tablo 14'te görülen doğru cevap dışındaki olası cevapların her biri için, cevabın görülme durumu ile toplam test puanı arasındaki nokta çift serili korelasyon katsayıları hesaplanmıştır. Maddelerin cevaplayıcıların kısmi bilgileriyle oluşturduğu olası

cevapları ile toplam puanlar arasındaki nokta çift serili korelasyon katsayıları Tablo 15'te yer almaktadır.

Tablo 15

Maddelerin Olası Cevapları ile Toplam Puanlar Arasındaki Nokta Çift Serili Korelasyon Katsayıları

Madde	Olası Cevap No		Toplam Puan	Madde	Olası Cevap No		Toplam Puan
Madde 1	M1_OC1	r_{pb}	0,44**	Madde 2	M2_OC1	r_{pb}	0,49**
	M1_OC2	r_{pb}	0,46**		M2_OC2	r_{pb}	0,52**
	M1_OC3	r_{pb}	0,47**		M2_OC3	r_{pb}	0,52**
	M1_OC4	r_{pb}	0,52**		M2_OC4	r_{pb}	0,47**
	M1_OC5	r_{pb}	0,48**		M2_OC5	r_{pb}	0,52**
	M1_OC6	r_{pb}	0,45**		M2_OC6	r_{pb}	0,54**
	M1_OC7	r_{pb}	0,46**	Madde 4	M4_OC1	r_{pb}	0,60**
Madde 3	M3_OC1	r_{pb}	0,54**		M4_OC2	r_{pb}	0,57**
	M3_OC2	r_{pb}	0,57**		M4_OC3	r_{pb}	0,50**
	M3_OC3	r_{pb}	0,55**		M4_OC4	r_{pb}	0,50**
	M3_OC4	r_{pb}	0,55**		M4_OC5	r_{pb}	0,62**
Madde 5	M5_OC1	r_{pb}	0,52**	Madde 6	M6_OC1	r_{pb}	0,30**
	M5_OC2	r_{pb}	0,56**		M6_OC2	r_{pb}	0,40**
	M5_OC3	r_{pb}	0,57**		M6_OC3	r_{pb}	0,46**
	M5_OC4	r_{pb}	0,61**		M6_OC4	r_{pb}	0,43**
	M5_OC5	r_{pb}	0,56**		M6_OC5	r_{pb}	0,44**
	M5_OC6	r_{pb}	0,56**	Madde 8	M8_OC1	r_{pb}	0,48**
Madde 7	M7_OC1	r_{pb}	0,50**		M8_C2	r_{pb}	0,47**
	M7_OC2	r_{pb}	0,54**		M8_OC3	r_{pb}	0,57**
	M7_OC3	r_{pb}	0,59**		M8_OC4	r_{pb}	0,56**
	M7_OC4	r_{pb}	0,42**		M8_OC5	r_{pb}	0,57**
	M7_OC5	r_{pb}	0,60**		M8_OC6	r_{pb}	0,58**
	M7_OC6	r_{pb}	0,55**		M10_OC1	r_{pb}	0,56**
	M7_OC7	r_{pb}	0,50**	Madde 10	M10_OC2	r_{pb}	0,65**
Madde 9	M9_OC1	r_{pb}	0,63**		M10_OC3	r_{pb}	0,66**
	M9_OC2	r_{pb}	0,41**		M10_OC4	r_{pb}	0,63**
	M9_OC3	r_{pb}	0,62**	Madde 12	M12_OC1	r_{pb}	0,61**
	M9_OC4	r_{pb}	0,61**		M12_OC2	r_{pb}	0,49**
	M9_OC5	r_{pb}	0,61**		M12_OC3	r_{pb}	0,56**
Madde 11	M11_OC1	r_{pb}	0,57**		M12_OC4	r_{pb}	0,56**
	M11_OC2	r_{pb}	0,59**	Madde 14	M14_OC1	r_{pb}	0,56**
	M11_OC3	r_{pb}	0,56**		M14_OC2	r_{pb}	0,58**
	M11_OC4	r_{pb}	0,61**		M14_OC3	r_{pb}	0,59**
	M11_OC5	r_{pb}	0,57**				

(Devam ediyor)

Tablo 15 (Devam)

Maddelerin Olası Cevapları ile Toplam Puanlar Arasındaki Nokta Çift Serili Korelasyon Katsayıları

Madde	Olası Cevap No	Toplam Puan	Madde	Olası Cevap No	Toplam Puan
Madde 13	M13_OC1	r_{pb} 0,60**	Madde 16	M14_OC4	r_{pb} 0,54**
	M13_OC2	r_{pb} 0,66**		M16_OC1	r_{pb} 0,51**
	M13_OC3	r_{pb} 0,60**		M16_OC2	r_{pb} 0,52**
	M13_OC4	r_{pb} 0,65**		M16_OC3	r_{pb} 0,57**
	M13_OC5	r_{pb} 0,67**		M16_OC4	r_{pb} 0,50**
Madde 15	M15_OC1	r_{pb} 0,65**	Madde 16	M16_OC5	r_{pb} 0,48**
	M15_OC2	r_{pb} 0,61**			
	M15_OC3	r_{pb} 0,62**			
	M15_OC4	r_{pb} 0,64**			

**p<0,01

Tablo 15'te görülen her bir maddenin her bir olası cevabı için belirlenen nokta çift serili korelasyon katsayıları incelenerek öncelikle her bir madde için büyüktен küçüğe doğru sıralanmıştır. Her bir madde için tam doğru cevabın ağırlıklı puanı dört puan olarak belirlendikten sonra, ilgili madde için Tablo 15'te görülen en büyük nokta çift serili korelasyon katsayısına sahip olası cevaba 3 puan verilmiştir. Hesaplanan nokta çift serili korelasyon katsayı bakımından ikinci ve üçüncü sırada yer alan olası cevaplar ise sırasıyla iki ve bir puan verilerek deneysel ağırlıklandırma işlemi yapılmıştır. Son olarak bu cevapların dışındaki tüm cevaplar için sıfır puan verilerek puanlama yapılmıştır. Bir maddede birden fazla olası cevabın aynı nokta çift serili korelasyon katsayısına sahip olması durumunda bu olası cevapların hepsi için aynı ağırlıklandırma işlemi yapılmıştır. Örneğin madde 9 için ağırlıklandırma yapılırken tam doğru cevaba dört puan, birinci olası cevaba üç puan, üçüncü olası cevaba iki puan, dördüncü ve beşinci olası cevaplara ise bir puan verilerek ağırlıklandırma yapılmıştır. Daha sonra her bir olası cevap için belirlenen ağırlıklı puana göre puanlama işlemi yapılmıştır. Böylece olası cevapların deneysel olarak ağırlıklandırılmasına dayalı puanlama yöntemine ilişkin veri seti hazırlanmıştır.

Çoklu Ağırlıklandırmaya Dayalı Ağırlıklı Puanların Elde Edilmesi

Çoklu ağırlıklandırmaya dayalı puanların elde edilmesi için, bir önceki adımda olası cevapların deneysel olarak ağırlıklandırılmasına dayalı puanlama yöntemiyle

hazırlanan veri seti kullanılmıştır. Böylece olası cevapların deneysel olarak ağırlıklandırıldığı yöntemin ve olası cevaplarla birlikte maddelerin de ağırlıklandırıldığı bir diğer yöntemin testin psikometrik özellikleri bakımından performansının incelenmesi hedeflenmiştir. Çoklu ağırlıklar elde edilirken puanlama yönteminin önerildiği araştırmada sıralanan adımlar takip edilmiştir (Haladyna ve Sympson, 1988; Sympson, 1993). Öncelikle cevaplayıcıların her birine, testte yer alan maddelerdeki doğru oranlarını gösteren $[0,1]$ aralığında bir geçici puan atanmıştır. Bu doğru oranları geleneksel puanlama yöntemi kullanıldığında hesaplanan toplam puanlar aracılığıyla hesaplanmıştır. İkinci aşamada doğru oranlarının “yüzdelik sıralama puanı”na dönüştürülmesi için, her bir cevaplayıcının bir önceki adımda hesaplanan doğru oranına eşit ve daha küçük doğru oranına sahip cevaplayıcıların sayısı, tüm cevaplayıcıların sayısına oranlanıp 100 ile çarpılmıştır. Bu çalışmada hesaplanan yüzdelik sıralama puanlarının 7,51 ile 100 arasında değiştiği görülmüştür. Üçüncü adımda her bir madde için, her bir olası cevabı oluşturan cevaplayıcıların bir önceki adımda hesaplanan yüzdelik sıralama puanlarının ortalaması hesaplanmıştır. Bu aşamada tam doğru cevap için hesaplanan yüzdelik sıralama puanlarının ortalamasının o madde için tüm cevaplayıcıların yüzdelik sıralama puanlarının ortalamasından küçük olup olmadığı kontrol edilmiştir. Testteki 16 maddenin hiçbirinde böyle bir durumla karşılaşmadığından maddelerin tamamına ilişkin ağırlıklandırma işlemi yapılmıştır.

Olası cevapların her biri için bu cevabı oluşturan cevaplayıcıların yüzdelik sıralama puanlarının ortalaması kullanılarak tüm maddeler ve tüm olası cevaplar için çoklu ağırlıklar hesaplanmıştır. Çoklu ağırlıklar hesaplanırken bir yanlış cevabın çoklu ağırlığının doğru cevabın çoklu ağırlığına eşit veya doğru cevabın çoklu ağırlığından büyük olup olmadığı kontrol edilmiştir. Bu araştırma kapsamında tüm maddeler ve tüm olası cevaplar için hesaplanan çoklu ağırlıklar Tablo 16’da verilmiştir.

Tablo 16

Maddelerin Olası Cevapları için Hesaplanan Çoklu Ağırlıklar

Madde	Olası Cevap	Çoklu Ağırlık	Madde	Olası Cevap	Çoklu Ağırlık
Madde 1	OC1	50	Madde 9	OC1	51
	OC2	57		OC2	52
	OC3	54		OC3	57
	OC4	58		OC4	56
	OC5	62		OC5	58
Madde 2	OC1	43	Madde 10	OC1	47
	OC2	51		OC2	48
	OC3	52		OC3	48
	OC4	54		OC4	52
	OC5	64		OC5	61
Madde 3	OC1	49	Madde 11	OC1	48
	OC2	50		OC2	49
	OC3	46		OC3	49
	OC4	53		OC4	55
	OC5	60		OC5	61
Madde 4	OC1	50	Madde 12	OC1	43
	OC2	57		OC2	43
	OC3	54		OC3	45
	OC4	54		OC4	50
	OC5	58		OC5	60
Madde 5	OC1	52	Madde 13	OC1	47
	OC2	50		OC2	50
	OC3	58		OC3	51
	OC4	55		OC4	52
	OC5	59		OC5	58
Madde 6	OC1	48	Madde 14	OC1	51
	OC2	52		OC2	52
	OC3	54		OC3	52
	OC4	56		OC4	55
	OC5	58		OC5	58
Madde 7	OC1	42	Madde 15	OC1	51
	OC2	48		OC2	52
	OC3	48		OC3	55
	OC4	51		OC4	57
	OC5	58		OC5	58
Madde 8	OC1	53	Madde 16	OC1	46
	OC2	53		OC2	48
	OC3	54		OC3	49
	OC4	57		OC4	51
	OC5	63		OC5	57

Tablo 16’da verilen olası cevaplardan birincisi yanlış cevabı, beşincisi ise tam doğru cevabı göstermektedir. İkinci, üçüncü ve dördüncü olası cevaplar da cevaplayıcıların kısmi bilgileriyle oluşturduğu olası cevapları ifade etmektedir. Olası

cevapların her birine ilişkin Tablo 16’da verilen çoklu ağırlıklar hesaplandıktan sonra bunlara göre cevaplayıcıların ağırlıklandırılmış toplam puanları hesaplanmıştır. Bunun için cevaplayıcıların her bir maddede oluşturduğu olası cevapların çoklu ağırlıklarının ortalaması alınmıştır. Böylece çoklu ağırlıklandırmaya dayalı puanlama yöntemine ilişkin veri seti de hazırlanmıştır.

Verilerin Çözümlemesi

Araştırmada geleneksel ve ağırlıklı puanlama yöntemleriyle elde edilen puanlar araştırmanın genel amacına ve alt amaçlarına uygun şekilde analiz edilmiştir. Testin uygulanmasından sonra ilk aşamada “1-0” puanlama yöntemi kullanılarak her bir maddeye ilişkin tam doğru cevaba bir puan, diğer tüm cevaplara ise sıfır puan verilmiş ve bu puanlama yöntemine ilişkin veri seti elde edilmiştir. Daha sonraki aşamalarda üç ayrı ağırlıklı puanlama yöntemine göre ağırlıklandırmalar yapılarak çok kategorili verilerden oluşan üç veri seti elde edilmiştir.

Araştırmanın genel amacı doğrultusunda kullanılan kısa yanıtlı testin puanlanmasında önce “1-0” puanlama, daha sonra ağırlıklı puanlama yöntemlerinin kullanılması durumunda klasik test kuramına göre test puanlarının geçerliğine ilişkin analizler yapılmıştır. Ayrıca aynı puanlama durumlarında test puanlarına ilişkin güvenilirlik kestirimleri elde edilmiştir.

Araştırmada geliştirilen kısa yanıtlı test aracılığıyla elde edilen puanlara ilişkin yapı geçerliği kapsamında açıklayıcı faktör analizinin sonucunda açıklanan varyans değerleri, öz değerler ve faktör yükleri elde edilmiştir. GP yönteminin kullanıldığı durumda elde edilen ikili veriyle tetrakorik, ağırlıklı puanlama yöntemlerinin kullanılması durumunda elde edilen çok kategorili veriyle polikorik korelasyon matrisine dayalı olarak AFA yapılmıştır. AFA’da kullanılan korelasyon matrisi, kategori sayısına göre farklılaşabilmektedir. Örneğin iki kategoriye sahip veri için tetrakorik, daha çok kategoriye sahip veri için polikorik korelasyon matrisiyle analiz yapılırken kategori sayısı arttıkça verinin sürekli olduğu kabulüyle Pearson korelasyon matrisi kullanılabilir. Finney ve DiStefano’e göre (2013) kategori sayısı altı ve daha çok ise veri sürekli olarak kabul edilebilmektedir. Tabachnik ve Fidell (2013) ise kategori sayısının yedi ve üzerinde olduğu zaman verinin sürekli kabul edileceğini ifade etmektedir. AFA’da kullanılan veri sürekli ise Pearson korelasyon matrisi, veri

kategorik olduğunda ise tetrakorik (iki kategoriye sahip veride) veya polikorik (çok kategorili veride) korelasyon matrisi kullanılmalıdır (Cooper, 2019).

AFA yapılırken faktör çıkarma (faktörleştirme) yaklaşımı olarak temel bileşen analizi yöntemi kullanılmıştır. AFA’da basit anlamda veri setinin daha az değişkenle özetlenmesi amaçlanıyorsa “temel bileşen analizi” yöntemi önerilmektedir. Eğer yapı geçerliği incelenecek ölçme aracı, önceden net olarak tanımlanmış ve bilinen bir yapıyı ölçüyorsa veri setinin ampirik bir özetinin elde edilmesi için TBA kullanılabilir (Tabachnik & Fidell, 2013). AFA ve paralel analiz yapılırken R (R core team, 2022)’da bulunan psych (Revelle, 2022) ve “hornpa” (Huang, 2015) paketleri kullanılmıştır.

AFA sonuçları ayrı ayrı incelendikten sonra, belirlenen açıklanan varyans değerlerinin ve madde faktör yüklerinin farklı puanlama durumlarında nasıl değişmekte olduğu incelenmiştir. Elde edilen faktör yük değerleri arasındaki ilişkinin incelenmesinde Spearman Brown sıra farkları korelasyon katsayısı hesaplanmıştır. Bunun yanında incelenen puanlama durumlarında belirlenen faktör yük değerleri arasındaki farkların anlamlılığını incelemek için ikili karşılaştırmalarda Mann-Whitney U testi, ağırlıklı puanlama yöntemlerine ilişkin karşılaştırmada ise Kruskal Wallis H testi kullanılmıştır. Verilerin analize hazırlığı ve maddelerin faktör yük değerlerine ilişkin incelemelerde IBM SPSS Statistics 22 kullanılmıştır.

Yapı geçerliğine ilişkin incelemelerin devamında kısa cevaplı testten elde edilen puanlara dayalı olarak DFA yapılmıştır. DFA’da çok değişkenli normallik gereğinin sağlanmaması nedeniyle tahmin yöntemi olarak Robust Maximum Likelihood (MLR) kullanılmıştır. Verilerin analizi sonucunda DFA için hesaplanan uyum indeksleri (RMSEA, AGFI, CFI, NNFI, Ki-kare/sd) elde edilmiştir. Bu analiz için LISREL 8.7 yazılımı kullanılmıştır. Elde edilen uyum indekslerinin puanlama yöntemlerine göre karşılaştırılmasının ardından her bir yöntem için yapılan DFA sonuçlarında görülen hata varyansları da karşılaştırılmıştır. Farklı puanlama durumlarında elde edilen hata varyans değerleri arasındaki ilişkinin incelenmesi amacıyla Spearman Brown sıra farkları korelasyon katsayısı hesaplanmıştır. Bunun yanında incelenen puanlama durumlarında belirlenen hata varyansları arasındaki farkların anlamlılığını incelemek için ikili karşılaştırmalarda Mann-Whitney U testi, ağırlıklı puanlama yöntemlerine ilişkin karşılaştırmada ise Kruskal Wallis H testi yapılmıştır.

Araştırmada incelenen bir diğer geçerlik türü olan uygunluk geçerliği için ölçüt puan olarak öğrencilerin matematik dersine ilişkin bir önceki öğretim yılına ait yılsonu puanları alınmıştır. Bu yıl sonu puanları araştırmada incelenen ölçme sonuçlarından

daha önce elde edildiği için uygunluk geçerliğine ilişkin sonuçlar verilmiştir (Baykul, 2015). Bu doğrultuda her bir puanlama yönteminin kullanıldığı durum için elde edilen cevaplayıcıların toplam puanları ile bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki ilişki Pearson korelasyon katsayısı yardımıyla incelenmiştir. Son olarak ağırlıklı puanlama yöntemlerinin her biriyle elde edilen puanların matematik dersine ilişkin not ortalamasıyla aralarındaki ilişkilerin nasıl değişmekte olduğu incelenmiştir. İki farklı puanlama durumuna ilişkin elde edilen korelasyon değerleri arasında manidar bir farklılık olup olmadığını araştırmak için bağımlı iki korelasyon katsayısının karşılaştırılmasına yönelik z-testi yapılmıştır (Meng, Rosenthal & Rubin, 1992). Bu teste ilişkin z istatistiği aşağıdaki gibi hesaplanmaktadır:

$$Z = (zr_{yx1} - zr_{yx2}) \sqrt{\frac{N - 3}{2(1 - r_{x1x2})h}}$$

Burada,

$$zr_{yx1} = 0,5 \left(\ln \left((1 + r_{yx1}) - (1 - r_{yx1}) \right) \right)$$

$$zr_{yx2} = 0,5 \left(\ln \left((1 + r_{yx2}) - (1 - r_{yx2}) \right) \right)$$

$$h = \frac{1 - f\bar{r}}{1 - \bar{r}^2}$$

$$\bar{r} = (r_{yx1}^2 + r_{yx2}^2)/2$$

$$f = \frac{1 - r_{x1x2}}{2(1 - \bar{r}^2)}$$

z testi ile bu şekilde hesaplanan değer z tablo değeri olan 1,96'dan büyük olması durumunda, karşılaştırılan iki korelasyon katsayısının %95 güven aralığında istatistiksel olarak anlamlı biçimde birbirinden farklı olduğu belirlenmiştir.

Araştırmada güvenirlikle ilgili incelemelerde, geleneksel ve ağırlıklı puanlama yöntemlerinden her birinin kullanıldığı durum için Cronbach α ve McDonald'ın ω güvenirlik katsayısı hesaplanmıştır. Söz konusu tüm puanlama durumlarında elde edilen puanlara ilişkin güvenirlik katsayıları hesaplanarak bu güvenirlik katsayıları arasında anlamlı bir fark olup olmadığı incelenmiştir. Güvenirlik katsayılarının hesaplanması için R (R core team, 2022)'da bulunan psych (Revelle, 2019) ve ltm (Rizopoulos, 2006) paketleri kullanılmıştır.

Cronbach α güvenirlik katsayısı hesaplanırken testin maddeleri k tane paralel ölçme olarak kabul edilmektedir. Testin maddeleri arasındaki korelasyon veya testin maddeleri ile test-toplam puanları arasındaki korelasyon arttıkça yüksek değer

vermektedir. Hata kaynağı olarak testin maddeleri arasındaki paralellik düzeyi dikkate alınmaktadır. Çok kategorili puanlama yöntemleri kullanılarak elde edilen puanlar için kullanılabilir (Büyüköztürk, Çakmak, Akgün, Karadeniz ve Demirel, 2012; Crocker ve Algina, 1986).

Cronbach α güvenirlik katsayısının formülü aşağıdaki gibi ifade edilmektedir:

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum s_i^2}{S_x^2} \right)$$

Cronbach alfa formülündeki K madde sayısını, s_i^2 madde varyansını ve S_x^2 test puanları varyansını göstermektedir.

Maddelerin faktör yüklerinin eşit olmadığı durumlarda Cronbach α katsayısının yanlış sonuçlar verebilmesinden (Yurdugül, 2006) dolayı güvenirlikle ilgili incelemelerde, geleneksel ve ağırlıklı puanlama yöntemlerinden her birinin kullanıldığı durum için McDonald'ın ω güvenirlik katsayısı da hesaplanmıştır. McDonald'ın ω katsayısında maddelerin faktör yükleri ayrı ayrı hesaba katılmaktadır. Maddelerin standartlaştırılmamış faktör yükleri ile ifade edilen ω katsayısının formülü aşağıda verilmiştir.

$$\omega = \frac{(\sum_{i=1}^k \lambda)^2}{(\sum_{i=1}^k \lambda)^2 + \sum_{i=1}^k \Psi_i}$$

Formülde λ maddelerin standartlaştırılmamış faktör yüklerini ve Ψ maddelere ilişkin hata varyanslarını ifade etmektedir. Puanlama yöntemlerinden her birinin kullanıldığı durum için Cronbach α ve McDonald'ın ω güvenirlik katsayısı hesaplandıktan sonra bu katsayıların birbirleri ile karşılaştırılması için Fisher'ın z-testi kullanılmıştır. z_1 birinci yönteme ait Fisher z değeri, z_2 ikinci yönteme ait Fisher z değeri, n_1 ve n_2 karşılaştırılan durumlardaki kişi sayılarını göstermek üzere, Fisher'ın z-testi sonucu aşağıdaki gibi hesaplanmaktadır:

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Yukarıdaki formüle göre elde edilen z istatistik değerinin z tablo değeri olan 1,96'dan büyük olması durumunda, karşılaştırılan iki güvenirlik katsayısının %95 güven aralığında istatistiksel olarak anlamlı biçimde birbirinden farklı olduğu söylenebilmektedir (Kenny, 1987).

Araştırmada kullanılan kısa yanıtli testin psikometrik özellikleri KTK ve MTK çerçevesinde incelenmiştir. MTK'ya ilişkin incelemelerde öncelikle puanların, MTK'ya ilişkin tek boyutluluk ve yerel bağımsızlık gereklerini sağlama durumu incelenmiştir. Tek boyutluluk için elde edilen AFA sonuçları dikkate alınmıştır. Yerel bağımsızlıkla ilgili olarak ise Yen'in Q3 istatistiği (Yen, 1993) kullanılmıştır. Yen'in Q3 istatistiği değerleri için kritik kesme noktası 0.30 olarak kabul edilmiştir (Røe, Damsgård, vd, 2014). Daha sonra ikili veri seti için 1 parametrelili lojistik model (1PLM) ve 2 parametrelili lojistik model (2PLM), çok kategorili veri seti için MTK modellerinden geliştirilmiş kısmi puan modeli (GKPM) ve kademeli tepki modeli (KTM) için kalibrasyonlar yapılarak en uygun MTK modeline karar verilmiştir. Bu sırada madde model uyumu için S_{χ^2} istatistiği ve RMSEA değerlerinden yararlanılmıştır. MTK kalibrasyonlarında R (R core team, 2022)'da bulunan mirt (Chalmers, 2012) ve ltm (Rizopoulos, 2006) paketleri kullanılmıştır.

Araştırmadaki puanlama sürecinde önce "1-0" puanlama, daha sonra OCAP yönteminin kullanılmasıyla elde edilen orijinal veri kullanılarak alınan sonuçların farklı veri setlerinin kullanılması durumundaki genellenebilirliğinin incelenmesi amacıyla R (R core team, 2022)'da her iki puanlama yöntemine ilişkin 100'er tane 300 kişilik veri seti oluşturulmuştur. Bu veri setleri üretilirken orijinal veri setinden rastgele çekilen bir birey tekrar seçilmeyecek şekilde fonksiyon yazılmıştır. Veri setleri oluşturulduktan sonra döngü kullanılarak AFA yapılırken R (R core team, 2022)'da bulunan psych (Revelle, 2022) ve DFA yapılırken lavaan (Rosseel, 2012) paketleri kullanılmıştır. Güvenirlik katsayılarının hesaplanması için R (R core team, 2022)'da bulunan psych (Revelle, 2019) ve ltm (Rizopoulos, 2006) paketleri kullanılmıştır.

GP ve OCAP yöntemlerinin kullanıldığı veri setleri için AFA ile belirlenen açıklanan varyans değerlerinin ve hesaplanan güvenirlik katsayılarının kullanılan puanlama yöntemine göre manidar bir fark gösterip göstermediği bağımsız gruplar için t-testi ile incelenmiştir. İncelenen puanlama durumlarında belirlenen ortalama faktör yük değerleri ile ortalama hata varyansları arasındaki farkların anlamlılığını incelemek için ise Mann-Whitney U testi kullanılmıştır.

BÖLÜM 4

BULGULAR VE YORUMLAR

Bu bölümde araştırmanın genel amacı ve alt problemlerine ilişkin bulgular ile bu bulgulara dayalı yorumlar alt problem sırasına göre verilmiştir. Bulgu ve yorumlara geçilmeden önce geleneksel puanlama (GP), madde güçlük indekslerine dayalı ağırlıklı puanlama (MGİAP), olası cevapların ağırlıklandırılmasına dayalı puanlama (OCAP) ve çoklu ağırlıklandırmaya dayalı puanlama (ÇAP) yöntemleriyle elde edilen betimsel istatistikler Tablo 17’de sunulmuştur.

Tablo 17

Puanlama Yöntemleriyle Elde Edilen Betimsel İstatistikler

Yöntem	N	Min	Max	$\bar{X}$	s_x	Çarpıklık	Basıklık
GP	586	0	16	6,12	4,42	0,39	-0,95
MGİAP		0	44	15,46	12,01	0,53	-0,80
OCAP		0	64	29,07	18,45	0,19	-1,02
ÇAP		49	60	54,14	2,93	0,21	-1,09

Tablo 17’de verilen puanlama yöntemlerinin her birine ilişkin betimsel istatistikler incelendiğinde OCAP yöntemi kullanıldığında elde edilen puanların maksimum değerinin diğer durumlara göre daha büyük olduğu görülmüştür. ÇAP yoluyla elde edilen puanların ortalaması diğer durumlara göre en büyük, standart sapması ise diğer puanlama durumlarına göre en küçük değeri almaktadır. Çarpıklık ve basıklık değerleri incelendiğinde ise tüm durumlarda elde edilen değerlerin (-1,1) civarında değer aldığı söylenebilmektedir.

Birinci Alt Probleme İlişkin Bulgular ve Yorumlar

Birinci alt problemle ilgili olarak araştırmada kullanılan kısa yanıtlı testin puanlanmasında önce “1-0” puanlama, daha sonra olası cevapların ağırlıklandırılmasına dayalı ağırlıklı puanlama yönteminin kullanılması durumunda klasik test kuramına göre öncelikle test puanlarının geçerliğinin nasıl değiştiği incelenmiştir. Daha sonra ise aynı

durumda test puanlarına ilişkin güvenilirlik kestirimlerin nasıl değişmekte olduğuna bakılmıştır.

Araştırmada geliştirilen kısa yanıtlı testin puanlanmasında önce GP, daha sonra OCAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin yapı geçerliği kapsamında AFA’da açıklanan varyans değerlerinin, elde edilen faktör yüklerinin ve DFA’da görülen uyum indekslerinin (RMSEA, AGFI, NNFI, CFI ve ki-kare/sd) nasıl değişmekte olduğu incelenmiştir.

GP yönteminin kullanılması durumunda elde edilen puanların yapı geçerliğine ilişkin tüm bilgiler Verilerin Toplanması başlıklı bölümde verilmiştir. OCAP yöntemiyle ilgili olarak nihai test uygulamasında elde edilen öğrenci yanıtları, olası cevapların deneysel olarak ağırlıklandırılması yaklaşımıyla puanlanmıştır. Daha sonra yanıtların bu yöntemle göre puanlanmasıyla elde edilen puanlara dayalı olarak AFA yapılmıştır. AFA sonucuna göre elde edilen özdeğerler ile yapının tek faktörlü olmasına dayanak oluşturmak amacıyla yapılan paralel analiz sonucunda elde edilen özdeğerlerin bulunduğu özdeğer grafiği ekte sunulmuştur. Analiz sonuçları incelendiğinde 16 maddenin tek faktör altında toplandığı ve tek faktöre ilişkin öz değerin 8,76 olduğu belirlenmiştir. Ayrıca analizdeki tek faktörün varyansın toplam %54,8’ini açıklamış olduğu görülmüştür.

GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin AFA sonuçları ayrı ayrı incelendikten sonra, belirlenen açıklanan varyans değerlerinin ve madde faktör yüklerinin bu iki puanlama durumunda nasıl değişmekte olduğu incelenmiştir. Tüm veri setiyle GP yönteminin kullanılması durumunda elde edilen puanların AFA sonucunda açıklanan varyans değerinin 0,491 olduğu görülmüştür. OCAP yöntemiyle elde edilen puanlara AFA yapıldığında ise açıklanan varyans oranı 0,548 bulunmuştur. Buna göre cevaplayıcıların kısmi bilgileri dikkate alınarak bir ağırlıklandırma işlemi yapıldığında, ağırlıklandırmanın yapılmadığı duruma göre AFA sonucunda açıklanan varyans miktarının arttığı söylenebilmektedir.

GP ve OCAP yöntemlerinin kullanıldığı durumlarda elde edilen puanlara ilişkin AFA sonucunda elde edilen faktör yükleri ise Tablo 18’de birlikte sunulmuştur.

Tablo 18

GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Maddelerin Faktör Yük Değerleri

Madde No	Faktör Yük Değeri (GP)	Faktör Yük Değeri (OCAP)
m1	0,61	0,71
m2	0,64	0,70
m3	0,68	0,69
m4	0,72	0,80
m5	0,69	0,75
m6	0,51	0,52
m7	0,74	0,73
m8	0,68	0,72
m9	0,74	0,80
m10	0,80	0,81
m11	0,72	0,76
m12	0,70	0,75
m13	0,80	0,83
m14	0,68	0,73
m15	0,76	0,80
m16	0,63	0,69

Tablo 18’de görüldüğü gibi GP yönteminin kullanılması durumunda elde edilen puanların AFA sonucunda görülen faktör yük değerleri 0,51 ile 0,80 arasında değişmektedir. OCAP yöntemiyle elde edilen puanlara dayalı olarak AFA yapıldığında ise faktör yük değerleri 0,52 ile 0,83 arasında değişmektedir. Öncelikle OCAP yönteminin kullanıldığı durumda, yedinci madde dışındaki maddelerin her biri için elde edilen faktör yük değerlerinin GP yönteminin kullanıldığı durumda elde edilen değerlerden daha büyük olduğu görülmektedir. Bu iki puanlama durumunda elde edilen faktör yük değerleri arasındaki ilişkinin incelenmesi için Spearman Brown sıra farkları korelasyon katsayısı hesaplanmıştır. Bu korelasyon değerine ($r_s = 0,91$) göre, iki puanlama yöntemi arasında faktör yükleri bakımından pozitif yönde yüksek düzeyde uyum olduğu görülmektedir. Bunun yanında incelenen puanlama durumlarında belirlenen faktör yük değerleri arasındaki farkların anlamlılığını incelemek için Mann-Whitney U testi kullanılmıştır. Tablo 19’da AFA sonucunda görülen faktör yük değerlerinin kullanılan ağırlıklı puanlama yöntemine göre karşılaştırılmasına ilişkin Mann-Whitney U testi sonuçları verilmiştir.

Tablo 19

GP ve OCAP Yöntemlerinde AFA Sonucunda Elde Edilen Faktör Yük Değerlerinin Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları

Yöntem	n	Sıra Ortalaması	Sıra Toplamı	z	p
GP	16	13,13	210	-2,04	0,041
OCAP	16	19,88	318		
Toplam	32				

Tablo 19 incelendiğinde AFA sonucunda elde edilen faktör yük değerlerinin kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaştığı görülmektedir ($z = -2,04, p < 0,05$). Sıra ortalamaları dikkate alındığında OCAP yöntemi kullanıldığında elde edilen faktör yük değerlerinin GP yönteminin kullanıldığı durumda elde edilen değerlerden anlamlı olarak daha büyük olduğu görülmektedir.

AFA sonuçlarının incelenmesinin ardından kısa yanıtlı testin yanıtlarının GP ve OCAP yöntemleriyle puanlanması durumlarında elde edilen puanlara dayalı olarak DFA yapıldığında ortaya çıkan sonuçlar karşılaştırılmıştır. OCAP yöntemiyle ilgili olarak ölçme aracındaki 16 maddeye ilişkin ölçümlerin faktör yapısı, tek faktörlü ölçme modeli tanımlanarak DFA ile test edilmiştir. Tanımlanan tek faktörlü modelin test edilmesiyle elde edilen yol şeması ekte sunulmuştur. OCAP yönteminin kullanıldığı durumda, ölçme aracındaki verilere ilişkin yapılan DFA sonuçlarında elde edilen tüm uyum indeksleri ve hata varyansları birlikte ele alındığında, tanımlanan modelin ilgili veri setine iyi uyum sağladığı ve göstergelerin ilgili değişkenlerin iyi birer temsilcisi olduğu sonucuna ulaşılmıştır. GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin DFA sonuçları ayrı ayrı incelendikten sonra, elde edilen uyum indeks değerlerinin kullanılan puanlama yöntemine göre nasıl değişmekte olduğu incelenmiştir. GP ve OCAP yöntemlerinin her birinden elde edilen puanlara uygulanan DFA sonucunda görülen uyum indeksleri birlikte Tablo 20’de yer almaktadır.

Tablo 1

<i>GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Uyum İndeksleri</i>					
Puanlama Yöntemi	RMSEA	CFI	AGFI	NNFI	χ^2/sd
GP	0,044	0,98	0,94	0,98	2,15
OCAP	0,055	0,98	0,92	0,98	2,72

Tablo 20’de verilen DFA sonuçlarına göre, GP yönteminin kullanıldığı durumda elde edilen RMSEA değeri, olası cevapların ağırlıklandırıldığı durumda elde edilen RMSEA değerinden daha küçüktür. GP ve OCAP yönteminin kullanıldığı durumlarda CFI ve NNFI değerinde değişiklik olmadığı görülmüştür. OCAP yönteminin kullanılması durumunda kestirilen AGFI değeri ise GP yönteminin kullanıldığı durumda elde edilen AGFI değerine göre daha küçüktür. Son olarak χ^2/sd oranları incelendiğinde, GP yöntemi kullanıldığında bu oranın OCAP yönteminin kullanıldığı duruma göre daha düşük olduğu belirlenmiştir.

GP ve OCAP yöntemlerinin kullanıldığı durumlarda elde edilen puanlara dayalı olarak yapılan DFA sonucunda görülen hata varyansları Tablo 21’de birlikte verilmiştir.

Tablo 21

GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda DFA Sonucunda Görülen Hata Varyansları

Madde No	Hata Varyans Değeri (GP)	Hata Varyans Değeri (OCAP)
m1	0,83	0,62
m2	0,76	0,66
m3	0,73	0,70
m4	0,72	0,55
m5	0,72	0,60
m6	0,86	0,83
m7	0,67	0,64
m8	0,71	0,65
m9	0,67	0,53
m10	0,57	0,51
m11	0,68	0,57
m12	0,70	0,59
m13	0,57	0,50
m14	0,73	0,64
m15	0,64	0,58
m16	0,76	0,66

Tablo 21’de görüldüğü gibi GP yönteminin kullanılması durumunda elde edilen puanlara dayalı olarak DFA yapıldığında görülen hata varyansları 0,57 ile 0,86 arasında değer almaktadır. OCAP yöntemiyle elde edilen puanlara dayalı olarak DFA yapıldığında hesaplanan hata varyansları ise 0,51 ile 0,83 arasında değişmektedir.

Öncelikle OCAP yönteminin kullanıldığı durumda, maddelerin her biri için hesaplanan hata varyanslarının GP yönteminin kullanıldığı durumda elde edilen değerlerden daha küçük olduğu görülmüştür. Daha sonra bu iki puanlama durumunda elde edilen hata varyans değerleri arasındaki ilişkinin incelenmesi için Spearman Brown sıra farkları korelasyon katsayısı hesaplanmıştır. Bu korelasyon değerine ($r_s = 0,78$) göre, iki puanlama yöntemi arasında hesaplanan hata varyansları bakımından pozitif yönde yüksek düzeyde uyum olduğu görülmektedir. Bunun yanında incelenen puanlama durumlarında belirlenen hata varyansları arasındaki farkların anlamlılığını incelemek amacıyla Mann-Whitney U testi kullanılmıştır. Tablo 22’de DFA sonucunda elde edilen hata varyanslarının kullanılan ağırlıklı puanlama yöntemine göre karşılaştırılmasına ilişkin Mann-Whitney U testi sonuçları verilmiştir.

Tablo 22

GP ve OCAP Yöntemlerinde DFA Sonucunda Elde Edilen Hata Varyanslarının Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları

Yöntem	n	Sıra Ortalaması	Sıra Toplamı	z	p
GP	16	21,75	348	-3,17	0,002
OCAP	16	11,25	180		
Toplam	32				

Tablo 22 incelendiğinde DFA sonucunda elde edilen hata varyanslarının kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaştığı görülmektedir ($z = -3,17, p < 0,05$). Sıra ortalamaları dikkate alındığında OCAP yöntemi kullanıldığında elde edilen hata varyanslarının GP yönteminin kullanıldığı durumda elde edilen değerlerden anlamlı olarak daha küçük olduğu görülmektedir.

GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin yapı geçerliği kanıtlarının karşılaştırılmasıyla ilgili bulgulara göre, genel olarak OCAP yönteminin test puanlarının yapı geçerliği bakımından daha iyi sonuçlar verdiği söylenebilmektedir. Bu araştırmada kullanılan kısa cevaplı bir testte olası cevapların deneysel olarak ağırlıklandırılmasına dayalı puanlama yaklaşımına göre yapı geçerliğiyle ilgili elde edilen sonuçlar, Erdem, Ertuna ve Doğan (2016)’ın çalışmalarının sonuçlarıyla AFA sonuçları bakımından farklılaşmakta ve DFA sonuçları bakımından örtüşmektedir. Erdem, Ertuna ve Doğan (2016)’ın araştırmalarında, bu araştırmanın aksine, AFA sonuçlarına göre deneysel puanlamanın açıklanan varyans oranında düşmeye sebep

olduğu gözlenmiştir. DFA sonuçlarında ise, benzer şekilde uyum indekslerinin deneysel puanlamada önemli düzeyde farklılaşmadığı görülmüştür.

Araştırmada puanlama sürecinde önce GP, daha sonra OCAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin yapı geçerliğinin incelenmesinin ardından uygunluk geçerliğine yönelik bulgular incelenmiştir. Uygunluk geçerliği kapsamında iki ayrı puanlama durumunda elde edilen puanlarla öğrencilerin bir önceki öğretim yılına ait matematik dersi yıl sonu puanı arasındaki ilişkilerin nasıl değişmekte olduğu incelenmiştir. Bu doğrultuda GP yönteminin kullanıldığı birinci durumda elde edilen toplam puanlar ile cevaplayıcıların bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon hesaplanmış ve bu değer $r = 0,65$ ($p < 0,05$) bulunmuştur. OCAP yönteminin kullanılması durumunda ise toplam puanlar ile öğrencilerin bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyonun $r = 0,70$ ($p < 0,05$) olduğu görülmüştür.

İki farklı puanlama durumuna ilişkin elde edilen korelasyon değerleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek amacıyla, iki bağımlı korelasyon katsayısı arasındaki farkın test edilmesine ilişkin yönelik test sonuçları (Meng, Rosenthal & Rubin, 1992) incelenmiştir. Böylece elde edilen iki korelasyon değerinin karşılaştırılmasına ilişkin hesaplanan Z değeri $Z = 7,93$ bulunmuştur. Hesaplanan Z değeri, $\alpha = 0,05$ düzeyinde kritik tablo değeri olan 1,96 değerinden büyük olduğundan incelenen iki farklı puanlama durumuna ilişkin elde edilen korelasyon değerleri arasında anlamlı bir fark olduğu söylenebilmektedir. Buna göre GP yönteminin kullanıldığı durumda elde edilen cevaplayıcıların toplam puanları ile bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon değeri, OCAP yöntemiyle bulunan toplam puanlar ile öğrencilerin bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon değerinden anlamlı olarak daha küçüktür.

Buraya kadar sunulan GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin geçerlik kanıtlarının karşılaştırılmasıyla ilgili sonuçlar doğrultusunda OCAP yönteminin test puanlarının geçerliği bakımından genel olarak daha avantajlı olduğu yorumu yapılabilmektedir. Daha önce belirtildiği gibi bu araştırmada kullanılan kısa cevaplı bir testte olası cevapların deneysel olarak ağırlıklandırılmasına dayalı olan puanlama yaklaşımına başka bir çalışmada rastlanmamıştır. Ancak bu araştırmada test puanlarının geçerliğine ilişkin elde edilen bulgular, Echternacht (1976) ile Diedenhofen ve Musch (2015)'un buradaki olası cevaplara benzer şekilde çoktan seçmeli bir testte seçenekleri

deneysel olarak ağırlıklandırdıkları çalışmalarının sonuçlarıyla tutarlılık göstermektedir. Gözen Çıtak (2007)'in çalışmasında ise bu araştırmanın bulgularından farklı olarak deneysel seçenek ağırlıklandırmanın test geçerliği üzerinde geleneksel puanlamayla benzer düzeyde etki yaptığı belirlenmiştir.

Birinci alt problemle ilgili son olarak puanlanmada önce GP, daha sonra OCAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin güvenilirlik katsayılarının nasıl değişmekte olduğu araştırılmıştır. Söz konusu iki puanlama durumunda elde edilen puanlara ilişkin güvenilirlik katsayıları hesaplanarak bu güvenilirlik katsayıları arasında manidar bir fark olup olmadığı incelenmiştir. Buna göre GP yönteminin kullanıldığı durumda elde edilen puanlara dayalı olarak güvenilirlik analizi yapılmış ve Cronbach alfa güvenilirlik katsayısı $\alpha = 0,86$ bulunmuştur. Aynı durumda güvenilirlik kestirimi için McDonald'ın ω katsayısı ise $\omega = 0,88$ hesaplanmıştır. OCAP yönteminin kullanıldığı durumda ise Cronbach alfa güvenilirlik katsayısının $\alpha = 0,91$ olduğu görülmüştür. Bu puanlama yöntemiyle ilgili olarak McDonald'ın ω katsayısı ise $\omega = 0,92$ hesaplanmıştır. Bu iki durumda elde edilen Cronbach alfa değerleri arasındaki farkın anlamlı olup olmadığını belirlemek için gerçekleştirilen Fisher'in z testi sonucu incelendiğinde, hesaplanan z değeri $z = 4,03$ bulunmuştur. Bu değer 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığının (Kenny, 1987) dışında olduğu görülmektedir. Buna göre GP ve OCAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenilirlik analizlerinin sonucunda hesaplanan Cronbach alfa değerleri arasında anlamlı bir farklılığın bulunduğu söylenebilmektedir. OCAP yönteminin kullanıldığı durumda kestirilen Cronbach alfa güvenilirlik katsayısı, GP yapılması durumunda kestirilen Cronbach alfa katsayısından anlamlı bir şekilde daha büyüktür. Benzer şekilde iki duruma ilişkin McDonald'ın ω katsayıları arasındaki farkın anlamlılığını incelemek için gerçekleştirilen Fisher'in z testi sonucunda $z = 3,69$ bulunmuştur. Bu değer de 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığının (Kenny, 1987) dışında olduğu görülmektedir. Böylece GP ve OCAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenilirlik analizlerinin sonucunda hesaplanan McDonald'ın ω katsayıları arasında anlamlı bir farklılığın bulunduğu söylenebilmektedir. OCAP yönteminin kullanıldığı durumda kestirilen McDonald'ın ω katsayısı, GP yapılması durumunda kestirilen McDonald'ın ω katsayısından istatistiksel olarak anlamlı bir şekilde daha büyüktür.

GP ve OCAP yöntemlerinin kullanıldığı durumlarda elde edilen güvenilirlik kestirimlerine ilişkin sonuçlar doğrultusunda OCAP yönteminin testin güvenilirliği

bakımından daha avantajlı olduğu söylenebilmektedir. Bu araştırmada klasik test kuramına göre güvenilirliğe ilişkin elde edilen bulgular, Echternacht (1976), Davis ve Fifer (1959), Crehan ve Haladyna (1994), Gözen Çıtak (2007) ile Diedenhofen ve Musch (2015)'un buradaki olası cevaplara benzer şekilde çoktan seçmeli bir testte seçenekleri deneysel olarak ağırlıklandırırdıkları çalışmalarının sonuçlarıyla tutarlılık göstermektedir.

İkinci Alt Probleme İlişkin Bulgular ve Yorumlar

İkinci alt problemle ilgili olarak araştırmada kullanılan kısa yanıtlı testin puanlanmasında önce GP, daha sonra OCAP yönteminin kullanılması durumunda madde tepki kuramına göre öncelikle test puanlarının geçerliğinin nasıl değiştiği, daha sonra aynı durumda test puanlarına ilişkin güvenirliliğin nasıl değişmekte olduğu incelenmiştir.

Öncelikle bu iki puanlama yöntemiyle elde edilen verinin MTK'nın tek boyutluluk ve yerel bağımsızlık varsayımlarını sağlayıp sağlamadığı incelenmiş ve ardından MTK kalibrasyonları yapılmıştır. Tek boyutluluğun her iki puanlama durumu için de sağlandığını gösteren AFA sonuçları Birinci Alt Probleme İlişkin Bulgular ve Yorumlar bölümünde verilmiştir. GP yönteminin kullanıldığı durumda, yapılan AFA sonuçlarına göre öz değeri 1'den büyük olan tek bir faktör olduğu (7.860) ve bu faktörün toplam varyansın %49,1'ini açıkladığı belirlenmiştir. Buna göre 16 maddenin tek bir faktör altında toplandığı görülmüştür. Benzer şekilde OCAP yönteminin kullanıldığı durumda yapılan AFA sonuçlarında, öz değeri 1'den büyük olan tek bir faktör olduğu (8.76) ve bu faktörün toplam varyansın %54,8'ini açıkladığı belirlenmiştir. Böylece hem GP, hem de OCAP yönteminin kullanılması durumunda elde edilen bulgular, testte yer alan 16 maddenin tek bir faktör altında toplandığını ve tek boyutluluk özelliğini sağladığını göstermektedir.

Tek boyutluluk varsayımının ardından yerel bağımsızlığın incelenmesi için Yen'in Q3 istatistiği hesaplanmıştır. GP yönteminin kullanıldığı birinci durumda testte yer alan maddeler arasındaki Q3 değerleri Tablo 23'te verilmiştir.

Tablo 23

GP Yönteminin Kullanıldığı Durumda Maddeler Arasındaki Q3 Değerleri

	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16
m1	,03	,03	,05	-,12	-,10	-,07	-,06	-,12	-,10	,01	-,05	-,09	,04	-,03	-,05
m2		,07	-,02	-,08	-,09	-,07	-,09	-,11	-,06	,00	-,07	-,07	-,01	-,07	-,13
m3			-,10	-,04	,00	,00	-,08	-,09	-,06	-,15	-,13	-,06	-,08	-,07	-,04
m4				-,08	,04	,03	-,02	-,04	-,14	-,06	,01	-,15	-,11	-,13	-,09
m5					,10	-,02	-,06	-,11	-,06	-,10	-,09	-,03	-,02	-,08	-,07
m6						-,03	-,05	-,04	-,10	-,01	-,06	-,01	-,02	-,09	-,10
m7							-,11	-,08	-,07	-,12	-,05	-,09	-,20	-,12	,00
m8								,00	,00	-,07	-,03	-,06	-,15	-,15	-,02
m9									-,20	-,15	-,09	-,06	,00	,03	,08
m10										,09	-,01	-,17	-,13	-,15	-,06
m11											-,03	-,07	-,01	-,15	-,19
m12												-,08	-,10	-,10	-,08
m13													-,07	-,06	-,13
m14														,07	-,01
m15															,08
m16															

Tablo 23'te verilen Q3 değerleri incelendiğinde tüm değerlerin 0,30 (Røe, Damsgård, vd, 2014)'dan küçük olduğu görülmektedir. Buna göre GP yönteminin kullanıldığı durumda elde edilen veri setinde yer alan 16 maddeden herhangi birinin yerel bağımsızlığı bozmadığı sonucuna ulaşılmıştır. İkinci durumda OCAP yöntemi kullanıldığında testte yer alan maddeler arasındaki Q3 değerleri Tablo 24'te verilmiştir.

Tablo 24

OCAP Yönteminin Kullanıldığı Durumda Maddeler Arasındaki Q3 Değerleri

	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16
m1	,03	,05	,10	-,06	-,06	-,06	-,08	-,11	-,18	-,03	-,12	-,18	-,01	,01	-,11
m2		,12	,01	-,11	-,09	-,06	-,08	-,11	-,07	-,06	-,13	-,09	-,06	-,06	-,10
m3			,04	-,09	,00	0,05	-,04	-,06	-,06	-,16	-,16	-,05	-,01	-,08	-,04
m4				-,08	,05	,03	-,08	-,07	-,18	-,13	-,14	-,12	-,07	-,09	-,03
m5					,07	-,04	-,03	-,11	-,04	,00	-,11	,04	-,05	-,11	-,13
m6						-,02	-,05	-,04	-,10	-,03	-,07	,02	-,04	-,04	-,06
m7							,12	-,04	-,04	-,11	-,07	-,07	-,15	-,07	-,02
m8								,01	,09	-,05	-,04	-,07	-,12	-,15	,03
m9									-,15	-,16	-,08	-,02	,01	,00	-,02
m10										,06	,08	-,10	-,19	-,17	,01
m11											,07	-,06	-,07	-,14	-,14
m12												-,09	-,11	-,08	,05
m13													-,05	-,09	-,14
m14														,09	-,03
m15															-,01
m16															

Tablo 24'te verilen Q3 değerlerinin tamamının 0,30'dan küçük olduğu görülmektedir. Buna göre OCAP yönteminin kullanıldığı durumda elde edilen veri setinde yer alan 16 maddeden herhangi birinin yerel bağımsızlığı bozmadığı sonucuna ulaşılmıştır. MTK varsayımlarından tek boyutluluk ve yerel bağımsızlığın test edilmesinin ardından GP ve OCAP yöntemleriyle elde edilen veriye ilişkin model veri uyumu incelemelerine geçilmiştir.

GP Yöntemiyle Elde Edilen Veri için MTK Kalibrasyonları

GP yönteminin kullanıldığı durumda elde edilen ikili veri seti için model veri uyumu incelemeleri kapsamında maddelerin kalibrasyonları MTK modellerinden 1PLM ve 2PLM'ye göre yapılmıştır. Öncelikle 1PLM ve 2PLM, model veri uyumu

bakımından karşılaştırılmıştır. Bu modellerin istatistiksel karşılaştırmaları Tablo 25’te sunulmuştur.

Tablo 25

GP Yöntemiyle Elde Edilen Veri için MTK Model Karşılaştırma Sonuçları

Model	AIC	BIC	logLik	χ^2	df	p
1PLM	10302,63	10376,98	-5134,31			
2PLM	10261,37	10401,31	-5098,68	71,26	15	<0,001

Tablo 25’te verilen iki modelin karşılaştırılmasına ilişkin sonuçlar incelendiğinde p değerinin 0,001’den küçük olduğu, ayrıca 1PLM’den 2PLM’ye geçilirken AIC ve logLikelihood değerlerinin düştüğü görülmüştür. Bu nedenle 2PLM’nin eldeki veri setine daha uygun olabileceği söylenebilmektedir. 1PLM ve 2PLM’nin istatistiksel karşılaştırmalarının ardından her iki model için elde edilen S χ^2 istatistikleri Tablo 26’da verilmiştir.

Tablo 26

GP Yöntemiyle Elde Edilen Veri için Madde Uyum İndeksleri

Madde	1PLM		2PLM	
	S χ^2	RMSEA	S χ^2	RMSEA
m1	29,59	0,047	21,50	0,033
m2	27,55	0,047	21,02	0,036
m3	28,01	0,048	25,84	0,044
m4	13,15	0,004	13,57	0,009
m5	20,24	0,034	20,13	0,034
m6	52,71	0,076	13,51	0,008
m7	12,05	0,003	10,30	0,000
m8	17,31	0,028	16,86	0,026
m9	6,98	0,000	5,14	0,000
m10	23,21	0,040	14,66	0,024
m11	9,16	0,000	8,83	0,000
m12	16,34	0,025	16,37	0,029
m13	19,42	0,033	9,74	0,000
m14	24,10	0,042	21,59	0,037
m15	19,10	0,032	17,50	0,028
m16	28,44	0,048	21,41	0,037

Tablo 26’da verilen değerlere göre hem 1PLM hem de 2PLM için madde uyum istatistiklerine ait RMSEA değerleri 0,05’ten küçüktür. Böylece testte yer alan maddelerin her iki modelle de uyum gösterdiği söylenebilmektedir. Bu durumda

istatistiksel karşılaştırma sonuçları da dikkate alınarak 2PLM'nin kullanılmasına karar verilmiştir. Bu modele göre elde edilen maddelere ait parametreler Tablo 27'de verilmiştir.

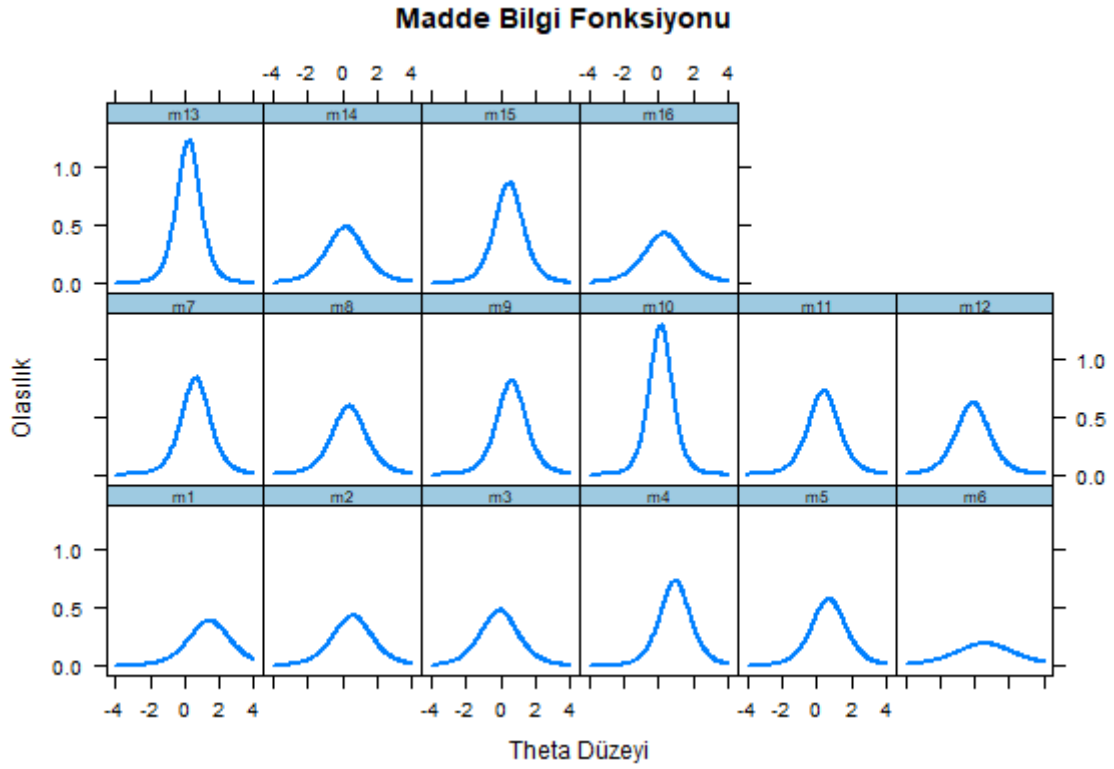
Tablo 27

2PLM'ye Göre Elde Edilen Madde Parametreleri

Madde	a	b
m1	1,253	1,421
m2	1,319	0,554
m3	1,388	-0,070
m4	1,720	0,967
m5	1,526	0,714
m6	0,881	0,571
m7	1,824	0,644
m8	1,537	0,371
m9	1,805	0,640
m10	2,277	0,153
m11	1,702	0,423
m12	1,575	-0,084
m13	2,218	0,248
m14	1,389	0,152
m15	1,866	0,499
m16	1,308	0,312

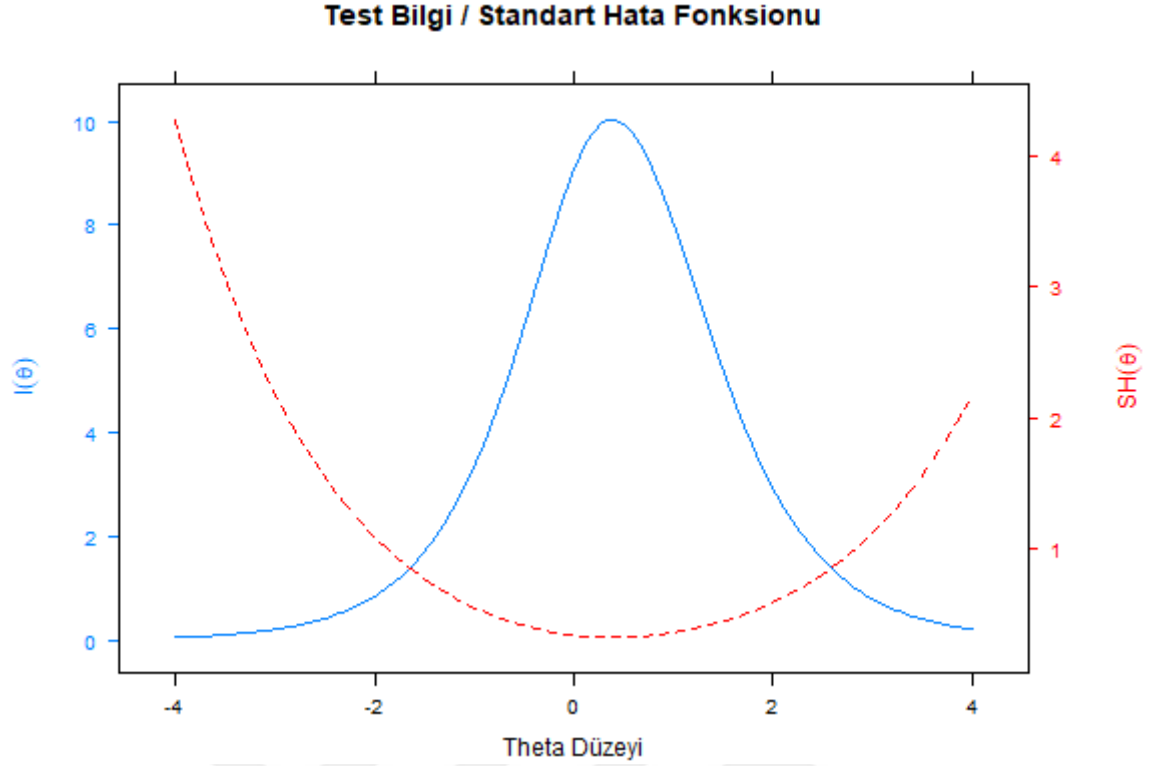
Tablo 27'de verilen maddelerin ayırt edicilik parametrelerinin (a) 0,881 ile 2,277 arasında değer aldığı gözlenmektedir. Baker'e (2001) göre a parametresi için 0,01-0,34 arası çok düşük, 0,35-0,64 arası düşük, 0,65-1,34 arası orta, 1,35-1,69 arası yüksek ve 1,70 ve üstü çok yüksek olarak değerlendirilmektedir. Buna göre dört madde (m1, m2, m6 ve m16) orta, beş madde (m3, m5, m8, m12 ve m14) yüksek, yedi madde (m4, m7, m9, m10, m11, m13 ve m15) çok yüksek ayırt ediciliğe sahiptir. Kestirilen diğer madde parametresi (b) madde güçlüğü değerleri incelendiğinde ise bu parametrelerin -0,084 ile 1,421 arasında değer aldığı görülmektedir.

Madde parametrelerinin ardından, madde ve test bilgi fonksiyonları 2PLM'ye göre incelenmiştir. Bu fonksiyonlara ait grafikler sırasıyla Şekil 9 ve Şekil 10'da verilmiştir.



Şekil 9. GP Yönteminin Kullanıldığı Durumda Madde Bilgi Fonksiyonları

Şekil 9’da verilen madde bilgi fonksiyonları incelendiğinde maddelerin çoğunun sıfır civarındaki yetenek puanlarına sahip olan bireyler için en fazla bilgi verdiği söylenebilmektedir. Şekil 10’da sunulan test bilgi fonksiyonunda da benzer bir durum görülmektedir.



Şekil 10. GP Yönteminin Kullanıldığı Durumda Test Bilgi Fonksiyonu

Şekil 10'a göre bu çalışmada kullanılan test, 0 civarı yetenek puanına sahip bireyler için en fazla bilgiyi vermekte, en az hata ile ölçüm yapmakta ve bu yetenek puanları için güvenilirlik en yüksek olmaktadır. Son olarak GP yönteminin kullanıldığı durumda test için marjinal güvenilirlik katsayısı hesaplanmış ve 0,83 değeri elde edilmiştir.

OCAP Yöntemiyle Elde Edilen Veri için MTK Kalibrasyonları

Model veri uyumu incelemeleri kapsamında, OCAP yönteminin kullanıldığı durumda elde edilen çok kategorili veri seti için maddelerin kalibrasyonları MTK modellerinden Kademeli Tepki Modeli (KTM - Graded Response Model) ve Genelleştirilmiş Kısmi Puan Modeline (GKPM - Generalized Partial Credit Model) göre yapılmıştır. KTM ve GKPM modellerinin istatistiksel karşılaştırmaları Tablo 28'de yer almaktadır.

Tablo 28

OCAP Yöntemiyle Elde Edilen Veri için MTK Model Karşılaştırma Sonuçları

Model	AIC	AICc	SABIC	BIC	logLik	χ^2	sd	p
KTM	16627,34	16654,06	16720,46	16974,43	-8233,67			
GKPM	16626,34	16653,06	16719,46	16973,43	-8233,16	1001,09	0	0,00

Tablo 28’de yer alan iki model için elde edilen AIC, BIC, logLikelihood değerlerinin birbirlerine çok yakın değerler aldığı görülmektedir. Bununla birlikte χ^2 istatistiğine ait p değerinin 0.05’ten küçük olması test maddeleri için GKPM’nin daha uygun bir model olabileceğini göstermektedir. KTM ve GKPM modellerinin istatistiksel karşılaştırmalarının ardından her iki model için elde edilen S_{χ^2} istatistikleri Tablo 29’da verilmiştir.

Tablo 29

OCAP Yöntemiyle Elde Edilen Veri için Madde Uyum İndeksleri

Madde	KTM		GKPM	
	S_{χ^2}	RMSEA	S_{χ^2}	RMSEA
m1	120,60	0,022	116,87	0,020
m2	147,34	0,036	144,78	0,035
m3	87,09	0,027	73,06	0,022
m4	95,39	0,018	88,77	0,015
m5	111,89	0,018	120,74	0,025
m6	73,92	0,025	65,49	0,021
m7	74,14	0,023	74,10	0,025
m8	68,45	0,019	68,46	0,022
m9	96,12	0,021	93,88	0,022
m10	108,19	0,028	106,49	0,028
m11	165,86	0,039	165,71	0,042
m12	100,46	0,007	88,61	0,000
m13	46,45	0,000	40,13	0,000
m14	74,03	0,022	73,07	0,022
m15	61,16	0,028	55,91	0,023
m16	121,81	0,026	116,01	0,022

Tablo 29’da verilen değerlere göre hem KPM hem de GKPM modeli için madde uyum istatistiklerine ait RMSEA değerleri 0,05’ten küçüktür. Bu doğrultuda testte yer alan maddelerin her iki modelle de uyumlu olduğu söylenebilmektedir. Model madde

uyumunun ardından KTM ve GKPM'ye göre madde parametreleri hesaplanmıştır. Elde edilen değerler incelendiğinde GKPM için elde edilen madde ayırt edicilik değerlerinin KTM ile elde edilen değerlere göre tüm maddelerde daha düşük olduğu görülmüştür. Ayrıca GKPM'de maddelerdeki kategorilere ilişkin adım güçlük değerlerinin birçok maddede monoton artan olmadığı görülmüştür. Bu durumda KTM'nin test maddelerine daha uygun olduğu değerlendirilmiştir. KTM'ye göre elde edilen madde parametreleri Tablo 30'da yer almaktadır.

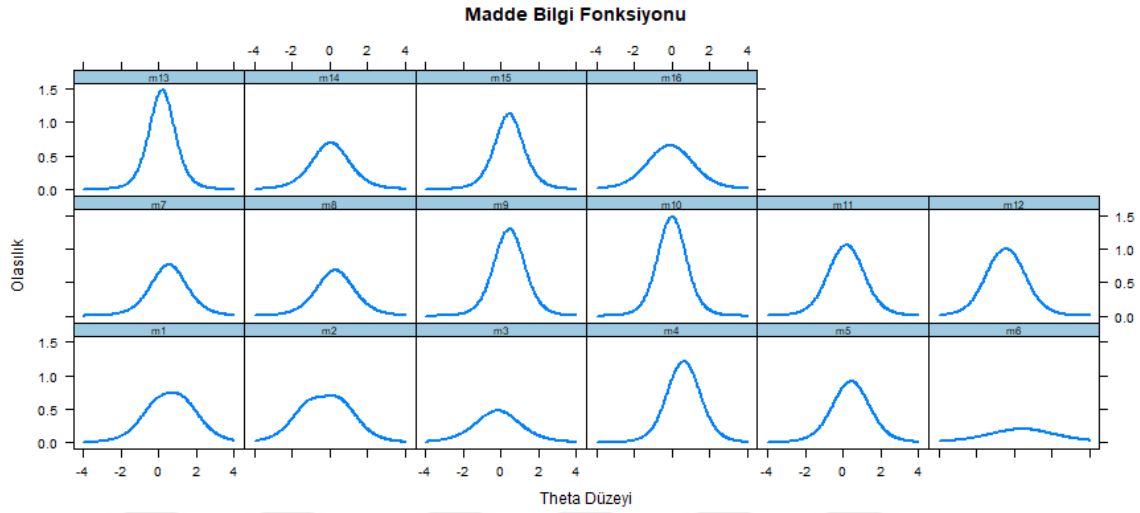
Tablo 30

KTM'ye Göre Elde Edilen Madde Parametreleri

Madde	b1	b2	b3	b4	a
m1	-0,012	0,008	0,962	1,292	1,557
m2	-1,052	0,282	0,479	0,549	1,538
m3	-0,335	-0,190	-0,079	-0,019	1,321
m4	0,333	0,444	0,736	0,926	2,004
m5	0,123	0,386	0,454	0,717	1,751
m6	0,248	0,335	0,384	0,571	0,871
m7	0,384	0,497	0,604	0,686	1,664
m8	0,162	0,316	0,379	0,427	1,583
m9	0,226	0,250	0,300	0,674	2,106
m10	-0,195	-0,062	0,003	0,210	2,250
m11	-0,168	0,271	0,366	0,495	1,873
m12	-0,876	-0,451	-0,253	-0,022	1,799
m13	0,090	0,175	0,240	0,283	2,325
m14	-0,092	0,063	0,083	0,182	1,595
m15	0,369	0,432	0,451	0,514	2,056
m16	-0,522	-0,410	-0,337	0,314	1,475

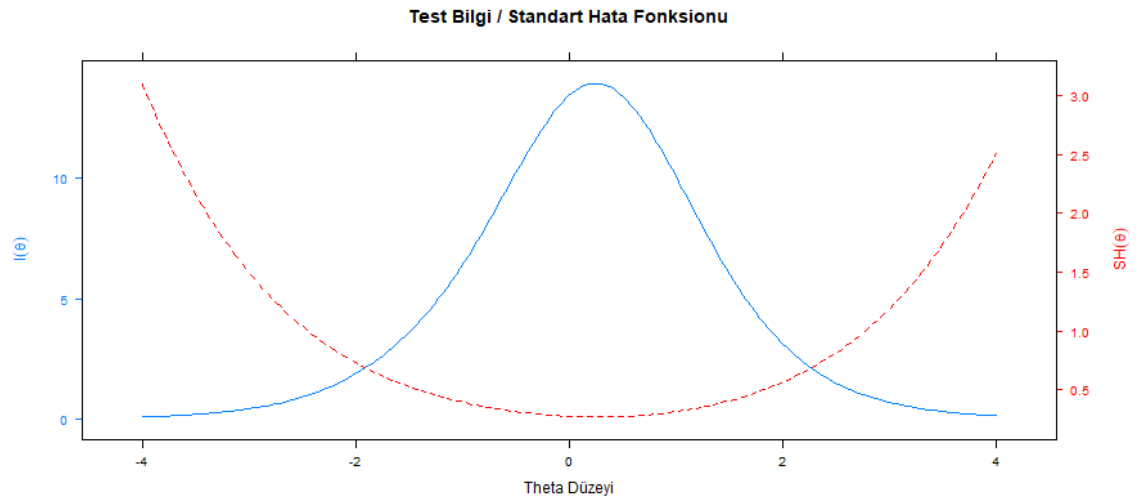
Tablo 30'da verilen maddelerin ayırt edicilik parametrelerinin (a) 0,871 ile 2,325 arasında değer aldığı gözlenmektedir. Baker'e (2001) göre a parametresi için 0,01-0,34 arası çok düşük, 0,35-0,64 arası düşük, 0,65-1,34 arası orta, 1,35-1,69 arası yüksek ve 1,70 ve üstü çok yüksek olarak değerlendirilmektedir. Buna göre iki madde (m3 ve m6) orta, altı madde (m1, m2, m7, m8, m14 ve m16) yüksek, sekiz madde (m4, m5, m9, m10, m11, m12, m13 ve m15) çok yüksek ayırt ediciliğe sahiptir. Kestirilen diğer madde parametresi (b_i) madde güçlüğü veya cevaplayıcıların maddeyi cevaplama frekansına ilişkin bilgi sağlayan eşik değerlerdir. Testteki 16 maddenin KTM'ye göre elde edilen eşik değerleri incelendiğinde bu değerlerin tüm maddelerde monoton artan olduğu görülmektedir.

Madde parametrelerinin ardından, madde ve test bilgi fonksiyonları KTM'ye göre incelenmiştir. Bu fonksiyonlara ait grafikler sırasıyla Şekil 11 ve Şekil 12'de verilmiştir.



Şekil 11. OCAP Yönteminin Kullanıldığı Durumda Madde Bilgi Fonksiyonları

Şekil 11'da verilen madde bilgi fonksiyonları incelendiğinde maddelerin çoğunlukla sıfır civarındaki yetenek puanlarına sahip olan cevaplayıcılar için en fazla bilgi verdiği söylenebilmektedir. Şekil 12'de sunulan test bilgi fonksiyonunda da bu durum benzer şekilde görülmektedir.

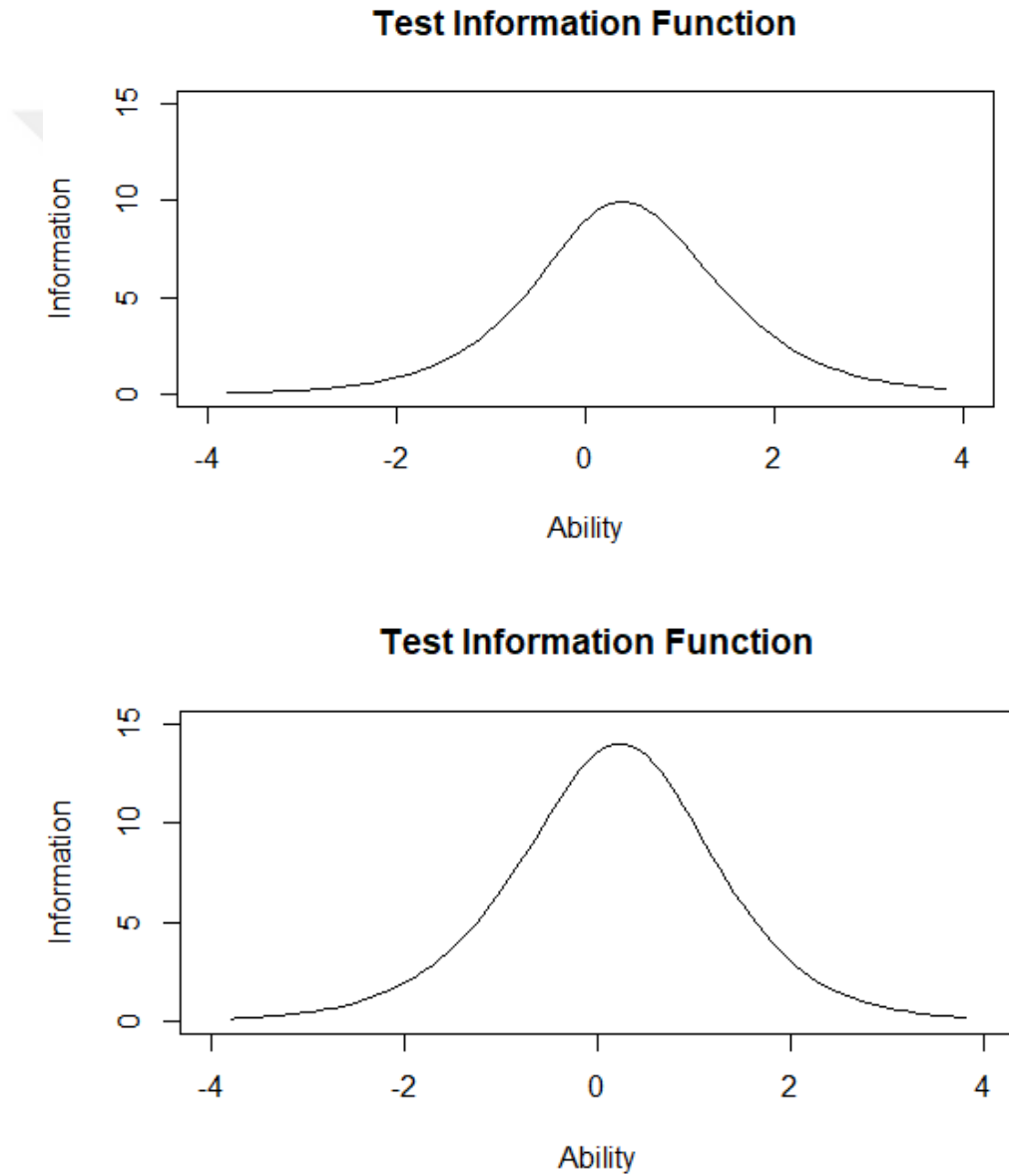


Şekil 12. OCAP Yönteminin Kullanıldığı Durumda Test Bilgi Fonksiyonu

Şekil 12'ye göre bu çalışmada kullanılan test, 0 civarı yetenek puanına sahip bireyler için en fazla bilgiyi vermekte, en az hata ile ölçüm yapmakta ve bu yetenek puanları için güvenirlik en yüksek olmaktadır. Son olarak OCAP yönteminin

kullanıldığı durumda test için marjinal güvenirlik katsayısı hesaplanmış ve 0,88 değeri elde edilmiştir.

GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin MTK'ya dayalı geçerlik ve güvenirlik analizleri ayrı ayrı incelendikten sonra, elde edilen sonuçların iki puanlama durumunda nasıl değişmekte olduğu incelenmiştir. Test puanlarının geçerliğiyle ilgili olarak GP ve OCAP yöntemlerinin kullanıldığı durumlarda elde edilen test bilgi fonksiyonları karşılaştırılmıştır. Şekil 13'te bu iki puanlama durumunda elde edilen test bilgi fonksiyonları birlikte verilmiştir.



Şekil 13. GP ve OCAP Yönteminin Kullanıldığı Durumlara İlişkin Test Bilgi Fonksiyonları

Şekil 13'te önce GP, daha sonra OCAP yönteminin kullanıldığı durumda elde edilen test bilgi fonksiyonlarının grafikleri görülmektedir. Bu grafikler incelendiğinde OCAP yönteminin kullanıldığı durumda alınan bilgi düzeyinin arttığı söylenebilmektedir. Bu bulgu Thissen (1976), Childs, Dunn, van Barneveld, Jaciw ve McIlroy (2003) ile Si ve Schumacker (2004)'ın çalışmalarında belirtilen çok kategorili puanlama yöntemlerinin kullanılmasıyla, özellikle düşük yetenek düzeyindeki kişiler olmak üzere, sınava giren cevaplayıcılar için daha fazla bilgi elde edildiği yönündeki bulgularla örtüşmektedir. Ayrıca bu çalışmanın bulgularıyla paralel olacak şekilde DeMars (2008), iki kategorili modele ilişkin hesaplanan standart hataların çok kategorili bir modeldeki değerlerden daha yüksek olduğunu belirtmiştir.

GP ve OCAP yöntemlerinin kullanıldığı durumlardaki yetenek kestirimlerinin dağılımı incelendiğinde, GP yöntemi kullanıldığında θ değerlerinin ortalamasının -0,01 ve OCAP yöntemi kullanıldığında ise kestirilen θ değerlerinin ortalamasının 0,029 olduğu görülmüştür. Ortalamalar arasındaki bu farkın OCAP yöntemiyle kısmi doğru cevaplara puan verilmesinden kaynaklandığı düşünülmektedir. GP yönteminde tam doğru cevabı yazamayan öğrenciler hiç puan alamazken OCAP yönteminde tam doğru olmayan cevaplar da kısmi puanlar verilerek puanlanmıştır. Bu bulguya benzer şekilde Zhai ve Li (2021)'nin çalışmalarında da çok kategorili puanlamanın kullanıldığı durumda ikili modele göre daha yüksek bir öğrenci yetenek ölçümü elde edilmiştir.

İkinci alt problemle ilgili son olarak puanlanmada önce GP, daha sonra OCAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin MTK'ya dayalı olarak elde edilen güvenilirlik katsayılarının nasıl değişmekte olduğu araştırılmıştır. Bu doğrultuda GP yönteminin kullanıldığı durumda elde edilen puanlara dayalı olarak güvenilirlik analizi yapılmış ve marjinal güvenilirlik katsayısı 0,83 bulunmuştur. OCAP yönteminin kullanıldığı durumda ise marjinal güvenilirlik katsayısının 0,88 olduğu görülmüştür. Bu durumda OCAP yönteminin kullanıldığı durumda güvenilirlik analizinde daha büyük güvenilirlik katsayısının elde edildiği görülmüştür.

Madde tepki kuramına dayalı olarak GP ve OCAP yöntemlerinin kullanıldığı durumlarda elde edilen geçerlik ve güvenilirlik kanıtlarına ilişkin sonuçlar doğrultusunda OCAP yönteminin testin psikometrik özellikleri bakımından daha güçlü olduğu söylenebilmektedir. Bu araştırmada madde tepki kuramına göre elde edilen bulgular, Young (1991)'in madde tepki kuramına dayalı aşamalı tepki modeli ile geçerliği daha yüksek sonuçlar elde ettiği çalışmasıyla tutarlılık göstermektedir. Bu araştırmanın bulgularıyla paralel olacak şekilde Donoghue (1994) ve Kinsey (2003) de çok kategorili

puanlamanın testten elde edilen bilgi düzeyini yükselttiğini belirtmiştir. Yamamoto ve Kulick (1992) tarafından yapılan bir araştırmada ise çok kategorili olarak puanlanan maddelerin ortalama olarak, ikili puanlanan maddelerden biraz daha az bilgi içerdiği belirtilmiştir. Özdemir (2002) ve Gözen Çıtak (2007)'ın çalışmalarında da bu çalışmadan farklı olarak “1-0” puanlamada geçerliği daha yüksek olan sonuçlar elde edilebildiği görülmüştür.

Üçüncü Alt Probleme İlişkin Bulgular ve Yorumlar

Araştırmadaki üçüncü alt problemle ilgili olarak kullanılan kısa yanıtli testin puanlanmasında önce “1-0” puanlama, daha sonra OCAP yönteminin kullanılmasıyla elde edilen orijinal veri setlerinden rastgele 300 kişilik 100'er tane veri seti oluşturulmuştur. Daha sonra bu veri setleri aracılığıyla “1-0” puanlama ve OCAP yönteminin kullanılması durumunda öncelikle test puanlarının geçerliğinin nasıl değiştiği incelenmiştir. Daha sonra ise aynı durumda test puanlarına ilişkin güvenilirlik kestirimlerin nasıl değişmekte olduğuna bakılmıştır.

Geçerlik bulgularının incelenmesi kapsamında, GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin 100 veri setinin her biri için AFA ile belirlenen açıklanan varyans değerlerinin ve madde faktör yüklerinin bu iki puanlama durumunda nasıl değişmekte olduğu incelenmiştir. GP yöntemi için elde edilen 100 veri setinin her birindeki puanların AFA sonucunda açıklanan varyans değerlerinin 0,449 ve 0,535 arasında olduğu görülmüştür. OCAP yöntemi için elde edilen 100 veri setinin her birindeki puanların AFA sonucunda görülen açıklanan varyans değerleri ise 0,505 ve 0,580 arasındadır.

GP ve OCAP yöntemlerinin kullanıldığı 100'er veri setinin her biri için AFA ile belirlenen açıklanan varyans değerleri arasındaki farkların anlamlı olup olmadığını belirlemek için bağımsız örneklemeler için t testi yapılmıştır. Öncelikle her bir puanlama yöntemine ilişkin elde edilen açıklanan varyans değerlerinin normal dağılım gereğini sağlayıp sağlamadığı kontrol edilmiştir. GP ve OCAP yöntemleri için AFA ile belirlenen açıklanan varyans değerlerinin dağılımına ilişkin Kolmogorov Smirnov normallik testi sonuçları incelenmiştir. Kolmogorov Smirnov Testi sonuçlarına göre puan dağılımının normal dağılımdan farklılaşmadığı görülmüştür ($p>0.05$). Böylece hem GP hem de OCAP yönteminin kullanıldığı durumda elde edilen açıklanan varyans

değerlerinin dağılımının normal olduğu görülmüştür. Parametrik test varsayımlarından normallığın incelenmesinden sonra varyansların homojenliği test edilmiştir. Varyansların eşit olup olmadığına Levene testi ile bakılmıştır. Bu incelemeler sonucunda açıklanan varyans değerlerine ilişkin alt grupların varyansları arasında anlamlı farklılık olmadığı ($p>0,05$) görülmüş, incelenen değişken açısından varyansların homojenliği gereğinin sağlandığı belirlenmiştir. GP ve OCAP yöntemleri için AFA ile belirlenen açıklanan varyans değerlerinin kullanılan puanlama yöntemine göre anlamlı farklılık gösterip göstermediği bağımsız gruplar için t-testi ile analiz edilmiştir. Elde edilen sonuçlar Tablo 31’de verilmiştir.

Tablo 31

GP ve OCAP Yöntemlerinde AFA Sonucunda Elde Edilen Açıklanan Varyans Değerlerinin Karşılaştırılmasına İlişkin t Testi Sonuçları

Yöntem	n	$\bar{X}$	ss	t	sd	p
GP	100	0,493	0,016	25,201	198	0,000
OCAP	100	0,548	0,014			

Tablo 31 incelendiğinde GP ve OCAP yöntemleri için AFA ile belirlenen açıklanan varyans değerlerinin kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaştığı görülmektedir ($t(198)= 25,201$; $p<0,05$). Buna göre GP ve OCAP yönteminin kullanılmasıyla elde edilen orijinal veri setlerinden rastgele 300 kişilik 100’er tane veri seti oluşturularak her biri için elde edilen AFA sonuçları incelendiğinde, OCAP yöntemi kullanıldığında elde edilen açıklanan varyans değerlerinin GP yöntemi kullanıldığında elde edilen değerlerden anlamlı olarak daha büyük olduğu görülmektedir.

GP ve OCAP yönteminin kullanılmasıyla elde edilen veriden rastgele oluşturulan veri setlerinin her biri için AFA sonuçlarında görülen ortalama faktör yükleri ise Tablo 32’de birlikte sunulmuştur.

Tablo 32

GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Maddelerin Ortalama Faktör Yük Değerleri

Madde No	Ortalama Faktör Yük Değeri (GP)	Ortalama Faktör Yük Değeri (OCAP)
m1	0,62	0,71
m2	0,65	0,70
m3	0,68	0,68
m4	0,73	0,80
m5	0,69	0,75
m6	0,51	0,53
m7	0,75	0,73
m8	0,69	0,71
m9	0,74	0,80
m10	0,80	0,81
m11	0,72	0,76
m12	0,70	0,75
m13	0,80	0,83
m14	0,68	0,73
m15	0,76	0,80
m16	0,64	0,68

Tablo 32’de görüldüğü gibi GP yönteminin kullanılması durumunda AFA sonuçlarında görülen ortalama faktör yük değerleri 0,51 ile 0,80 arasında değer almaktadır. OCAP yöntemiyle elde edilen puanlara dayalı olarak üretilen veriyle AFA yapıldığında ise ortalama faktör yük değerleri 0,53 ile 0,83 arasında değer almaktadır. İncelenen puanlama durumlarında belirlenen ortalama faktör yük değerleri arasındaki farkların anlamlılığını incelemek için Mann-Whitney U testi kullanılmıştır. Tablo 33’te AFA sonucunda elde edilen ortalama faktör yük değerlerinin kullanılan puanlama yöntemine göre karşılaştırılmasına ilişkin Mann-Whitney U testi sonuçları verilmiştir.

Tablo 33

GP ve OCAP Yöntemlerinde AFA Sonucunda Elde Edilen Ortalama Faktör Yük Değerlerinin Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları

Yöntem	n	Sıra Ortalaması	Sıra Toplamı	Z	p
GP	16	13,63	218	-1,74	0,082
OCAP	16	19,38	310		
Toplam	32				

Tablo 33 incelendiğinde AFA sonucunda elde edilen ortalama faktör yük değerlerinin kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaşmadığı görülmektedir ($z = -1,74, p > 0,05$).

AFA sonuçlarının incelenmesinin ardından kısa yanıtlı testin yanıtlarının GP ve OCAP yöntemleriyle puanlanması durumlarında elde edilen puanlara dayalı olarak üretilen veriyle DFA yapıldığında ortaya çıkan sonuçlar karşılaştırılmıştır. GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin DFA sonuçlarında elde edilen uyum indeks değerlerinin kullanılan puanlama yöntemine göre nasıl değişmekte olduğu incelenmiştir.

GP ve OCAP yönteminin kullanılmasıyla elde edilen veriden rastgele oluşturulan veri setlerinin her biri için DFA sonucunda görülen ortalama uyum indeksleri birlikte Tablo 34’te yer almaktadır.

Tablo 34

GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Ortalama Uyum İndeksleri

Puanlama Yöntemi	RMSEA	CFI	AGFI	NNFI	χ^2/sd
GP	0,047	0,94	0,91	0,93	1,67
OCAP	0,056	0,94	0,90	0,93	1,94

Tablo 34’te verilen DFA sonuçlarına göre, GP yönteminin kullanıldığı durumda elde edilen ortalama RMSEA değeri, OCAP yönteminin kullanıldığı durumda elde edilen ortalama RMSEA değerinden daha küçüktür. GP ve OCAP yönteminin kullanıldığı durumlarda ortalama CFI ve ortalama NNFI değerlerinde değişiklik olmadığı görülmüştür. OCAP yönteminin kullanılması durumunda elde edilen ortalama AGFI değeri ise GP yönteminin kullanıldığı durumda elde edilen ortalama AGFI değerine göre daha küçüktür. Son olarak χ^2/sd oranları incelendiğinde, GP yöntemi kullanıldığında ortalama χ^2/sd değerinin OCAP yönteminin kullanıldığı duruma göre daha düşük olduğu belirlenmiştir.

GP ve OCAP yönteminin kullanılmasıyla elde edilen veriden rastgele oluşturulan veri setlerinin her biri için DFA yapıldığında görülen ortalama hata varyansları Tablo 35’te birlikte verilmiştir.

Tablo 35

GP ve OCAP Yöntemlerinin Kullanıldığı Durumlarda DFA Sonucunda Görülen Ortalama Hata Varyansları

Madde No	Ortalama Hata Varyans Değeri (GP)	Ortalama Hata Varyans Değeri (OCAP)
m1	0,83	0,62
m2	0,76	0,66
m3	0,73	0,70
m4	0,72	0,55
m5	0,72	0,59
m6	0,86	0,83
m7	0,66	0,64
m8	0,71	0,65
m9	0,66	0,53
m10	0,57	0,51
m11	0,67	0,57
m12	0,70	0,60
m13	0,58	0,50
m14	0,73	0,64
m15	0,64	0,57
m16	0,76	0,66

Tablo 35 incelendiğinde GP yönteminin kullanılması durumunda elde edilen puanlara dayalı olarak DFA yapıldığında görülen ortalama hata varyans değerlerinin 0,57 ile 0,86 arasında değer aldığı görülmektedir. OCAP yöntemiyle elde edilen puanlara dayalı olarak DFA yapıldığında hesaplanan ortalama hata varyansları ise 0,50 ile 0,83 arasında değişmektedir. İncelenen puanlama durumlarında belirlenen ortalama hata varyansları arasındaki farkların anlamlılığını incelemek için Mann-Whitney U testi yapılmıştır. Tablo 36’da DFA sonucunda elde edilen ortalama hata varyanslarının kullanılan puanlama yöntemine göre karşılaştırılmasıyla ilgili Mann-Whitney U testi sonuçları verilmiştir.

Tablo 36

GP ve OCAP Yöntemlerinde DFA Sonucunda Elde Edilen Ortalama Hata Varyanslarının Karşılaştırılmasına İlişkin Mann-Whitney U Testi Sonuçları

Yöntem	n	Sıra Ortalaması	Sıra Toplamı	z	p
GP	16	21,75	348	-3,17	0,002
OCAP	16	11,25	180		
Toplam	32				

Tablo 36 incelendiğinde DFA sonucunda elde edilen ortalama hata varyanslarının kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaştığı görülmektedir ($z = -3,17, p < 0,05$). Sıra ortalamaları dikkate alındığında OCAP yöntemi kullanıldığında elde edilen ortalama hata varyanslarının GP yönteminin kullanıldığı durumda elde edilen değerlerden anlamlı olarak daha küçük olduğu görülmektedir.

Üçüncü alt probleme ilişkin son olarak GP ve OCAP yönteminin kullanılmasıyla elde edilen veriden rastgele oluşturulan veri setleri için hesaplanan güvenilirlik katsayılarının puanlama yöntemine göre nasıl değişmekte olduğu araştırılmıştır. Söz konusu iki puanlama durumunda elde edilen puanlara ilişkin güvenilirlik katsayıları hesaplanarak bu güvenilirlik katsayıları arasında anlamlı bir farklılık olup olmadığı incelenmiştir.

GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin rastgele oluşturulan veri setlerinin her biri için hesaplanan Cronbach alfa ve McDonald'ın ω katsayıları arasındaki farkların anlamlılığını incelemek için bağımsız örneklem için t testi kullanılmıştır. Öncelikle her bir puanlama yöntemine ilişkin elde edilen güvenilirlik katsayılarının normal dağılım gereğini sağlayıp sağlamadığı kontrol edilmiştir. GP ve OCAP yöntemleri için hesaplanan Cronbach alfa ve McDonald ω katsayılarının dağılımına ilişkin Kolmogorov Smirnov normallik testi sonuçları incelenmiştir. Kolmogorov Smirnov Testi sonuçlarına göre, her bir puanlama yöntemine ilişkin elde edilen güvenilirlik katsayılarının dağılımıyla ilgili anlamlılık düzeylerinin 0.05'ten küçük olmasından dolayı dağılımların normal dağılımdan farklılaştığı görülmüştür. Çarpıklık ve basıklık değerlerine bakıldığında ise bu değerlerin (-1,1) aralığında olduğu görüldüğünden güvenilirlik katsayılarının dağılımının normal olduğu kabul edilmiştir (Büyüköztürk, Çokluk ve Köklü; 2014). Daha sonra varyansların homojenliğine Levene testi ile bakılmıştır. Bu incelemeler sonucunda güvenilirlik katsayılarına ilişkin alt grupların varyansları arasında anlamlı farklılık olmadığı ($p > 0.05$) görülerek incelenen değişken açısından varyansların homojenliği gereğinin sağlandığı belirlenmiştir. GP ve OCAP yöntemleri için güvenilirlik katsayılarının kullanılan puanlama yöntemine göre anlamlı bir fark gösterip göstermediği bağımsız gruplar için t-testi ile analiz edilmiştir. Elde edilen sonuçlar Tablo 37'de yer almaktadır.

Tablo 37

GP ve OCAP Yöntemlerinde Güvenirlik Katsayılarının Karşılaştırılmasına İlişkin t Testi Sonuçları

	Yöntem	n	$\bar{X}$	ss	t	sd	p
Cronbach alfa	GP	100	0,87	0,007	44,613	198	0,000
	OCAP	100	0,91	0,005			
McDonald ω	GP	100	0,88	0,007	42,756	198	0,000
	OCAP	100	0,92	0,005			

Tablo 37 incelendiğinde GP ve OCAP yöntemleri için hesaplanan Cronbach alfa değerlerinin kullanılan puanlama yöntemine göre anlamlı bir biçimde farklılaştığı görülmektedir ($t(198)= 44,613$; $p<0,05$). Buna göre GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin rastgele oluşturulan veri setlerinin her biri için güvenilirlik analizi yapıldığında OCAP yöntemi kullanıldığında elde edilen Cronbach alfa değerlerinin GP yöntemi kullanıldığında elde edilen değerlerden anlamlı olarak daha büyük olduğu görülmektedir.

McDonald ω katsayısına ilişkin sonuçlar incelendiğinde ise benzer şekilde GP ve OCAP yöntemleri için hesaplanan değerlerin birbirinden anlamlı bir biçimde farklılaştığı görülmektedir ($t(198)= 42,756$; $p<0,05$). Böylece GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin rastgele oluşturulan veri setlerinin her biri için güvenilirlik analizi yapıldığında OCAP yöntemi kullanıldığında elde edilen McDonald ω değerlerinin GP yöntemi kullanıldığında elde edilen değerlerden anlamlı olarak daha büyük olduğu söylenebilmektedir.

Araştırmadaki üçüncü alt probleme ilişkin bulgularla puanlama sürecinde önce “1-0” puanlama, daha sonra OCAP yönteminin kullanılmasıyla elde edilen orijinal veri kullanılarak alınan sonuçların farklı veri setlerinin kullanılması durumundaki genellenebilirliği incelenmiştir. Her iki puanlama yöntemine ilişkin rastgele veri üretilmesi durumundaki tüm geçerlik ve güvenilirlik çalışmaları göz önünde bulundurularak rastgele oluşturulan veri setleri için madde faktör yükleri dışındaki tüm sonuçların orijinal verinin kullanıldığı durumdaki sonuçlarla benzerlik gösterdiği görülmüştür. Diedenhofen ve Musch (2019)’un çalışmalarında da bu araştırmaya benzer şekilde rastgele alt örnekleme doğrulaması gerçekleştirilmiş ve deneysel seçenek ağırlıklandırmanın kullanıldığı durumda, geleneksel puanlamadan istatistiksel olarak anlamlı derecede daha yüksek α katsayısı elde edildiği belirlenmiştir.

KTİK'ya ve deneysel ağırlıklandırma kullanılmasına dayalı olarak elde edilen bu sonuçlar, Erdem, Ertuna ve Doęan (2016)'ın alıřmalarının sonuçlarıyla AFA sonuçları bakımından farklılařmakta ve DFA sonuçları bakımından örtüşmektedir. Erdem, Ertuna ve Doęan (2016)'ın arařtırmalarında, bu arařtırmanın aksine, AFA sonuçlarına göre deneysel puanlamanın açıklanan varyans oranında düşmeye sebep olduęu belirlenmiştir. DFA sonuçlarında ise, benzer şekilde uyum indekslerinin deneysel puanlamada önemli düzeyde farklılařmadıęı görülmüřtür. Bu alt probleme iliřkin test puanlarının geçerlięiyle ilgili elde edilen bulgular genel olarak deęerlendirildięinde, Echternacht (1976) ile Diedenhofen ve Musch (2015)'un deneysel ağırlıklandırma yaptıkları alıřmalarının sonuçlarıyla tutarlılık göstermektedir. Gözen ıtak (2007)'ın alıřmasında ise bu arařtırmanın bulgularından farklı olarak deneysel ağırlıklandırmanın test geçerlięi üstünde geleneksel puanlamayla benzer düzeyde etki yaptıęı ifade edilmiştir. Klasik test kuramına dayalı olarak güvenilirlięe iliřkin elde edilen bulgular ise Echternacht (1976), Davis ve Fifer (1959), Crehan ve Haladyna (1994), Gözen ıtak (2007) ile Diedenhofen ve Musch (2015)'un deneysel ağırlıklandırma kullanarak yürüttükleri alıřmalarının sonuçlarıyla tutarlılık göstermektedir.

Dördüncü Alt Probleme İliřkin Bulgular ve Yorumlar

Dördüncü alt problemle ilgili olarak arařtırmada ele alınan ağırlıklı puanlama yöntemlerinin (OCAP, MGİAP ve AP) kullanılması durumlarında klasik test kuramına göre öncelikle test puanlarının geçerlięinin nasıl deęiřtięi incelenmiştir. Daha sonra ise test puanlarına iliřkin güvenilirlik kestirimlerin nasıl deęiřmekte olduęu arařtırılmıştır. Puanlamada OCAP, MGİAP ve AP yöntemlerinin kullanılması durumunda elde edilen puanlara iliřkin yapı geçerlięi kapsamında AFA sonucunda açıklanan varyans deęerlerinin, elde edilen faktör yüklerinin ve DFA sonucunda elde edilen uyum indekslerinin (RMSEA, AGFI, NNFI, CFI ve ki-kare/sd) nasıl deęiřmekte olduęu incelenmiştir.

OCAP yönteminin kullanılması durumunda elde edilen puanların yapı geçerlięine iliřkin tüm sonuçlar Birinci Alt Probleme İliřkin Bulgular ve Yorumlar başlıklı bölümde verilmiştir. MGİAP yöntemiyle ilgili olarak öncelikle GP yönteminin kullanılmasının ardından bu puanlara dayalı olarak hesaplanan madde güçlük indekslerine göre madde ağırlıkları belirlenmiştir. Daha sonra bu madde ağırlıkları

kullanılarak aynı öğrenci yanıtları tekrar puanlanmıştır. Öğrenci yanıtlarının MGİAP yöntemine göre puanlanmasıyla elde edilen puanlara dayalı olarak AFA yapılmıştır. AFA sonucuna göre elde edilen özdeğerlerle yapının tek faktörlü olmasına dayanak oluşturmak amacıyla yapılan paralel analiz sonucunda elde edilen özdeğerlerin birlikte verildiği özdeğer grafiği ekte sunulmuştur. AFA sonuçlarına göre toplam 16 maddenin tek faktör altında toplandığı ve bu faktöre ilişkin öz değerin 7,86 olduğu belirlenmiştir. Ayrıca 16 madde ile yapılan analizdeki tek faktörün varyansın toplam %49,1'ini açıklamış olduğu görülmüştür.

Üçüncü ağırlıklı puanlama yöntemi olan ÇAP ile ilgili olarak hem maddeler hem de olası cevaplar göz önünde bulundurularak bir ağırlıklandırma işlemi yapılmıştır. Nihai uygulamada elde edilen öğrenci yanıtları bu kez ÇAP yaklaşımıyla tekrar puanlanmıştır. Ardından yanıtların bu yöneme göre puanlanmasıyla elde edilen puanlara dayalı olarak AFA yapılmıştır. AFA sonucuna göre elde edilen özdeğerlerle yapının tek faktörlü olmasına dayanak oluşturmak amacıyla yapılan paralel analiz sonucunda elde edilen özdeğerlerin birlikte verildiği özdeğer grafiği ekte sunulmuştur. ÇAP yöntemiyle ilgili olarak elde edilen sonuçlara göre toplam 16 maddenin tek faktör altında toplandığı ve bu faktöre ilişkin öz değerin 5,95 olduğu belirlenmiştir. Ayrıca 16 madde ile yapılan analizdeki tek faktörün varyansın toplam %37,2'sini açıklamış olduğu görülmüştür.

OCAP, MGİAP ve ÇAP yöntemlerinin kullanıldığı durumlara ilişkin AFA sonuçları ayrı ayrı incelendikten sonra, elde edilen puanların AFA sonuçlarında belirlenen açıklanan varyans değerlerinin ve madde faktör yüklerinin bu üç puanlama durumunda nasıl değişmekte olduğu incelenmiştir. OCAP yönteminin kullanılmasıyla elde edilen puanların AFA sonucunda açıklanan varyans değerinin 0,548 olduğu görülmüştür. MGİAP yöntemi kullanılarak elde edilen puanlara AFA yapıldığında ise açıklanan varyans oranı 0,491 bulunmuştur. Buna göre mevcut araştırmada cevaplayıcıların kısmi bilgileri dikkate alınarak bir ağırlıklandırma işlemi yapıldığında, maddelerin ağırlıklandırıldığı duruma göre AFA sonucunda açıklanan varyans miktarının arttığı söylenebilmektedir. Ancak hem maddelerin hem de olası cevapların ağırlıklandırılmasına dayalı olarak ÇAP yöntemiyle puanlama yapılması durumunda ise AFA sonucunda açıklanan varyans değerinin 0,372 olduğu görülmüştür. Buna göre OCAP yönteminin kullanıldığı durumda AFA'da ortaya çıkan açıklanan varyans değeri, araştırma kapsamında incelenen diğer iki ağırlıklı puanlama yöntemine kıyasla daha büyük bir değer almaktadır.

OCAP, MGİAP ve ÇAP yöntemlerinin kullanıldığı durumlarda elde edilen puanların açımlayıcı faktör analizinin sonucunda elde edilen faktör yükleri Tablo 38’de birlikte verilmiştir.

Tablo 38

Ağırlıklı Puanlama Yöntemlerinin Kullanıldığı Durumlarda Maddelerin Faktör Yük Değerleri

Madde No	Faktör Yük Değeri (OCAP)	Faktör Yük Değeri (MGİAP)	Faktör Yük Değeri (ÇAP)
m1	0,71	0,61	0,62
m2	0,70	0,64	0,62
m3	0,69	0,68	0,58
m4	0,80	0,72	0,68
m5	0,75	0,69	0,57
m6	0,52	0,51	0,41
m7	0,73	0,74	0,55
m8	0,72	0,68	0,57
m9	0,80	0,74	0,68
m10	0,81	0,80	0,61
m11	0,76	0,72	0,62
m12	0,75	0,70	0,60
m13	0,83	0,80	0,68
m14	0,73	0,68	0,60
m15	0,80	0,76	0,67
m16	0,69	0,63	0,58

Tablo 38’de görüldüğü gibi OCAP yöntemiyle elde edilen puanlara dayalı olarak AFA yapıldığında faktör yük değerleri 0,52 ile 0,83 arasında değer almaktadır. MGİAP yönteminin kullanılması durumunda elde edilen puanların AFA sonucunda görülen faktör yük değerleri 0,51 ile 0,80 arasında değer almaktadır. Üçüncü ağırlıklı puanlama yöntemiyle elde edilen puanlara dayalı olarak AFA yapıldığında ise faktör yük değerleri 0,41 ile 0,68 arasında değer almaktadır. Öncelikle OCAP yönteminin kullanıldığı durumda, yedinci madde dışındaki maddelerin her biri için elde edilen faktör yük değerlerinin MGİAP ve ÇAP yöntemlerinin kullanıldığı durumda elde edilen değerlerden daha büyük olduğu görülmektedir. İncelenen puanlama durumlarında elde edilen faktör yük değerleri arasındaki ilişkinin incelenmesi için yöntemler ikiyeşerli olarak gruplanarak Spearman Brown sıra farkları korelasyon katsayısı hesaplanmıştır. OCAP ve MGİAP yöntemleri için elde edilen korelasyon değerine ($r_s = 0,91$) göre, iki puanlama yöntemi arasında faktör yükleri bakımından pozitif yönde yüksek düzeyde uyum olduğu görülmektedir. OCAP ve ÇAP yöntemleriyle ilgili olarak hesaplanan

Spearman Brown sıra farkları korelasyon katsayısı ise $r_s = 0,66$ bulunmuştur. Buna göre bu iki puanlama yöntemi arasında faktör yükleri bakımından pozitif yönde yüksek düzeyde uyum olduğu görülmektedir. MGİAP ve ÇAP yöntemleri için elde edilen korelasyon değerine ($r_s = 0,48$) bakıldığında ise iki puanlama yöntemi arasında faktör yükleri bakımından pozitif yönde orta düzeyde uyum olduğu söylenebilmektedir.

Bu üç puanlama durumunda AFA sonucunda elde edilen faktör yük değerleri arasındaki farkların istatistiksel olarak anlamlılığını incelemek için ayrıca Kruskal Wallis H testi yapılmıştır. Tablo 39’da AFA sonucunda elde edilen faktör yük değerlerinin kullanılan ağırlıklı puanlama yöntemine göre karşılaştırılmasına ilişkin Kruskal Wallis H testi sonuçları yer almaktadır.

Tablo 39

AFA Sonucunda Elde Edilen Faktör Yük Değerlerinin Ağırlıklı Puanlama Yöntemine Göre Karşılaştırılmasına İlişkin Kruskal Wallis H Testi Sonuçları

Yöntem	n	Sıra Ortalaması	sd	χ^2	p	Anlamlı Fark
OCAP	16	34,94	2	23,005	0,000	OCAP>ÇAP MGİAP>ÇAP
MGİAP		26,94				
ÇAP		11,63				

Tablo 39 incelendiğinde AFA sonucunda elde edilen faktör yük değerlerinin kullanılan ağırlıklı puanlama yöntemine göre anlamlı bir şekilde farklılaştığı görülmektedir

($\chi^2(2) = 23,005, p < 0,05$). Ağırlıklı puanlama yöntemlerinin arasında gözlenen bu farkın hangi yöntemler arasında olduğunu bulmak için yapılan Mann Whitney U testi sonuçları ise bu anlamlı farkın OCAP ve ÇAP yöntemleri ($z = -4,266, p < 0,016$) ile MGİAP ve ÇAP yöntemleri arasında olduğunu göstermiştir ($z = -3,52, p < 0,016$). Buna göre mevcut araştırmada ÇAP yöntemi kullanıldığında AFA sonucunda elde edilen faktör yük değerlerinin diğer iki ağırlıklı puanlama yönteminin kullanıldığı durumlarda elde edilen faktör yük değerlerinden anlamlı olarak daha küçük olduğu söylenebilmektedir.

AFA sonuçlarının incelenmesinin ardından kısa yanıtlı testin yanıtlarının OCAP, MGİAP ve ÇAP yöntemiyle puanlanması sonucunda elde edilen puanlara dayalı olarak DFA yapıldığında ortaya çıkan sonuçlar karşılaştırılmıştır. Öncelikle cevaplayıcı yanıtlarının MGAİP yöntemiyle puanlanmasıyla elde edilen puanlara dayalı olarak DFA yapılmıştır. Ölçme aracındaki 16 maddeye ilişkin ölçümlerin faktör yapısı, tek faktörlü

ölçme modeli tanımlanarak DFA ile test edilmiştir. MGİAP yöntemiyle puanlanan maddeler için tanımlanan tek faktörlü modelin test edilmesiyle elde edilen yol şeması ekte sunulmuştur. MGİAP yönteminin kullanıldığı durumda, ölçme aracındaki verilere ilişkin yapılan DFA sonuçlarında elde edilen tüm uyum indeksleri ve hata varyansları birlikte ele alındığında, tanımlanan modelin ilgili veri setine iyi uyum sağladığı ve göstergelerin ilgili değişkenlerin iyi birer temsilcisi olduğu sonucuna ulaşılmıştır.

DFA sonuçlarının incelenmesiyle ilgili son aşamada ÇAP yönteminin kullanıldığı durumda elde edilen puanlara dayalı olarak DFA yapılmıştır. Ölçme aracındaki 16 maddeye ilişkin ölçümlerin faktör yapısı, tek faktörlü ölçme modeli tanımlanarak DFA ile test edilmiştir. ÇAP yöntemiyle puanlanan maddeler için tanımlanan tek faktörlü modelin test edilmesiyle elde edilen yol şeması ekte sunulmuştur. ÇAP yönteminin kullanıldığı durumda, ölçme aracındaki verilere ilişkin yapılan DFA sonuçlarında elde edilen tüm uyum indeksleri ve hata varyansları birlikte ele alındığında, tanımlanan modelin ilgili veri setine iyi uyum sağladığı ve göstergelerin ilgili değişkenlerin iyi birer temsilcisi olduğu sonucuna ulaşılmıştır.

OCAP, MGİAP ve ÇAP yöntemlerinin kullanıldığı durumlara ilişkin DFA sonuçları ayrı ayrı incelendikten sonra, elde edilen uyum indeks değerlerinin ağırlıklı puanlama yöntemine göre nasıl değişmekte olduğu incelenmiştir. Ağırlıklı puanlama yöntemlerinin her birinden elde edilen puanlara uygulanan DFA sonucunda elde edilen uyum indeksleri birlikte Tablo 40'ta verilmiştir.

Tablo 40

OCAP, MGİAP ve ÇAP Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Uyum İndeksleri

Puanlama Yöntemi	RMSEA	CFI	AGFI	NNFI	χ^2 / sd
OCAP	0,055	0,98	0,92	0,98	2,72
MGİAP	0,044	0,98	0,94	0,98	2,15
ÇAP	0,043	0,99	0,94	0,98	2,07

Tablo 40'ta verilen DFA sonuçlarına göre, yanıtların MGİAP ve ÇAP yöntemiyle puanlanması durumlarında RMSEA birbirine çok yakın değerler almaktadır. Ayrıca bu RMSEA değerleri, olası cevapların ağırlıklandırıldığı durumda elde edilen RMSEA değerinden daha küçüktür. Birinci ve ikinci ağırlıklı puanlama yönteminin kullanıldığı durumlarda CFI değerinde değişiklik olmadığı, çoklu ağırlıklandırma yapıldığında ise CFI değerinin küçük bir farkla arttığı görülmüştür. Madde güçlük

indekslerine dayalı ağırlıklandırma ile çoklu ağırlıklandırma yapıldığı durumlarda elde edilen AGFI değerlerinin birbirine eşit olduğu belirlenmiştir. OCAP yönteminin kullanılması durumunda kestirilen AGFI değeri ise bu değerlere göre daha küçüktür. NNFI değerinin kullanılan ağırlıklı puanlama yöntemlerinin her birine ilişkin sonuçlarda eşit olduğu görülmüştür. Son olarak χ^2/sd oranları incelendiğinde, çoklu ağırlıklandırma yapıldığında bu oranın en düşük, OCAP yönteminin kullanılması durumunda ise bu oranın en yüksek olduğu belirlenmiştir.

OCAP, MGİAP ve ÇAP yöntemlerinin kullanıldığı durumlarda elde edilen puanların DFA sonucunda hesaplanan hata varyansları Tablo 41’de birlikte verilmiştir.

Tablo 41

Ağırlıklı Puanlama Yöntemlerinin Kullanıldığı Durumlarda DFA Sonucunda Görülen Hata Varyansları

Madde No	Hata Varyans Değeri (OCAP)	Hata Varyans Değeri (MGİAP)	Hata Varyans Değeri (ÇAP)
m1	0,62	0,83	0,65
m2	0,66	0,76	0,65
m3	0,70	0,73	0,71
m4	0,55	0,72	0,57
m5	0,60	0,72	0,72
m6	0,83	0,86	0,86
m7	0,64	0,67	0,73
m8	0,65	0,71	0,71
m9	0,53	0,67	0,57
m10	0,51	0,57	0,67
m11	0,57	0,68	0,66
m12	0,59	0,70	0,68
m13	0,50	0,57	0,57
m14	0,64	0,73	0,67
m15	0,58	0,64	0,59
m16	0,66	0,76	0,70

Tablo 41’de görüldüğü gibi olası cevapların ağırlıklandırılmasıyla elde edilen puanlara dayalı olarak DFA yapıldığında hesaplanan hata varyansları 0,51 ile 0,83 arasında değişmektedir. MGİAP yönteminin kullanılması durumunda elde edilen puanların DFA sonucunda görülen hata varyansları 0,57 ile 0,86 arasında değişmektedir. Üçüncü ağırlıklı puanlama yöntemiyle elde edilen puanlara dayalı olarak DFA yapıldığında ise hata varyansları 0,57 ile 0,86 arasında değer almaktadır. Öncelikle OCAP yönteminin kullanıldığı durumda, ikinci madde dışındaki maddelerin her biri için

elde edilen hata varyanslarının MGİAP ve ÇAP yöntemlerinin kullanıldığı durumda elde edilen hata varyanslarından daha küçük olduğu görülmektedir. İncelenen ağırlıklı puanlama durumlarında elde edilen hata varyansları arasındaki ilişkinin incelenmesi için yöntemler ikiye ayrılarak gruplanarak Spearman Brown sıra farkları korelasyon katsayısı hesaplanmıştır. OCAP ve MGİAP yöntemleri için elde edilen korelasyon değerine ($r_s = 0,78$) göre, iki puanlama yöntemi arasında hata varyansları bakımından pozitif yönde yüksek düzeyde uyum olduğu görülmektedir. OCAP ve ÇAP yöntemleriyle ilgili olarak hesaplanan Spearman Brown sıra farkları korelasyon katsayısı ise $r_s = 0,68$ bulunmuştur. Buna göre bu iki puanlama yöntemi arasında hata varyansları bakımından pozitif yönde yüksek düzeyde uyum olduğu görülmektedir. MGİAP ve ÇAP yöntemleri için elde edilen korelasyon değerine ($r_s = 0,34$) bakıldığında ise iki puanlama yöntemi arasında DFA sonuçlarında görülen hata varyansları bakımından pozitif yönde düşük düzeyde uyum olduğu söylenebilmektedir.

Bu üç ağırlıklı puanlama durumunda DFA sonucunda elde edilen hata varyansları arasındaki farkların istatistiksel olarak anlamlı olup olmadığını belirlemek için ayrıca Kruskal Wallis H testi yapılmıştır. Tablo 42’de DFA sonucunda elde edilen hata varyanslarının kullanılan ağırlıklı puanlama yöntemine göre karşılaştırılmasına ilişkin Kruskal Wallis H testi sonuçları verilmiştir.

Tablo 42

<i>DFA Sonucunda Elde Edilen Hata Varyanslarının Ağırlıklı Puanlama Yöntemine Göre Karşılaştırılmasına İlişkin Kruskal Wallis H Testi Sonuçları</i>						
Yöntem	n	Sıra Ortalaması	sd	χ^2	p	Anlamlı Fark
OCAP	16	15,47	2	11,985	0,002	OCAP<MGİAP
MGİAP		32,47				
ÇAP		25,56				

Tablo 42 incelendiğinde DFA sonucunda elde edilen hata varyanslarının kullanılan ağırlıklı puanlama yöntemine göre anlamlı bir şekilde farklılaştığı görülmektedir ($\chi^2(2) = 11,985, p < 0,05$). Ağırlıklı puanlama yöntemlerinin arasında gözlenen bu farkın hangi yöntemler arasında olduğunu bulmak için yapılan Mann Whitney U testi sonuçları ise bu anlamlı farkın sadece OCAP ve MGİAP yöntemleri arasında olduğunu göstermiştir ($z = -3,17, p < 0,016$). Buna göre mevcut araştırmada OCAP yöntemi kullanıldığında DFA sonucunda elde edilen hata varyanslarının MGİAP yönteminin kullanıldığı durumda elde edilen hata varyanslarından anlamlı olarak daha küçük olduğu söylenebilmektedir.

Üçüncü alt problemle ilgili olarak OCAP, MGİAP ve ÇAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin yapı geçerliğinin incelenmesinin ardından uygunluk geçerliğine yönelik bulgular incelenmiştir. Uygunluk geçerliği kapsamında, ağırlıklı puanlama yöntemlerinin her birinden elde edilen toplam puanların öğrencilerin matematik dersine ilişkin bir önceki öğretim yılı sonundaki not ortalamasıyla arasındaki ilişkilerin nasıl değişmekte olduğu incelenmiştir. Bu doğrultuda OCAP yönteminin kullanılması durumunda toplam puanlar ile öğrencilerin bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyonun $r=0,70$ ($p<0,05$) olduğu görülmüştür. MGİAP yönteminin kullanıldığı ikinci durumda elde edilen toplam puanlar ile cevaplayıcıların bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon hesaplanmış ve bu değer $r= 0,64$ ($p<0,05$) bulunmuştur. Son olarak puanlama yapılırken çoklu ağırlıklandırmanın kullanıldığı durumda hesaplanan toplam puanlar ile öğrencilerin bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon değeri $r=0,66$ ($p<0,05$) bulunmuştur.

Üç farklı puanlama durumuna ilişkin elde edilen bağımlı korelasyon değerleri arasında istatistiksel olarak anlamlı bir fark olup olmadığını belirlemek için bu korelasyon değerleri ikiyeşerli gruplar halinde karşılaştırılmıştır. Bu doğrultuda ağırlıklı puanlama yöntemlerinden elde edilen puanların öğrencilerin matematik dersine ilişkin bir önceki öğretim yılı sonundaki not ortalamasıyla aralarındaki ilişkilerin karşılaştırılmasına yönelik test sonuçları (Meng, Rosenthal & Rubin, 1992) incelenmiştir. MGİAP yöntemi kullanıldığında bulunan toplam puanlar ile cevaplayıcıların matematik dersi not ortalamaları arasındaki korelasyon değeri ($r=0,64$) ve OCAP yönteminin kullanılması durumunda toplam puanlar ile öğrencilerin matematik dersi not ortalamaları arasındaki korelasyon değeri ($r=0,70$) karşılaştırıldığında hesaplanan Z değeri $Z= 14,73$ bulunmuştur. Hesaplanan Z değeri, $\alpha=0,05$ düzeyinde kritik tablo değeri olan 1,96 değerinden büyük olduğundan incelenen iki farklı puanlama durumuna ilişkin elde edilen korelasyon değerleri arasında anlamlı bir fark olduğu söylenebilmektedir. Buna göre MGİAP yönteminin kullanıldığı durumda elde edilen cevaplayıcıların toplam puanları ile bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon değeri, OCAP yöntemiyle bulunan toplam puanlar ile öğrencilerin bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon değerinden anlamlı olarak daha küçüktür.

Bir sonraki aşamada MGİAP yöntemi kullanıldığında bulunan toplam puanlar ile cevaplayıcıların matematik dersi not ortalamaları arasındaki korelasyon değeri ($r=0,64$) ve ÇAP yönteminin kullanılması durumunda toplam puanlar ile öğrencilerin matematik dersi not ortalamaları arasındaki korelasyon değeri ($r=0,66$) karşılaştırılmıştır. Söz konusu iki korelasyon değerine göre hesaplanan Z değeri $Z= 7,72$ bulunmuştur. Hesaplanan Z değeri, $\alpha=0,05$ düzeyinde kritik tablo değeri olan 1,96 değerinden büyük olduğundan incelenen iki farklı puanlama durumuna ilişkin elde edilen korelasyon değerleri arasında anlamlı bir fark olduğu söylenebilmektedir. Böylece MGİAP yönteminin kullanıldığı durumda elde edilen cevaplayıcıların toplam puanları ile bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon değeri, ÇAP yöntemiyle bulunan toplam puanlar ile öğrencilerin bir önceki öğretim yılı sonundaki matematik dersi not ortalamaları arasındaki korelasyon değerinden anlamlı olarak daha küçüktür.

Dördüncü alt problemle ilgili son olarak OCAP, MGİAP ve ÇAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin güvenilirlik katsayılarının nasıl değişmekte olduğu araştırılmıştır. Söz konusu üç puanlama durumunda elde edilen puanlara ilişkin güvenilirlik katsayıları hesaplanarak bu güvenilirlik katsayıları arasında anlamlı bir fark olup olmadığı incelenmiştir. Bu doğrultuda OCAP, MGİAP ve ÇAP yönteminin kullanılması durumunda elde edilen puanlara dayalı olarak hesaplanan güvenilirlik katsayıları Tablo 43'te verilmiştir.

Tablo 43

Ağırlıklı Puanlama Yöntemlerinin Kullanıldığı Durumlarda Elde Edilen Güvenirlik Katsayıları

Yöntem	Cronbach α	McDonald ω	Anlamlı fark (Cronbach α)	Anlamlı fark (McDonald ω)
OCAP	0,91	0,92	OCAP>MGİAP OCAP>ÇAP	OCAP>MGİAP OCAP>ÇAP
MGİAP	0,84	0,88		
ÇAP	0,87	0,89		

Tablo 43'e göre OCAP yönteminin kullanıldığı durumda Cronbach alfa güvenilirlik katsayısının $\alpha= 0,91$ olduğu görülmüştür. Bu puanlama yöntemiyle ilgili olarak McDonald'ın ω katsayısı ise $\omega= 0,92$ hesaplanmıştır. MGİAP yönteminin kullanıldığı durumda elde edilen puanlara dayalı olarak güvenilirlik analizi yapıldığında Cronbach alfa güvenilirlik katsayısı $\alpha= 0,84$ bulunmuştur. Aynı durumda güvenilirlik

kestirimi için McDonald'ın ω katsayısı ise $\omega = 0,88$ hesaplanmıştır. Bu iki durumda elde edilen Cronbach alfa değerleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için gerçekleştirilen Fisher'in z testi sonucu incelendiğinde, hesaplanan z değeri $z = 5,34$ bulunmuştur. Bu değer 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığının (Kenny, 1987) dışında olduğu görülmektedir. Buna göre OCAP ve MGİAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenilirlik analizlerinin sonucunda hesaplanan Cronbach alfa değerleri arasında anlamlı bir farklılığın bulunduğu söylenebilmektedir. OCAP yönteminin kullanıldığı durumda kestirilen Cronbach alfa güvenilirlik katsayısı, MGİAP kullanılması durumunda kestirilen Cronbach alfa katsayısından istatistiksel olarak anlamlı bir şekilde daha büyüktür. Benzer şekilde iki duruma ilişkin McDonald'ın ω katsayıları arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için gerçekleştirilen Fisher'in z testi sonucunda $z = 3,69$ bulunmuştur. Bu değer de 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığının (Kenny, 1987) dışında olduğu görülmektedir. Böylece OCAP ve MGİAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenilirlik analizlerinin sonucunda hesaplanan McDonald'ın ω katsayıları arasında anlamlı bir farklılığın bulunduğu söylenebilmektedir. OCAP yönteminin kullanıldığı durumda kestirilen McDonald'ın ω katsayısı, MGİAP kullanılması durumunda kestirilen McDonald'ın ω katsayısından istatistiksel olarak anlamlı bir şekilde daha büyüktür.

Tablo 43'te verilen ÇAP yöntemiyle ilgili sonuçlar incelendiğinde Cronbach alfa güvenilirlik katsayısının $\alpha = 0,87$ olduğu görülmektedir. Aynı durumda güvenilirlik kestirimi için McDonald'ın ω katsayısı ise $\omega = 0,89$ hesaplanmıştır. OCAP ve ÇAP kullanılması durumlarında elde edilen Cronbach alfa değerleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için gerçekleştirilen Fisher'in z testi sonucu incelendiğinde, hesaplanan z değeri $z = 3,44$ bulunmuştur. Bu değer 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığının (Kenny, 1987) dışında olduğu görülmektedir. Buna göre OCAP ve ÇAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenilirlik analizlerinin sonucunda hesaplanan Cronbach alfa değerleri arasında anlamlı bir farklılığın bulunduğu söylenebilmektedir. OCAP yönteminin kullanıldığı durumda kestirilen Cronbach alfa güvenilirlik katsayısı, ÇAP kullanılması durumunda kestirilen Cronbach alfa katsayısından istatistiksel olarak anlamlı bir şekilde daha büyüktür. Benzer şekilde iki duruma ilişkin McDonald'ın ω katsayıları arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için

gerçekleştirilen Fisher'in z testi sonucunda $z= 2,91$ bulunmuştur. Bu değerin de 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığının (Kenny, 1987) dışında olduğu görülmektedir. Böylece OCAP ve ÇAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenirlik analizlerinin sonucunda hesaplanan McDonald'ın ω katsayıları arasında anlamlı bir farklılığın bulunduğu söylenebilmektedir. OCAP yönteminin kullanıldığı durumda kestirilen McDonald'ın ω katsayısı, ÇAP kullanılması durumunda kestirilen McDonald'ın ω katsayısından istatistiksel olarak anlamlı bir şekilde daha büyüktür.

Son olarak MGİAP ve ÇAP kullanılması durumlarında elde edilen Cronbach alfa değerleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için gerçekleştirilen Fisher'in z testi sonucu incelendiğinde, hesaplanan z değeri $z= 1,89$ bulunmuştur. Bu değerin 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığında (Kenny, 1987) olduğu görülmektedir. Buna göre MGİAP ve ÇAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenirlik analizlerinin sonucunda hesaplanan Cronbach alfa değerleri arasında anlamlı bir farklılığın olmadığı söylenebilmektedir. Benzer şekilde iki duruma ilişkin McDonald'ın ω katsayıları arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için gerçekleştirilen Fisher'in z testi sonucunda $z= 0,69$ bulunmuştur. Bu değerin de 0,05 anlamlılık düzeyinde kritik değerler olan -1.96 ile +1.96 aralığında (Kenny, 1987) olduğu görülmektedir. Böylece MGİAP ve ÇAP yönteminin kullanıldığı durumlarda elde edilen puanların güvenirlik analizlerinin sonucunda hesaplanan McDonald'ın ω katsayıları arasında anlamlı bir farklılığın bulunmadığı söylenebilmektedir.

Dördüncü alt probleme ilişkin elde edilen bulgulara göre, genel olarak OCAP yönteminin kullanılmasının klasik test kuramına dayalı olarak elde edilen geçerlik ve güvenirlik kanıtları bakımından diğer ağırlıklı puanlama yöntemlerine kıyasla daha büyük bir avantaj sağlayabileceği söylenebilmektedir. Alanyazın incelendiğinde ağırlıklı puanlama yöntemlerinin kullanıldığı durumlarda sadece maddelerin, sadece seçeneklerin veya hem maddelerin hem de seçeneklerin ağırlıklandırıldığı araştırmaların gerçekleştirildiği görülmüştür. Ancak bu araştırmalarda genellikle ağırlıklı puanlama yöntemlerinin geleneksel puanlama yöntemine kıyasla testin psikometrik özelliklerine katkılarının incelenmesi amaçlamıştır. Madde ağırlıklandırmayla ilgili olarak Gözen (2002), madde ağırlıklandırmanın test güvenirliği ve geçerliğine ilişkin geleneksel puanlamayla benzer sonuçlar verdiğini ifade etmiştir. Quist (2004) madde ağırlıklandırma yöntemiyle puanlamanın geleneksel puanlamaya göre puanların

özellikle ölçüte dayalı geçerliğini ve yordama geçerliğini artırmada önemli bir katkı sağladığını belirtmiştir. Kılıç ve Doğan (2019) önerdikleri madde ağırlıklandırma yöntemiyle elde edilen puanların geleneksel yöntemle elde edilen puanlara kıyasla güvenilirlik katsayısının daha yüksek olduğunu belirlemiştir. Çalışmada ayrıca kullanılan madde ağırlıklandırmanın testin yapı geçerliğine katkı sağladığı da belirtilmiştir. Çoklu ağırlıklandırma yöntemiyle ilgili olarak ise Haladyna ve Sympson (1988) ve Sympson (1993) çalışmalarında bu puanlama yaklaşımının geleneksel puanlamaya göre puanların güvenilirliğini ve geçerliğini arttırdığını belirlemiştir. Başka bir çalışmada Crehan ve Haladyna (1994), çoklu ağırlıklandırma yöntemiyle birlikte deneysel seçenek ağırlıklandırma yöntemlerini ve geleneksel puanlamayı kullanmış ve bu çalışmada elde edilen bulgulardan farklı bir şekilde güvenilirlik ve geçerlik kanıtları bakımından çoklu ağırlıklandırma ve deneysel seçenek ağırlıklandırma yöntemlerinin birbirine çok yakın sonuçlar verdiği görülmüştür.

BÖLÜM 5

SONUÇLAR VE ÖNERİLER

Bu bölümde, araştırma kapsamında elde edilen bulgulara dayalı sonuçlar verilmiş ve öneriler sunulmuştur.

Sonuçlar

- Klasik test kuramına dayalı olarak elde edilen kanıtlara göre genel olarak GP yöntemi yerine OCAP yönteminin kullanılmasının geçerlik ve güvenirlik bakımından avantaj sağlayabileceği görülmüştür.
- GP ve ardından OCAP yönteminin kullanılması durumunda, klasik test kuramına göre yapı geçerliğiyle ilgili olarak OCAP yöntemi kullanıldığında AFA sonucunda görülen açıklanan varyans değerinin daha büyük olduğu belirlenmiştir. Araştırma bulguları doğrultusunda cevaplayıcıların kısmi bilgileri dikkate alınarak bir ağırlıklandırma işlemi yapıldığında, AFA sonucunda açıklanan varyans miktarının ilk duruma göre arttığı görülmüştür.
- Maddelerin faktör yük değerleri incelendiğinde, OCAP yöntemi kullanıldığında elde edilen faktör yük değerlerinin GP yönteminin kullanıldığı durumda elde edilen değerlerden anlamlı olarak daha büyük olduğu görülmüştür.
- GP ve OCAP yöntemleriyle puanlama yapılması durumlarında elde edilen puanlara dayalı olarak DFA yapıldığında, uyum indekslerinden RMSEA değerinin GP yönteminin kullanıldığı durumda, OCAP yönteminin kullanıldığı duruma göre daha küçük olduğu görülmüştür. GP ve OCAP yönteminin kullanıldığı durumlarda CFI ve NNFI değerlerinde herhangi bir değişiklik olmadığı belirlenmiştir. OCAP yönteminin kullanılması durumunda kestirilen AGFI değerinin GP yönteminin kullanıldığı durumda elde edilen AGFI değerine göre daha küçük olduğu görülmüştür. Son olarak DFA sonuçlarına dayalı olarak hesaplanan χ^2/sd oranları incelendiğinde, GP yöntemi kullanıldığında bu

oranın OCAP yönteminin kullanıldığı duruma göre daha düşük olduğu belirlenmiştir.

- GP ve OCAP yöntemlerinin kullanıldığı durumlarda elde edilen puanlara dayalı olarak yapılan DFA sonucunda elde edilen hata varyanslarının kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaştığı görülmüştür. OCAP yöntemi kullanıldığında elde edilen hata varyanslarının GP yönteminin kullanıldığı durumda elde edilen değerlerden anlamlı olarak daha küçük olduğu belirlenmiştir.
- GP ve OCAP yöntemlerinin kullanılması durumunda, klasik test kuramına göre uygunluk geçerliğiyle ilgili olarak iki ayrı puanlama durumunda elde edilen toplam puanlarla öğrencilerin bir önceki öğretim yılına ait matematik dersi yıl sonu puanları arasındaki korelasyon değerleri arasında anlamlı bir fark olduğu görülmüştür. OCAP yönteminin kullanıldığı durumdaki toplam puanlar ile ders puanları arasındaki korelasyon değeri, GP yöntemine ilişkin hesaplanan korelasyon değerinden anlamlı olarak daha büyük bulunmuştur.
- Önce GP, daha sonra OCAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin klasik test kuramına dayalı olarak kestirilen güvenilirlik katsayıları karşılaştırıldığında, Cronbach alfa ve McDonald ω değerleri arasında anlamlı bir farklılığın bulunduğu görülmüştür. OCAP yönteminin kullanıldığı durumda kestirilen Cronbach alfa ve McDonald ω güvenilirlik katsayılarının GP durumunda kestirilen değerlerden anlamlı bir şekilde daha büyük olduğu belirlenmiştir.
- GP ve OCAP yöntemlerinin kullanılması durumunda, madde tepki kuramına dayalı olarak elde edilen geçerlik ve güvenilirliğe ilişkin bulgular incelendiğinde, OCAP yönteminin kullanıldığı durumda cevaplayıcıların yeteneğine ilişkin testten alınan bilgi düzeyinin arttığı görülmüştür. Ayrıca OCAP yönteminin kullanıldığı durumda kestirilen güvenilirlik katsayısının GP yapılması durumunda kestirilen değerden daha büyük olduğu belirlenmiştir. Böylece madde tepki kuramına dayalı çalışmalar sonucunda OCAP yönteminin GP yöntemine göre geçerlik ve güvenilirlik kanıtlarında bir iyileşme ortaya koyabileceği düşünülmüştür.
- Araştırmada puanlama sürecinde önce “1-0” puanlama, daha sonra OCAP yönteminin kullanılmasıyla elde edilen orijinal veriden rastgele veri setleri

üretilek bu aşamaya kadar elde edilen sonuçların tutarlılığı incelendiğinde, GP ve OCAP yöntemleri için AFA ile belirlenen açıklanan varyans değerklerinin kullanılan puanlama yöntemine göre anlamlı olarak farklılaştığı görölmüştür. OCAP yöntemi kullanıldığında elde edilen açıklanan varyans değerklerinin GP yöntemi kullanıldığında elde edilen değerklerden anlamlı olarak daha büyük olduğu belirlenmiştir.

- GP ve OCAP yönteminin kullanılmasıyla elde edilen veriden rastgele oluşturulan veri setlerinin her biri için AFA sonuçlarında görölen ortalama faktör yük değerkleri ise kullanılan puanlama yöntemine göre anlamlı bir şekilde farklılaşmamıştır.
- Üretilen veriyle DFA yapıldığında elde edilen ortalama uyum indeks değerkleri puanlama yöntemlerine göre karşılaştırıldığında, GP ve OCAP yönteminin kullanılmasıyla elde edilen orijinal veriden elde edilen sonuçların değışmediğı görölmüştür. Benzer şekilde OCAP yöntemi kullanıldığında elde edilen ortalama hata varyanslarının GP yönteminin kullanıldığı durumda elde edilen değerklerden anlamlı olarak daha küçük olduğu belirlenmiştir.
- GP ve OCAP yöntemlerinin kullanıldığı durumlara ilişkin rastgele oluşturulan veri setlerinin her biri için güvenilirlik analizi yapıldığında OCAP yöntemi kullanıldığında elde edilen Cronbach alfa ve McDonald ω değerklerinin GP yöntemi kullanıldığında elde edilen değerklerden anlamlı olarak daha büyük olduğu görölmüştür. Böylece bu aşamadaki tüm geçerlik ve güvenilirlik çalışmaları göz önünde bulundurularak rastgele oluşturulan veri setleri için madde faktör yükleri dışındaki tüm sonuçların orijinal verinin kullanıldığı durumdaki sonuçlarla örtüştüğü görölmüştür.
- OCAP, MGİAP ve ÇAP yöntemlerinin kullanıldığı durumlarda elde edilen puanların AFA sonuçlarına göre, OCAP yönteminin kullanıldığı durumdaki açıklanan varyans değerklerinin araştırma kapsamında incelenen diğerk iki ağırlıklı puanlama yöntemine kıyasla daha büyük olduğu belirlenmiştir. Ayrıca ÇAP yöntemi kullanıldığında AFA sonucunda elde edilen faktör yük değerklerinin diğerk iki ağırlıklı puanlama yönteminin kullanıldığı durumlarda elde edilen faktör yük değerklerinden anlamlı olarak daha küçük olduğu görölmüştür.
- OCAP, MGİAP ve ÇAP yöntemlerinin kullanıldığı durumlara ilişkin DFA sonuçlarında elde edilen uyum indeks değerklerinin söz konusu puanlama

durumlarında birbirlerine oldukça yakın değerler aldığı belirlenmiştir. Ayrıca OCAP yöntemi kullanıldığında DFA sonucunda elde edilen hata varyanslarının MGİAP yönteminin kullanıldığı durumda elde edilen hata varyanslarından anlamlı olarak daha küçük olduğu görülmüştür.

- OCAP, MGİAP ve ÇAP yönteminin kullanılması durumunda elde edilen puanlara ilişkin uygunluk geçerliğiyle ilgili olarak OCAP yönteminin kullanıldığı durumdaki toplam puanlar ile ders puanları arasındaki korelasyon değerinin, MGİAP ve ÇAP yöntemlerine ilişkin hesaplanan korelasyon değerlerinden anlamlı olarak daha büyük olduğu görülmüştür.
- Son olarak OCAP yöntemi kullanıldığında elde edilen Cronbach alfa ve McDonald ω değerlerinin MGİAP ve ÇAP yöntemleri kullanıldığında elde edilen değerlerden anlamlı olarak daha büyük olduğu belirlenmiştir. Genel olarak OCAP yönteminin kullanılmasının klasik test kuramına dayalı olarak elde edilen kanıtlara göre geçerlik ve güvenirlik bakımından diğer ağırlıklı puanlama yöntemlerine kıyasla daha büyük bir avantaj sağlayabileceği ifade edilmiştir.

Öneriler

Araştırmadan elde edilen sonuçlar doğrultusunda, uygulayıcılar ve araştırmacılar için aşağıdaki öneriler sunulmuştur.

Uygulayıcılar İçin Öneriler

- Araştırmanın sonuçlarına göre, kısa cevaplı testlerde olası cevapların deneysel olarak ağırlıklandırılmasına dayalı olan ve bu araştırma kapsamında önerilen puanlama yönteminin geleneksel puanlama yöntemi ile incelenen diğer iki ağırlıklı puanlama yöntemine kıyasla geçerlik ve güvenirlik bakımından daha iyi sonuçlar ortaya koyabileceği söylenebilmektedir. Buna göre öncelikle bu ağırlıklandırma yönteminin uygulayıcılar tarafından kullanılması önerilebilir.
- İncelenen puanlama yöntemi kısa cevaplı testler için de puanlama kolaylığı sağlayan aynı zamanda öğrencilerin kısmi bilgisini dikkate alan bir puanlama sağladığı için okul genelinde yapılacak ölçme ve değerlendirme uygulamalarında kullanılabilir.

Araştırmacılar İçin Öneriler

- Araştırmada incelenen ağırlıklandırma yönteminin testin psikometrik özelliklerine katkısı, bu araştırma kapsamında tek boyutlu bir yapıyla araştırılmıştır. Bu doğrultuda çok boyutlu yapılar için söz konusu ağırlıklandırma yönteminin üreteceği sonuçların incelenmesi araştırmacılara önerilebilir.
- Farklı sınıf düzeyi, farklı ders veya konuya ilişkin bu araştırmada incelenen ağırlıklandırma yönteminin testin psikometrik özellikleriyle ilgili performansı değerlendirilebilir.
- Matematik dışındaki bir derse yönelik geliştirilecek bir kısa cevaplı testte bu araştırmada incelenen ağırlıklandırma yönteminin kullanılabileceği maddeler ve olası cevapların araştırılabileceği düşünülmektedir.
- Bu araştırmada cevaplayıcıların kısmi bilgileriyle oluşturduğu olası cevaplar, cevabın görülme durumu ile toplam test puanı arasındaki nokta çift serili korelasyon katsayıları aracılığıyla deneysel olarak ağırlıklandırılmıştır. Olası cevapların bu yöntemin dışında örneğin uzman yargısına dayalı olarak ağırlıklandırılması durumunda veya deneysel ağırlıklandırmanın başka bir şekilde yapılması durumunda elde edilecek sonuçların araştırılması önerilebilmektedir.
- İncelenen puanlama yöntemlerinin yapı geçerliğine ilişkin sonuçlar ölçme değişmezliği çalışmalarıyla karşılaştırılabilir.
- Olası cevapların deneysel olarak ağırlıklandırılmasına dayalı puanlama yöntemi ile bu araştırmada kullanılan yöntemlerden farklı ağırlıklı puanlama yöntemleri kullanılarak elde edilecek güvenirlik ve geçerlik kanıtlarının karşılaştırmalı olarak incelenmesi araştırmacılara önerilebilir.

KAYNAKLAR

- Abu-Sayf, F.K. (1979). The Scoring for Multiple-Choice Tests: A Closer Look. *Educational Technology*, 19, 5-15.
- Acar, T. (2018). Yazılı Yoklamalarda Puanlama Yönteminin Test ve Madde İstatistiklerine Etkisi. *İlköğretim Online*, 17(2), 500-509.
- Akkuş, O. (2000). Çoktan seçmeli test maddelerini puanlamada, seçenekleri farklı biçimlerde ağırlıklandırmanın madde ve test istatistiklerine olan etkisinin incelenmesi. (Yüksek Lisans Tezi). Hacettepe Üniversitesi, Ankara.
- Akkuş, O., & Baykul, Y. (2001). Çoktan seçmeli test maddelerini puanlamada, seçenekleri farklı biçimlerde ağırlıklandırmanın madde ve test istatistiklerine olan etkisinin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 20(9), 15.
- Andrade, H. L., & Cizek, G. J. (2010). *Handbook of Formative Assessment*. New York, NY: Routledge.
- Atılğan, H., Kan, A., & Doğan, N. (2011). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Anı Yayıncılık.
- Baker, F. B. (2001). The basics of item response theory. ERIC.
- Bauer, D., Holzer, M., Kopp, V., & Fischer, M. R. (2011). Pick-N multiple choice-exams: a comparison of scoring algorithms. *Advances in health sciences education*, 16, 211-221.
- Baykul, Y. (2015). Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması. Ankara: Pegem.
- Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A Comparative Study of Measures of Partial Knowledge in Multiple- Choice Tests. *Applied Psychological Measurement*, 21 (1), 65-88.
- Birnbaum, A. (1968). Some latent models and their use in inferring an examinee's ability.
- Burt, C. (1950). The influence of differential weighting. *British Journal of Statistical Psychology*, 3(2), 105–125. <https://doi.org/10.1111/j.2044-8317.1950.tb00288.x>
- Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., Karadeniz, Ş. ve Demirel, F. (2012). *Bilimsel Araştırma Yöntemleri*. Ankara: Pegem Akademi.
- Cannon, G. E. (1994). *Empirical option weighting and polytomous scoring of multiple choice tests*. Doctor of Philosophy, The University of Arizona, Arizona.

- Cherkas, B. M. and Roitberg, J. (1993). Humanizing the multiple-choice test with partial credit. *International Journal of Mathematical Education in Science and Technology*, 24(6), 799-812.
- Childs, R. A., Dunn, J. L., van Barneveld, C., Jaciw, A. P., & McIlroy, J. H. (2003). Differential weighting of errors on a test of clinical reasoning skills. *Academic Medicine*, 78 (10), 62-64.
- Cohen. J. R., Swerdlik M. E., & Phillips, S. M. (1996). *Psychological testing and assessment*. London: Mayfield.
- Coombs, C. H., Milholland, J. E. & Womer, F.B. (1956). The Assessment of Partial Knowledge. *Educational and Psychological Measurement*, 16, 13-37.
- Cooper, C. (2019). *Psychological testing: Theory and practice*. Abingdon, Oxon: Routledge.
- Crehan, K. D. & Haladyna, T. M. (1994). A Comparison of Three Linear Polytomous Scoring Methods. *Educational Resources Information Center (ERIC)*, Report No: ED 377 246.
- Crocker, L. ve Algina J. (1986). *Introduction to classical and modern test theory*. Orlando: Harcourt Brace Jovanovich Inc.
- Cross, L. H., Ross, F. K., & Geller, E. S. (1980). Using choice-weighted scoring of multiple-choice tests for determination of grades in College courses. *The Journal of Experimental Education*, 48(4), 296-301.
- Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2012). *Sosyal Bilimler İçin Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları*. Ankara: Pegem.
- Davis, F. B. & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 19, 159-170.
- Davis, L.L.(1992). "Instrument review: Getting the most from a panel of experts". *Applied Nursing Research*, 5, 194-197.
- de Ayala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement & Evaluation in Counseling & Development*, 25(4), 172-189.
- DeMars, C. (2016). *Madde tepki kuramı*. H. Kelecioğlu (Çev. Ed.). Ankara: Nobel.
- DeMars, C. E. (2008, March). Scoring multiple choice items: A comparison of IRT and classical polytomous and dichotomous methods. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
- Deniz, K. Z. (2020). *Herkes için istatistikolay*. Ankara: Nobel Yayınevi.
- Dick, W. (1965). Item weighting: Test parameter effects and comparison of the efficiency of various weighting methods. (Doktora Tezi).

- Diedenhofen, B. & Musch, J. (2015). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment*, 33, 336-344.
- Diedenhofen, B., & Musch, J. (2019). Option weights should be determined empirically and not by experts when assessing knowledge with multiple-choice items. *International Journal of Selection and Assessment*. doi:10.1111/ijsa.12252
- Doğan, N. (2020). Geleneksel Ölçme ve Değerlendirme Teknikleri II: Yanıtı Yapılandırmayı Gerektiren Ölçme Araçları. N. Doğan (Editör), *Eğitimde Ölçme ve Değerlendirme* (ss. 139-180). Ankara: Pegem Akademi.
- Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, 31, 295-311.
- Echternacht, G. (1976). Reliability and validity of item option weighting schemes. *Educational and Psychological Measurement*, 36, 301-309.
- Erdem, B., Ertuna, L., ve Doğan, N. (2016). Çoktan Seçmeli Testlerde Seçenek Ağırlıklandırma Yöntemlerinin Testin Faktör Yapısına Etkisinin İncelenmesi. V. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi* (ss. 148-149). Erişim adresi: <http://epod2016.akdeniz.edu.tr/dinamik/333/53.pdf>.
- Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.). Charlotte, NC: IAP.
- Frery, R. (1980). The Effect of Misinformation, Partial Information and Guessing on Expected Multiple-Choice Test Item Scores. *Applied Psychological Measurement*, 4 (1), 79-90.
- Frery, R. (1982). A simulation study of reliability and validity of multiple choice test scores under six response-scoring modes. *Journal of Educational Statistics*. 7(4), 23-41.
- Frery, R. (1989). Partial Credit Scoring Methods For Multiple Choice Tests. *Applied Psychological Measurement*, 2 (1), 79-96.
- García-Pérez, M. A., & Frery, R. B. (1989). Psychometric properties of finite-state scores versus number-correct and formula scores: A simulation study. *Applied Psychological Measurement*, 13(4), 403-417.
- Garrett, P. (2009). A Monte Carlo study investigating missing data, differential item functioning and effect size (Doctoral Dissertation). Georgia State University, ABD.
- Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York: McGraw-Hill.

- Gözen Çıtak, G. (2007). Klasik test ve madde-tepki kuramlarına göre çoktan seçmeli testlerde farklı puanlama yöntemlerinin karşılaştırılması (Doktora Tezi). Ankara Üniversitesi, Ankara.
- Gözen, G. (2002). Kısa cevaplı ve çoktan seçmeli maddelerin "1-0" ve ağırlıklı puanlama yöntemleriyle puanlanmasının testin psikometrik özellikleri açısından incelenmesi. (Yüksek Lisans Tezi). Hacettepe Üniversitesi, Ankara.
- Grunert, M. L., Raker, J. R., Murphy, K. L. and Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: development of a rubric for rating partial credit. *Journal of Chemical Education*, 90(10), 1310-1315.
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
- Guilford, J. P., Lovell, C. & Williams, R. M. (1942). Completely weighted versus unweighted scoring in an achievement examination. *Educational and Psychological Measurement*, 2, 15-21.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Gültekin, S. (2017). Testlerde Kullanılacak Madde Türleri, Hazırlama İlkeleri ve Puanlaması. R. N. Demirtaşlı (Editör), *Eğitimde Ölçme ve Değerlendirme* (ss. 145-222). Ankara: Anı Yayıncılık.
- Gültekin, S., & Çıkrıkçı Demirtaşlı, N. (2012). Çoktan seçmeli, açık uçlu ve karma testlerden sağlanan bilginin madde tepki kuramına dayalı olarak karşılaştırılması. *İlköğretim Online*, 11(1), 251-263.
- Haladyna, T. M. (1990). Effects of Empirical Option Weighting on Estimating Domain Scores and Making Pass/Fail Decisions. *Applied Measurement in Education*, 3 (3), 231-244.
- Haladyna, T. M. (1994). *Developing and validating multiple choice items*. Hishalde: Lawrence Erlbaum Associates, Publishers.
- Haladyna, T. M. (1997). Writing test items to evaluate higher order thinking. Boston: Allyn and Bacon.
- Haladyna, T. M., & Sympson, J. B. (1988). Empirically based polychotomous scoring of multiple-choice items: Historical overview. *In an-nual meeting of the American Educational Research Association*, New Orleans.
- Hambleton, R. K. (1994). Item response theory: a broad psychometric framework for measurement advances. *Psicothema*, 6(3), 535-556.
- Hambleton, R.K. ve Swaminathan, H. (1989). *Item Response Theory. Principles and Applications*. Kluwer-Nijhoff Publishing. Boston-USA.
- Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 32, 179-185.

- Horst, P. (1966). *Psychological measurement and prediction*. California: Wadsworth Publishing.
- Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology*, 35(1), 71–89.
- Jaradat, D., & Tollefson, N. (1988). The impact of alternative scoring procedures for multiple-choice items on test reliability, validity, and grading. *Educational and psychological Measurement*, 48(3), 627-635.
- Jöreskog, K.G.& Sörbom, D. (1993). *LISREL 8: Structural equation modeling with SIMPLIS command language*. Chicago: Scientific Software International, Inc.
- Kan, A. (2005). Yazılı yoklamaların puanlamasında puanlama cetveli ve yanıt anahtarı kullanımının (aynı) puanlayıcı güvenilirliğine etkisi. *Eurasian Journal of Educational Research*, (20), 166-177.
- Kan, A., & Kayapınar, U. (2006). Yabancı dil eğitiminde aynı davranışları yoklayan çoktan seçmeli ve kısa cevaplı iki testin madde ve test özelliklerinin karşılaştırılması. *Eğitim ve Bilim*, 31(142).
- Karasar, N.(2016). *Bilimsel Araştırma Yöntemi*. Ankara: Nobel Akademik Yayıncılık.
- Karataş, F. Ö., Köse, A. G. S., & Coştu, A. G. B. (2003). Öğrenci yanılgılarını ve anlama düzeylerini belirlemede kullanılan iki aşamalı testler. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 13(13), 54-69.
- Kenny, D.A. (1987). *Statistics for the social and behavioral science*. Little, Brown.
- Kılıç, A. F., & Doğan, N. (2019). The effect of item weighting on reliability and validity. *Journal of Measurement and Evaluation in Education and Psychology*, 10(2), 149-164. DOI: 10.21031/epod.516057
- Kinsey, T. L. (2003), A Comparison of IRT and Rasch Procedures in a Mixed-Item Format Test. University of North Texas. Doctoral Dissertation.
- Kline, R.B. (2011). *Principles and practice of structural equation modeling* (3rd Edition). New York: The Guilford Press.
- Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice*. Hoboken, NJ: John Wiley & Sons, Inc.
- Kurz, T.B. (1999). A review of scoring algorithms for multiple-choice test items. *EDRS Publications*, Report No: ED 428 076.
- Lancaster, D. 1987, November). A Comparison of Item Type and Source on Difficulty and Discrimination Ability. Paper presented at the annual meeting of the Mid-South Educational Research Association.

- Lau, P. N. K., Lau, S. H., Hong, K. S. and Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology and Society*, 14, 99–110.
- Leech, N. L., Barrett, K. C., & Morgan G. A. (2005). *SPSS for intermediate statistics: Use and interpretation*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika*, 48, 377-391.
- Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1), 172.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, 1992(1), i-30.
- Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132.
- Nitko, A. (2004). *Educational assessments of students*. Englewood Cliffs, NJ: Prentice Hall.
- Özdemir, D. (1997). Kısa cevaplı, çoktan seçmeli ve kompozisyon tipi maddelerden oluşan testlerin psikometrik özelliklerinin karşılaştırmalı olarak incelenmesi. (Yayımlanmamış Yüksek Lisans Tezi). Hacettepe Üniversitesi, Ankara, Türkiye.
- Özdemir, D. (2002). Çoktan seçmeli testlerin klasik test teorisi ve örtük özellikler teorisine göre hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlanması yönünden karşılaştırılması (Doktora Tezi). Hacettepe Üniversitesi, Ankara.
- Özdemir, D. (2003). Çoktan seçmeli testlerde iki kategorili ve önsel ağırlıklı puanlamanın diferansiyel madde fonksiyonuna etkisi ile ilgili bir araştırma. *Eğitim ve Bilim*, 28(129), 37–43.
- Özdoğan, D., & Doğan, N. (2018). Öz değerlendirmeye dayalı şans başarısı düzeltmenin testin psikometrik özelliklerine etkisi. *Pegem Eğitim ve Öğretim Dergisi*, 8(3), 567-598.
- Özgüven, İ. E. (1998). *Psikolojik testler*. Ankara: PDREM.
- Phillips, A. J. (1943). Further evidence regarding weighted versus unweighted scoring of examinations. *Educational and Psychological Measurement*, 3, 151-155.

- Pinckard, R. N., McMahan, C. A., Prihoda, T. J., Littlefield, J. H., & Jones, A. C. (2009). Short-answer questions and formula scoring separately enhance dental students academic performance. *Journal of Dental Education*, 73, 950-961.
- Quist, J. S. (2004). Differential item weighting: improving assessment validity and reducing the impact of applicant faking. (Doctoral dissertation). University of Utah, Florida.
- R Core Team. (2022). R: A language and environment for statistical computing.
- R e, C., Damsg rd, E., Fors, T., & Anke, A. (2014). Psychometric properties of the pain stages of change questionnaire as evaluated by Rasch analysis in patients with chronic musculoskeletal pain. *BMC Musculoskeletal Disord*, 15(1), 95.
- Salehi, M., Sadighi, F. and Bagheri, M. S. (2015). Comparing confidence-based and conventional scoring methods: The case of an English grammar class. *Journal of Teaching Language Skills*, 6(4), 123-152.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.
- Sax, G. (1997). Principles of educational and psychological measurement and evaluation. Wadsworth Publishing Company.
- Saygı, B. (2004). “1-0” ve ağırlıklı puanlama yöntemleri ile puanlanan çoktan seçmeli testlerin madde ve test özelliklerinin karşılaştırılması. (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.
- Selvi, H. (2013). Klasik test ve madde tepki kuramlarına dayalı değişen madde fonksiyonu belirleme tekniklerinin farklı puanlama durumlarında incelenmesi. (Doktora Tezi). Mersin Üniversitesi, Mersin.
- Selvi, H. ve Derici Yıldırım, D. (2020). Örtük sınıf analizinin farklı puanlama durumlarında incelenmesi. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fak ltesi Dergisi*, 20 (2), 754-766.
- Si, C.-F., & Schumacker, R. E. (1983). Ability estimation under different item parameterization and scoring methods. *International Journal of Testing*, 4, 137-181.
- So an, G. (2015). Empirical option weights for multiple-choice items: Interactions with item properties and testing design. *Metodoloski Zvezki*, 12, 25–43.
- Stanley, J. C., & Wang, M. D. (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, 30(1), 21-35.
- Sympson, J. B. (1993). A procedure for linear polychotomous scoring of test items. Navy Personnel Research And Development Center San Diego Ca.

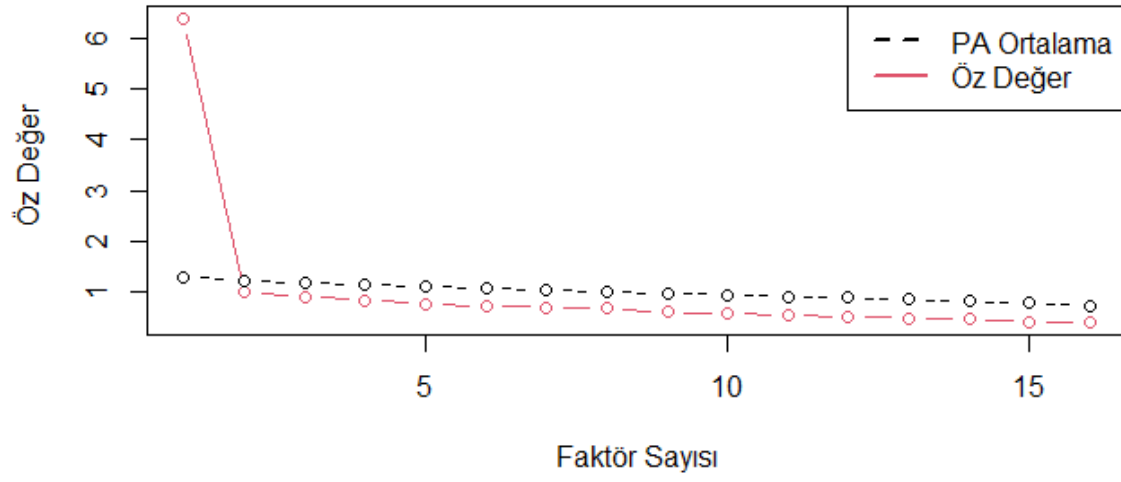
- Tabachnick, B. G. & Fidell, L. S. (2013). *“Using Multivariate Statistics”* Boston: Pearson Education.
- Tan, Ş. & Erdoğan, A. (2004). *Öğretimi planlama ve değerlendirme*. Ankara: PEGEMA.
- Tekin, H. (2000). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı yayınları.
- Tennant, A., McKenna, S. P. and Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health*, 7(1), 22-26.
- Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Education Measurement*, 14, 201-214.
- Thorndike, R. L. & Hagen, E. P. (1961). *Measurement and evaluation in psychology and education*. New York: John Wiley & Sons.
- Tunç, E. B. (2016). İki ve çok kategorili puanlanan maddelerde değişen madde fonksiyonlarının karşılaştırılması. (Doktora Tezi). Ankara Üniversitesi, Ankara.
- Turgut, M. F. (1997). *Eğitimde ölçme ve değerlendirme metotları* (10. Baskı). Ankara: Gül.
- Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem.
- Umay, A. (1997). Yanıtlayıcı davranışlarının analizi yolu ile matematikte problem çözümleri için bir güvenilirlik ve geçerlik araştırması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*. 13, 47-56.
- Umay, A. (1998). Seçmeli derslerde yanıtlayıcı davranışları ve şans başarısının elimine edilmesi işlemlerine ilişkin bazı öneriler. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 14(14), 54-61.
- Wang, M. W., & Stanley, J. C. (1970). Differential Weighting: A Review of Methods and Empirical Studies. *Review of Educational Research*, 40(5), 663–705.
- Warrens, M. J., de Gruijter, D. N. M., & Heiser, W. J. (2007). A systematic comparison between classical optimal scaling and the two-parameter IRT model. *Applied Psychological Measurement*, 31, 106-120.
- Waters, B. K. (1976). The Measurement of Partial Knowledge: A Comparison Between Two Empirical Option-Weighting Methods and Rights-Only Scoring. *The Journal of Educational Research*, 69 (7), 256-260.
- Wetzel, E., Böhnke, J.R., Carstensen, C.H., Zeigler, M., Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*, 34(2), 2013, 69-81. Doi: 10.1027/1614-0001/a000102.

- Wongwiwatthanakit, S., Popovich, N. G. and Bennett, D. E. (2000). Assessing pharmacy student knowledge on multiple-choice examinations using partial credit scoring of combined-response multiple-choice items. *American Journal of Pharmaceutical Education*, 64(1), 1-10.
- Yamamoto, K., & Kulick, E. (1992, April). An information-based approach to maintaining content validity and determining the relative value of polytomous and dichotomous items. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Yaman, S. (2016). Çoktan seçmeli madde tipleri ve fen eğitiminde kullanılan örnekleri. *Gazi Eğitim Bilimleri Dergisi*, 2(2), 151-170.
- Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item independence. *Journal of Educational Measurement*, 30(3). 187-213.
- Young, John W. (1991). Gender bias in predicting college academic performance: a new approach using Item Response Theory. *Journal of Educational Measurement*. 28(1), 37- 47.
- Yurdugöl, H. (2006). Paralel, eşdeğer ve konjenerik ölçmelerde güvenilirlik katsayılarının karşılaştırılması. *Ankara Üniversitesi Eğitim Fakültesi Dergisi*, 39 (1), 15-37.
- Yurdugöl, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1, 1-8.
- Zhai, X., & Li, M. (2021). Validating a partial-credit scoring approach for multiple-choice science items: an application of fundamental ideas in science. *International Journal of Science Education*, 43(10), 1640–1666.

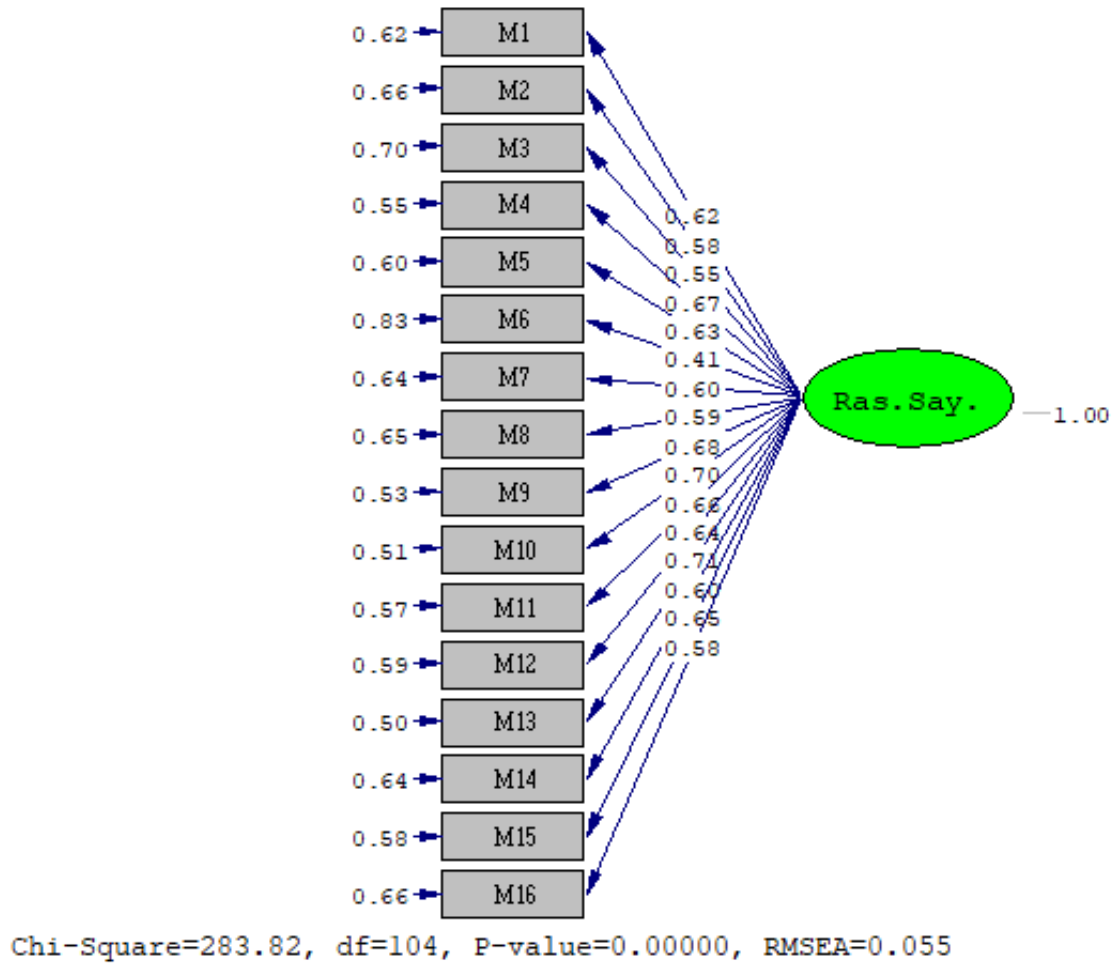


EKLER

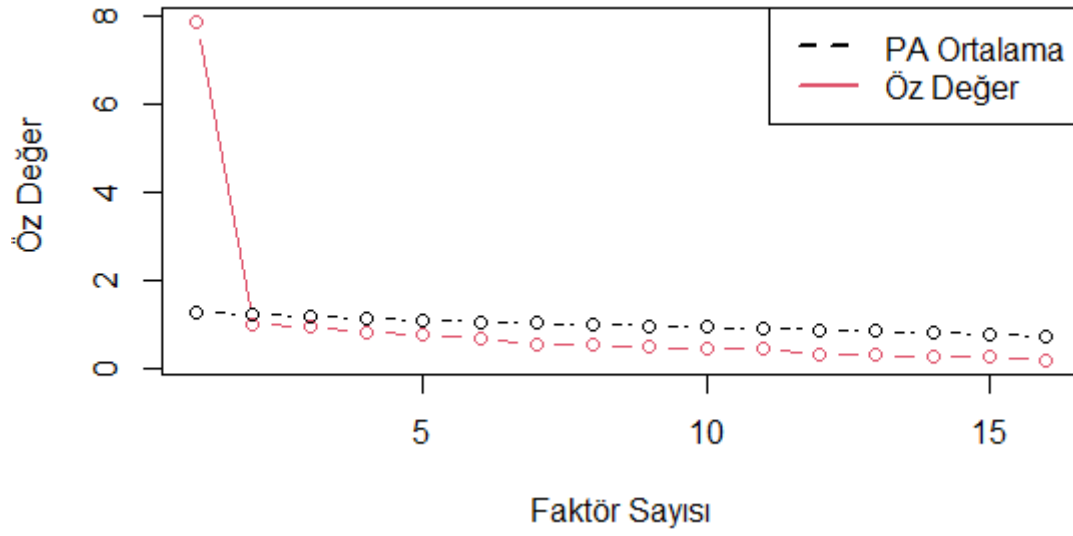
EK 1. OCAP Yönteminin Kullanıldığı Duruma İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları



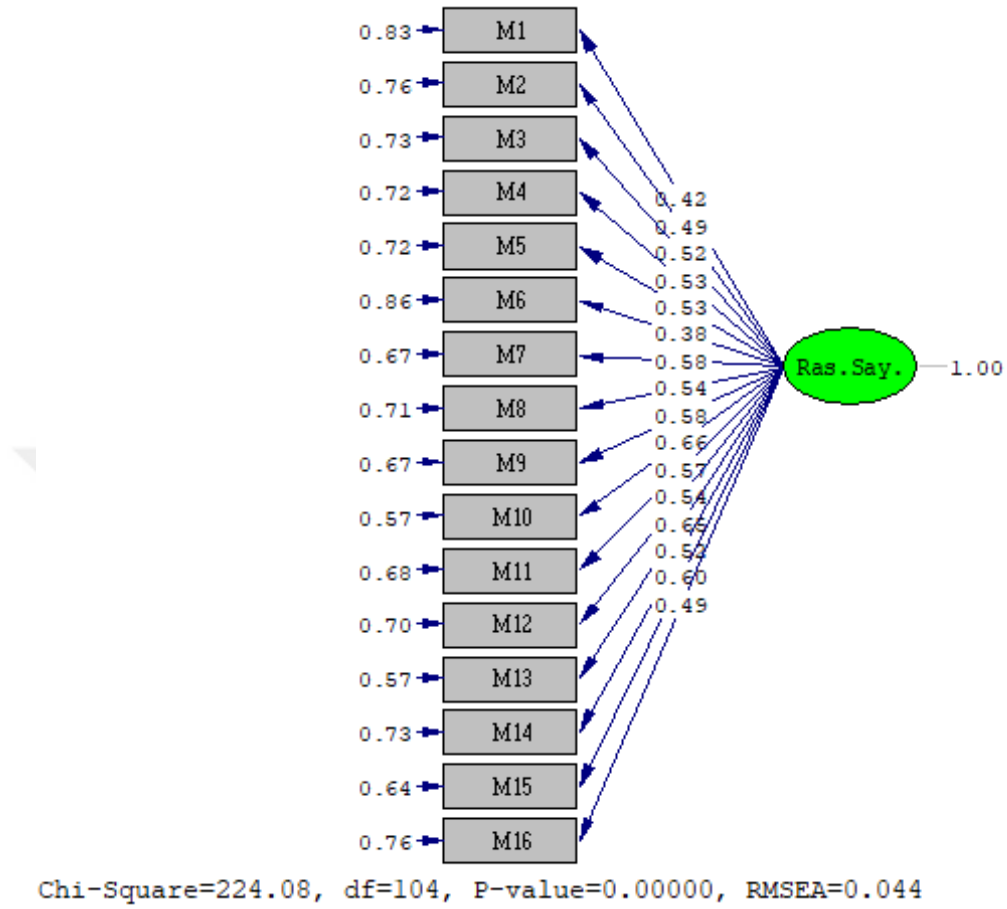
EK 2. OCAP Yönteminin Kullanıldığı Durumda Ölçme Aracının Tek Faktörlü Modeli için Yol Şeması



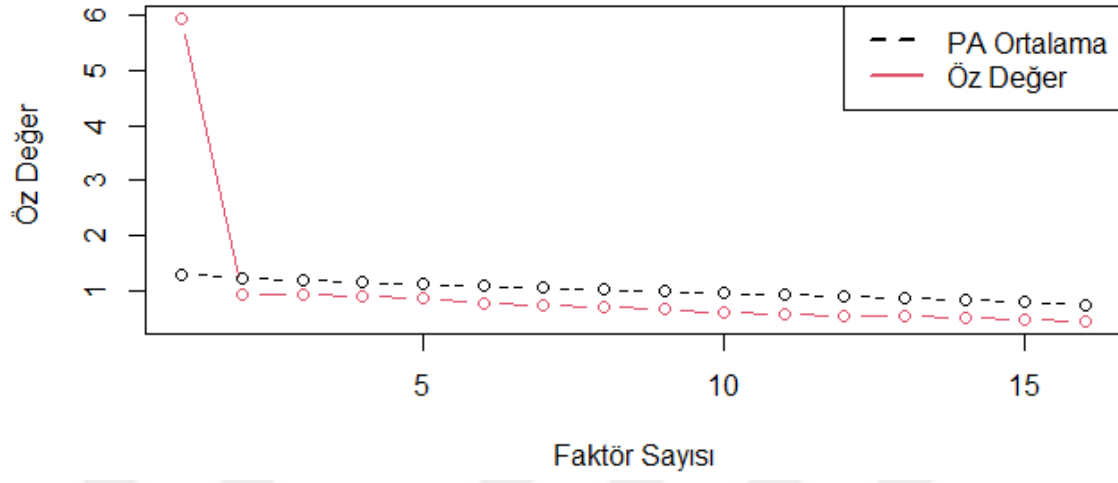
EK 3. MGİAP Yönteminin Kullanıldığı Duruma İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları



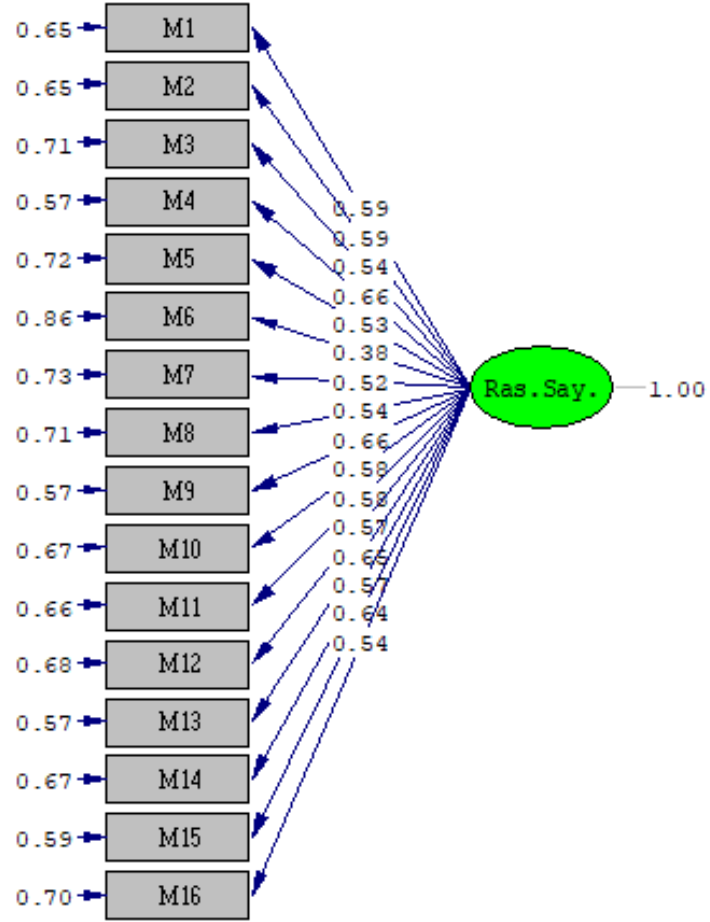
EK 4. MGİAP Yönteminin Kullanıldığı Durumda Ölçme Aracının Tek Faktörlü Modeli İçin Yol Şeması



EK 5. ÇAP Yönteminin Kullanıldığı Duruma İlişkin Özdeğer Grafiği ve Paralel Analiz Sonuçları



EK 6. ÇAP Yönteminin Kullanıldığı Durumda Ölçme Aracının Tek Faktörlü Modeli İçin Yol Şeması



Chi-Square=214.91, df=104, P-value=0.00000, RMSEA=0.043

EK 7. Etik Kurul Onayı

ANKARA ÜNİVERSİTESİ SOSYAL BİLİMLER ALT ETİK KURULU KARAR ÖRNEĞİ

Karar Tarihi : 03/01/2022

Toplantı Sayısı : 1

Karar Sayısı : 3

3-Üniversitemiz Eğitim Bilimleri Enstitüsü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı doktora öğrencilerinden **Emel Ilıcan**'ın "Kısa Yanıtlı Testlerde Maddelerin Ağırlıklı Puanlanmasının Test Puanlarının Güvenirliğine ve Geçerliğine Etkisi" başlıklı tezi ile ilgili "İnsan Üzerinde Yapılan Klinik Dışı Araştırmalar Başvuru Formu" Etik Kurulumuzca incelendi.

Üniversitemiz Eğitim Bilimleri Enstitüsü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı doktora öğrencilerinden **Emel Ilıcan**'ın "Kısa Yanıtlı Testlerde Maddelerin Ağırlıklı Puanlanmasının Test Puanlarının Güvenirliğine ve Geçerliğine Etkisi" başlıklı tezinin araştırma protokolüne uyulması ve etik onay tarihinden itibaren geçerli olması koşuluyla uygulanmasının etik açıdan uygun olduğuna oybirliği ile karar verildi.

ASLININ AYNIDIR
03/01/2022

EK 8. Millî Eğitim Bakanlığı Araştırma İzni



T.C.
ANKARA VALİLİĞİ
Millî Eğitim Müdürlüğü

Sayı : E-14588481-605.99-62891310
Konu : Araştırma İzni

07.11.2022

ANKARA ÜNİVERSİTESİ REKTÖRLÜĞÜNE
(Eğitim Bilimleri Enstitüsü Müdürlüğü)

İlgi: a) MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğünün 2020/2 sayılı Genelgesi.
b) 01.11.2022 tarihli ve 703341 sayılı yazınız.

Üniversiteniz Eğitim Bilimleri Anabilim Dalı Doktora Programı öğrencisi Emel ILICAN'ın "Kısa Yanıtlı Testlerde Maddelerin Ağırlıklı Puanlanmasının Test Puanlarının Güvenirliliğine ve Geçerliliğine Etkisi" konulu çalışması kapsamında merkez ilçelere bağlı resmi ortaokullarda uygulama yapma talebi ilgi (a) Genelge çerçevesinde incelenmiştir.

Yapılan inceleme sonucunda, söz konusu araştırmanın Müdürlüğümüzde muhafaza edilen ölçme araçlarının; Türkiye Cumhuriyeti Anayasası, Millî Eğitim Temel Kanunu ile Türk Millî Eğitiminin genel amaçlarına uygun olarak, ilgili yasal düzenlemelerde belirtilen ilke, esas ve amaçlara aykırılık teşkil etmeyecek, eğitim-öğretim faaliyetlerini aksatmayacak şekilde okul ve kurum yöneticilerinin sorumluluğunda, gönüllülük esasına göre uygulanması Müdürlüğümüzce uygun görülmüştür.

Bilgilerinizi ve gereğini rica ederim.

Harun FATSA
Vali a.
Millî Eğitim Müdürü

Ek: Uygulama Araçları (5 Sayfa)

Dağıtım:
Gereği:
Ankara Üniversitesi

Bilgi:
B Planı

Bu belge güvenli elektronik imza ile imzalanmıştır.

Adres :

Belge Doğrulama Adresi : <https://www.turkiye.gov.tr/meb-ehys>

Telefon No :

Bilgi için:

E-Posta:

Unvan : Memur

Kep Adresi : mebi@n01.kep.tr

İnternet Adresi: Faks:

Bu evrak güvenli elektronik imza ile imzalanmıştır. <https://evrak.meb.gov.tr> adresinden 1506-b0ca-3bdf-9aa6-5c99 koda ile teyit edilebilir.

BENZERLİK BİLDİRİMİ

“Kısa Yanıtlı Maddelerde Farklı Ağırlıklı Puanlama Yöntemlerinin Testin Psikometrik Özellikleri Açısından İncelenmesi” başlıklı tezimin ana bölümü (ön bölüm, kaynaklar ve ekler hariç) Turnitin İntihal Tespit Programı aracılığıyla incelenmiş ve ilgili rapor danışmanım tarafından da kontrol edilmiştir. Kontrol sırasında (1) “Yedi sözcükten daha az olan benzeşmeler” (2) “Kaynaklar” (3) “Doğrudan alıntılar” ve (4) “Lisansüstü tezim ile ilgili yapmış olduğum kendi yayınlarım ve yararlandığım mevzuat metinleri gibi kaynaklar” dışarıda tutulmuştur. Benzerlik kontrolüne ilişkin rapordan elde edilen sonuçlar aşağıda sunulmuştur.

Rapor Tarihi	: 08/06/2024
Gönderim Numarası	: 2398334849
Sayfa Sayısı	: 126
Sözcük Sayısı	: 25557
Karakter Sayısı	: 187381
Benzerlik Oranı	: % 9
Savunma Tarihi	: 30/04/2024

Yukarıda belirtilen sonuçları gösteren Turnitin İntihal Tespit Programı’na ilişkin orijinal raporu, sonuçlarda herhangi bir değişiklik yapmaksızın bu beyanım ekinde Enstitüye teslim ettiğimi, tezimin %10’dan fazla benzerlik oranı içerdiğinin, tek bir kaynakla eşleşme oranının ise %2’den fazla olduğunun belirlenmesi durumunda, bundan doğabilecek tüm yasal sorumluluğu kabul ettiğimi bildirir, saygılarımı sunarım.

Öğrencinin Adı Soyadı: Emel Ilıcan

Tarih: 08.06.2024

İmza:

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı ve Soyadı : Emel Ilıcan

E-Posta Adresi :

İş Deneyimi :

Unvan	Görev Yeri	Yıl
Öğretmen	MEB, İbn-i Sina Mesleki ve Teknik Anadolu Lisesi	2012-2014
Uzman Öğretmen	MEB, Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü	2014 -

Akademik Bilgiler

Öğrenim Durumu:

Derece	Bölüm/Program	Üniversite	Yıl
Lisans	Matematik Öğretmenliği	Gazi Üniversitesi	2012
Yüksek Lisans	Eğitimde Ölçme ve Değerlendirme	Gazi Üniversitesi	2017

Yayınlar:

Deniz, K. Z., & Ilıcan, E. (2021). Comparison of G and Phi coefficients estimated in generalizability theory with real cases. *International Journal of Assessment Tools in Education*, 8(3), 583-595.

Ilıcan, E. (2021). Diskriminant Analizi. K. Z. Deniz (Editör), *İstatistikolay 2 Çok Değişkenli İstatistik*. (ss. 71-116). Ankara: Nobel.