AÇIK VE UZAKTAN ÖĞRENMEDE ÇOK AŞAMALI ÖLÇMENİN ETKİLİLİĞİNİN İNCELENMESİ: BİR SİMÜLASYON ÇALIŞMASI

Doktora Tezi

Gülgün BULUT

Eskişehir 2023

AÇIK VE UZAKTAN ÖĞRENMEDE ÇOK AŞAMALI ÖLÇMENİN ETKİLİLİĞİNİN İNCELENMESİ: BİR SİMÜLASYON ÇALIŞMASI

Gülgün BULUT

DOKTORA TEZİ

Uzaktan Eğitim Anabilim Dalı

Danışman: Doç. Dr. Murat AKYILDIZ

Eskişehir

Anadolu Üniversitesi

Sosyal Bilimler Enstitüsü

Haziran 2023

JÜRİ VE ENSTİTÜ ONAYI

Gülgün BULUT'ın ''Açık ve Uzaktan Öğrenmede Çok Aşamalı Ölçmenin Etkililiğinin İncelenmesi: Bir Simülasyon Çalışması'' başlıklı tezi **13 Haziran 2023** tarihinde, aşağıdaki jüri tarafından Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin 37. Maddesi uyarınca ilgili maddeleri uyarınca **Uzaktan Eğitim Anabilim Dalında, Doktora** tezi olarak değerlendirilerek kabul edilmiştir.

		İmza
Üye (Tez Danışmanı)	**: Doç. Dr. Murat AKYILDIZ**	.....
Üye	**: Prof. Dr. Evrim GENÇ KUMTEPE**	.....
Üye	**: Prof. Dr. Eralp Hüseyin ALTUN**	.....
Üye	**: Doç. Dr. Tarkan GÜRBÜZ**	.....
Üye	**: Dr. Öğr. Üy. Nejdet KARADAĞ**	.....

Prof. Dr. Saime ÖNCE
Anadolu Üniversitesi
Sosyal Bilimler Enstitüsü Müdürü

ÖZET

AÇIK VE UZAKTAN ÖĞRENMEDE ÇOK AŞAMALI ÖLÇMENİN ETKİLİLİĞİNİN İNCELENMESİ: BİR SİMÜLASYON ÇALIŞMASI

Gülgün BULUT

Uzaktan Eğitim Anabilim Dalı
Anadolu Üniversitesi, Sosyal Bilimler Enstitüsü, Haziran 2023

Danışman: Doç. Dr. Murat AKYILDIZ

Ölçme ve değerlendirmenin özünde bireyin zihninde gizil (örtük) durumda bulunan bilgi birikimini en az hata ile gerçeğe en yakın doğrulukta kestirebilmek yatmaktadır. Geçmişten günümüze bu konu ile ilgili pek çok yöntem uygulanmış olmakla birlikte en yaygın kullanılan ölçme aracı sınavla ölçme ve değerlendirme yöntemidir. Bu sınavların özellikle geniş kitlelere uygulanma biçimi çoktan seçmeli testler şeklindedir. Araştırma kapsamında MST örüntüsü test üretme yönteminin açıköğretim sınavlarına uygulanması durumunda elde edilen olası sonuçların simülatif bir ortamda farklı yöntem ve tekniklerle incelenerek Açık ve Uzaktan Öğrenme (AUÖ) sistemi için optimum algoritmanın keşfedilmesi amaçlanmıştır. Araştırma sürecinde simülasyona dayalı olarak üretilen verilerin KTK, IRT ve MST (MST-R ve MST-S) yöntemlerinin her biri için ayrı ayrı standart hata değerleri, korelasyon katsayısı, AIC değerleri ve farklı kuramlara göre puan sıralaması farkları bağlamında analizleri gerçekleştirilmiştir. Analiz sonuçlarına göre KTK verilerle en az uyumlu yöntem olarak tespit edilirken IRT 3PL yöntemi en az standart hata değerine sahip optimum yöntem olarak tespit edilmiştir. MST test sunum yöntemi ise KTK'dan önemli oranlarda uzaklaşırken IRT 3PL yöntemine oldukça yakın sonuçlar vermektedir. Araştırma sonuçlarından elde edilen bulgular, MST yönteminin açık ve uzaktan öğrenme sistemleri dahil olmak üzere geniş kitlelere çoktan seçmeli testler ile sınav uygulayan herhangi bir sistemin ölçüm kesinliğini artırabileceğini göstermektedir.

Anahtar Sözcükler: Çok aşamalı testler, Madde tepki kuramı, Klasik test kuramı, Açık ve uzaktan öğrenme

ABSTRACT

INVESTIGATION OF THE EFFICIENCY OF MULTI-STAGE TESTING IN OPEN AND DISTANCE LEARNING: A SIMULATION STUDY

Gülgün BULUT

Department of Distance Education

Anadolu University, Graduate School of Social Sciences, Haziran 2023

Supervisor: Assoc. Prof. Dr. Murat AKYILDIZ

The main objective of measurement and assessment is to estimate the latent ability with a small amount of error and great precision. Although many methods have been applied for this purpose from the past to the present, the most widely used measurement tool is testing. These assessments are administered, especially to a high number of students, using multiple-choice question formats. Within the scope of the research, it is aimed to discover the optimum algorithm for the Open and Distance Learning (ODL) system by examining the possible results obtained in the case of applying the test generation method of the Multistage to the open education exams in a simulative environment in different conditions. In the research process, the simulation-based data were analyzed and compared by using Classical Test Theory, Item Response Theory, and Multistage testing conditions (MST-R and MST-S) estimations in terms of standard error values, correlation coefficients, AIC values, and score rank differences. Classical Test Theory was found to be the approach that fit the data the least well, and the IRT 3PL method was found to be the best one with the lowest error rate. The Multistage test presentation-based estimations, differs significantly from the Classical Test Theory-based estimations, giving results very close to the IRT 3PL model. The conclusions drawn from the study's findings demonstrate that the Multistage testing method can improve the measurement precision of any system, including open and remote learning systems, that use multiple-choice test techniques on big number of students.

Keywords: Multistage testing, Item response theory, Classical test theory, Open and distance learning

ÖNSÖZ

Teknolojik ilerlemelerin beraberinde getirdiği dijital dönüşüm yaşamın pek çok evresinde olduğu gibi eğitim öğretim süreçlerinde de değişim ve dönüşüme yol açmıştır. Söz konusu değişim dönüşümün katkılarıyla eğitim öğretime dair pek çok işlem adımı çevrimiçi platformlar aracılığıyla yürütülebilmektedir. Bu işlem adımlarının en önemli basamağını temsil eden ölçme ve değerlendirme aşamasının çevrimiçi platformlarda yürütülmesi konusu ise son zamanlarda sıklıkla gündeme gelmektedir. Çevrimiçi ölçme ve değerlendirmenin geniş kitlelere uygulanan sınavlarda yaygın olarak kullanılmasında 2019 yılında başlayıp tüm dünyayı saran Covid-19 pandemisinin ve 2023 yılının başlarında ülkemizde büyük yıkımlara sebep olan 6 Şubat depreminin yaşam koşullarına getirmiş olduğu zorunlulukların etkisi büyüktür. Son zamanlarda gerek dünya gerek ülkemiz özelinde salgın ve doğal afetlerin sonuçları en acı salt gerçeklikle tecrübe edilmiştir. Kendi başına mevcut sonuçları ağır olan böylesi durumlarda eğitim öğretimin yürütülmesinde yaşanacak aksamaların mininuma indirilmesi eğitim-öğretim ile ölçme ve değerlendirmenin çevrimiçi ortamlarda yapılabilmesi konusunda hazır olunmasına bağlıdır. Hazırlıksız yakalandığımız bu süreçte özellikle hızlı bir biçimde uygulamaya konulan çevrimiçi sınavlarda güvenlik, maliyet, ölçüm hassasiyeti gibi konular sıklıkla karşılaşılan problemler arasında yer almaktadır. Söz konusu problemlerden yola çıkarak hazırlamış olduğum "Açık ve Uzaktan Öğrenmede Çok Aşamalı Ölçmenin Etkililiğinin İncelenmesi: Bir Simülasyon Çalışması" konulu tez çalışmamın araştırmacılara, karar alıcılara, uygulayıcılara, öğrenenlere, öğretenlere ve alana katkı sağlaması umuduyla keyifli okumalar dilerim.

"The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires."
William Arthur Ward

William Arthur Ward'ın son cümlesinde ifade ettiği gibi öğretmenleri hayatına ilham kaynağı olmuş şanslı öğrencilerdenim. Öğrenmenin sonsuzluğu ve sınırsızlığında öğrenciliğimin tamamlandığı bugünlerde bana ilham olan tüm öğretmelerime saygı ve minnetlerimi sunuyorum.

Doktora tezi gibi zorlu ve kapsamlı bir sürecin her aşamasında keyifle öğrenerek birikimli bir şekilde ilerlememi sağlayan danışman hocam Doç. Dr. Murat AKYILDIZ'a ilk ve en öncelikli teşekkürü borç bilirim. Akademik birikimi, yol göstericiliği, anlayışı ve insan odaklı yaklaşımı ile daha nice öğrencilerin hayatına değer katması dileğiyle…

Tez izleme komitemde katkılarını hiçbir zaman esigemeyen farklı bakış açıları ve önerileri ile motivasyonumu hep diri tutarak bir sonraki izleme toplantısına kadar heyacanla çalışmama vesile olan değerli hocalarım Prof. Dr. Evrim GENÇ KUMTEPE ve Doç. Dr. Tarkan GÜRBÜZ'e sonsuz teşekkürlerimi sunuyorum.

Lisansüstü eğitimim boyunca akademik birikiminden ve pozitif enerjisinden beslendiğim değerli hocam Prof. Dr. Evrim GENÇ KUMTEPE tez komitemde de yer alarak her zaman olduğu gibi değerli katkıları ile bu zorlu yolu benim için kolaylaştırdı ve zenginleştirdi.

Doktora eğitimim sırasında birikiminden yararlanma fırsatı bulduğum için kendimi şanslı addettiğim, bir işi samimiyetle ve severek yapmanın özelliği ve güzelliğinin farkına varmamı sağlayan kıymetli hocam Doç. Dr. Tarkan GÜRBÜZ tez komitemde de yer alarak farklı bakış açısı ve yönlendirmeleri ile eşsiz katkılar sağladı.

Tez savunma jürimdeki katkılarından dolayı Prof. Dr. Eralp ALTUN ve Dr. Öğr. Üyesi Nejdet KARADAĞ hocalarıma teşekkürlerimi sunarım.

Tez yazma sürecimin büyük bölümüne tanıklık eden ve nezaketli naif duruşu ile desteğini esirgemeyen saygıdeğer hocam Prof. Dr. Saime ÖNCE'ye teşekkürlerimi sunuyorum.

Yollarımız kesiştiği için hep şükrettiğim yol arkadaşlarım, candostlarım, kardeşten ayırmadıklarım her birinize sonsuz teşekkürlerimle…

Sözlerimi sonlandırdığım bu satırlarda en anlamlı teşekkürü hayatlarını evlatlarına adayan annem ve babama; onların nezdinde de her daim tek yumruk olabildiğimiz her birini sırtımı yasladığım dağ olarak nitelendirdiğim kardeşlerime etmek istiyorum. Teşekkürlerin en özeli ise elbetteki ailemizin en kıymetlisi en özel kardeşim Müfid'ime gelsin..varlığına bin şükür, sen bizim nefesimizsin…

..../..../20....

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

.

Gülgün BULUT

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

GÖRSELLER DİZİNİ

Sayfa

Sayfa

Sayfa

Sayfa

Sayfa

Sayfa

Sayfa

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

θ	: Theta (Öğrenci Yetenek Parametresi)
1 PL	: 1 Parametreli Lojistik Model
2 PL	: 2 Parametreli Lojistik Model
3 PL	: 3 Parametreli Lojistik Model
4 PL	: 4 Parametreli Lojistik Model
AIC	: Akaike Information Criterion
CAT	: Computer Adaptive Testing (Bilgisayar Uyarlamalı Test)
CSEE	: Conditional Standard Errors of Estimation
CSEM	: Conditional Standard Error of Measurement
CTT	: Classical Test Theory
ETS	: Educational Testing Service
GMAT	: Graduate Management Admission Test
GRE	: Graduate Record Examination
ICC	: Item Characteristic Curve (Madde Karakteristik Eğrisi)
IIF	: Item Information Functions (Madde Bilgi Fonksiyonu)
IRT	: Item Response Theory (Madde Tepki Kuramı)
KTK	: Klasik Test Kuramı
MST	: Multistage Testing
MST-R	: Multistage Testing by Routing
MST-S	: Multistage Testing by Shaping
MTK	: Madde Tepki Kuramı (Item Response Theory)
NEAP	: National Assessment of Educational Progress
P&P	: Paper and Pencil

PISA : Programme for International Student Assessment

SE : Standard Error

SEM : Standard Error of Measurement

TDK : Türk Dil Kurumu

TIF : The Test Information Function (Test Bilgi Fonksiyonu)

TSM : True Score Model

1. GİRİŞ

Ölçme ve değerlendirmenin özünde bireyin zihninde gizil (örtük) durumda bulunan bilgi birikimini en az hata ile gerçeğe en yakın doğrulukta kestirebilmek yatmaktadır. Ölçme ve değerlendirme; eğitim öğretim sürecinin nihai çıktısının görülebilmesi ve sistemin gelecekteki gereksinimleri hakkında fikir sahibi olunabilmesi için önemli veriler sağlayan bir işlem adımıdır. Bu sebeple en doğru kestirimi yapabilmek adına ölçme ve değerlendirme işleminin en az standart hata değeriyle gerçekleştirilmesi sürece dahil olan tüm paydaşlar tarafından temel amaç olarak hedeflenmektedir. Ölçme ve değerlendirme alanında bu amaca yönelik hazırlanmış çok sayıda ölçme aracı bulunmakla birlikte geçmişten günümüze en yayın kullanılan ölçme aracı sınavla ölçme ve değerlendirme yöntemidir. Sınavla ölçme ve değerlendirme yöntemleri arasında ise özellikle geniş kitlelerin katılımıyla gerçekleştirilen sınavlarda sıklıkla çoktan seçmeli soru tekniğinin uygulandığı testler ölçme aracı olarak kullanılmaktadır (Adom, Mensah ve Dake, 2020). Testler özellikle eğitim uygulamalarında bir ölçme aracı olmanın ötesinde bir dizi soru ile bir davranış örneğini sistematik olarak ölçmenin yanı sıra bireyin yeteneğini, becerisini ve bilgisini kabul edilebilir veya kabul edilemez olarak belirli bir standarda göre ölçmek için tasarlanmış yapılardır (Miller, Linn ve Gronlund, 2009). Dolayısıyla testler, eğitim sistemlerine bireyin yeteneği, becerisi ve bilgisi hakkında karar verebilmeleri için "puan" olarak nitelendirilen veriler sunmaktadır. Bu puanlar bireyin beklenen yeteneğe sahip olup olmadığına dair verilecek karar/kararları etkileyeceği için mümkün olan en yüksek doğruluk oranında tespit edilmiş olması tüm paydaşlar açısından önem arz etmektedir.

Ölçme ve değerlendirme alanında gerçekleştirilen araştırma ve süreçlerin çıkış noktası; "Bireyin puanını gerçeğe en yakın doğrulukta nasıl kestirebiliriz?" ve "Bireyin puanını en az hatayla nasıl ölçebiliriz?" soruları üzerine şekillenmektedir. Benzer şekilde çoktan seçmeli bir testin ölçme aracı olarak uygulandığı bir sınavın özünde de bu gerekçe yatmaktadır. Bilindiği üzere uygulamada öğrenen başarısının kestiriminde gerçek değere ulaşma konusunda ölçülen özellikten, ölçme aracından, ölçme işlemini gerçekleştiren kişi ya da kurumlardan kaynaklı bir takım ölçüm güçlükleri ve dolayısıyla beraberinde getirdikleri hatalarla karşı karşıya kalınmaktadır. Bu ölçüm güçlüklerinin yol açtığı ölçme hatalarını minimize etme görevi ise ölçme işlemini gerçekleştiren kişi ya da kurumların sorumluluğundadır (Doğan, 2019; Howel ve Hricko, 2005; Nunnally ve Berstein, 1994).

Dolayısıyla ölçme işlemini gerçekleştiren kişi ya da kurumlar tarafından benimsenen puanlama ve test sunum yöntemleri de gerçek değere ulaşmada önemli bir etken konumundadır. Bu anlamda ölçme ve değerlendirme alanında bireyin gerçek puanını kestirebilmek için Klasik Test Kuramı (KTK)'ndan öğrenenin şans başarısının değerlendirmeye dahil edildiği Madde Tepki Kuramı/Item Response Theory (MTK/IRT) 3 PL (parametreli) modele kadar pek çok yöntem kullanılmaktadır. Son zamanlarda ise birey içi ve bireyler arası farkları duyarlı bir şekilde ortaya çıkarmayı amaçlayan MST (Multistage Testing) yöntemi (Han, 2013) sıklıkla gündeme gelen ölçme teknikleri arasında yer almaktadır. MST, her bir sınav katılımcısına yönelik olarak verilecek başarılı-başarısız kararının doğruluğunu en üst düzeye çıkarmak için CAT (Computer Adaptive Testing)'in verimliliğinden ve doğrusal testin test formu montaj kontrollerinden yararlanarak (Zenisky, Hambleton ve Luecht, 2009) bireysel test programlarının özelliklerine, gereksinimlerine ve hedeflerine uygun test sunumu gerçekleştirme çabasında olan bir yöntemdir.

Araştırma kapsamında MST örüntüsünün test üretme yönteminin açıköğretim sınavlarına uygulanması durumunda elde edilecek olası sonuçların simülatif bir ortamda farklı yöntem ve tekniklerle incelenerek Açık ve Uzaktan Öğrenme (AUÖ) sistemi için optimum algoritmanın keşfedilmesi planlanmaktadır. Söz konusu süreç dahilinde uygun simülasyon yöntemleriyle MST koşullarında üretilen veriler ile bir simülasyon çalışması gerçekleştirilmesi hedeflenmektedir.

1.1. Problem

Ölçme ve değerlendirme alanı geçmişten günümüze eğitimin tüm kademelerinde "Öğrenen becerisi en az hata ile gerçeğe en yakın doğrulukta nasıl ölçülebilir?" sorusuna yanıt aramaktadır. Bu durum gerek sürecin karmaşıklığı gerekse bazı yöntemlerin uygulanmasındaki güçlükler sebebiyle eğitim öğretim sisteminin en temel problemleri arasında yer almaktadır. Özellikle 2019 yılında başlayıp tüm dünyayı saran, bu çalışmanın yapıldığı tarihte de kısmen devam eden covid-19 pandemisi sebebiyle sınavların çevrimiçi ortamlarda maksimum güvenlikle ve gerçeğe yakın sonuç verecek şekilde nasıl gerçekleştirilebileceği sorusu sıklıkla gündeme gelmektedir.

Pandemi döneminde sınavlar çevrimiçi ortamlara taşınmış ancak sınav yöntemi/test sunum yöntemi olarak tüm soruların herkese açık olduğu geleneksel yüzyüze sınav kültürü devam ettirilmiştir. Bu durum özellikle eş zamanlı çevrimiçi sınavlarda haberleşme aygıtlarıyla soruların tümünün afişe olmasına yol açarak sınav güvenliği ve ölçüm kalitesi açısından önemli bir tehdit unsuru oluşturmaktadır. Çevrimiçi sınavlarda tüm soruların herkese aynı anda sunulması uzman işgücü ve yüksek maliyetlerle hazırlanmış olan soruların tek seferde kullanılmaz hale gelmesine yol açacaktır. Ayrıca ölçme ve değerlendirme sürecinin amacına ulaşmasına engel olmasının yanı sıra birçok yönden maliyetini (emek, zaman, mali vb.) de artırmış olacaktır. Bu sebeple sınavlarda öncelikle kopya vb. durumlara karşı güvenlik sağlanırken aynı zamanda gerçek değere yakın doğrulukta ölçümlerin ne şekilde yapılabileceği konusu süreçlerinde işlem basamağı olarak ölçme ve değerlendirmeyi barındıran tüm yapıların önemle üzerinde durduğu konular arasında yer almaktadır. Araştırmanın sorunsalı olarak büyük kitlelere eğitim öğretim hizmeti veren açık ve uzaktan öğrenme sistemlerinin sınavlarında MST yönteminin uygulanması durumunda olası sonuçlar simülatif ortamda incelenerek optimum algoritmanın tespiti için aşağıda yer alan araştırma soruları bağlamında yanıt aranmaya çalışılacaktır.

Araştırma Soruları:

Araştırmanın ana sorusu şöyledir: MST (Multistage Testing) test sunum yöntemi kullanılarak elde edilen yetenek kestirimleri (puanlar) ile aynı verilerden elde edilen KTK (Klasik Test Kuramı) ve MTK/IRT (Madde Tepki Kuramı/Item Response Theory)'ya dayalı yetenek kestirimleri arasında simülatif bir ortamda MST lehine anlamlı bir farklılık var mıdır?

Bu ana soru bağlamında aşağıdaki alt sorulara da yanıt aranmıştır:

Farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli);

- MST yöntemi ile elde edilen puanların standart hataları geleneksel KTK yöntemi ile elde edilen puanların standart hatalarından farklılık göstermekte midir?

- MST yöntemi ile elde edilen puanların standart hataları IRT yöntemi ile elde edilen puanların standart hatalarından farklılık göstermekte midir?

- MST yöntemi ile elde edilen puanların standart hataları MST-R (Multistage Testing by Routing) ve MST-S (Multistage Testing by Shaping) yöntemi ile elde edilme durumuna göre farklılık göstermekte midir?

- KTK, IRT ve MST yöntemlerinden hangisi en az standart hata ile gerçeğe en yakın kestirimi yapabilmektedir?

- KTK, IRT ve MST yöntemleri ile elde edilen puanlarla yapılan sıralamalar farklılık göstermekte midir?

1.2. Amaç

Bu çalışmanın amacı MST test üretme yöntemi temel alınarak tasarlanan simülatif bir ortamda bireyin yetenek kestiriminin gerçeğe en yakın doğrulukta tahmin edilebilirliğini farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) ve farklı yöntemlerle ele alarak MST yöntemi ile test üretmenin büyük sistemlerde geleneksel yöntemden daha düşük hata ile puanlama yapmaya imkân tanıyıp tanımadığını ortaya koymaktır. Araştırma sürecinde simülyasona dayalı olarak üretilecek veriler KTK, IRT ve MST yöntemlerinin her biri için ayrı ayrı standart hatalar hesaplanarak en az hata ile gerçek değere en yakın doğrulukta kestirim sağlayan yöntem araştırma soruları çerçevesinde tespit edilmeye çalışılmıştır. Simülasyonlar, pek çok alanda olduğu gibi ölçme ve değerlendirme sürecinde de geniş kitlelere hizmet veren kurumlara sınavlarda kullanılacak en uygun yöntemi minimum maliyetle görebilme olanağı sunmaktadır. Böylece kurumlar bu türden uygulamaları yüksek maliyetlerle doğrudan sürece dahil etmeden önce simülasyonlar aracılığıyla test ederek sınavlarında kullanılabilecek en uygun yöntemi minimum maliyetle belirleme olanağı elde etmiş olacaklardır. Bu çalışmada oluşturulacak simülatif durumlar Açık ve Uzaktan Öğrenme (AUÖ) sistemlerinden sıklıkla gözlenen durumlarla birebir örtüşme gösterecektir. Dağılımın normal olduğu veya normalden sapma gösterdiği, homojen veya heterojen olduğu, örneklemin büyük ya da küçük olduğu durumlar hemen hemen tüm uzaktan öğretim sistemlerindeki ölçme süreçlerinde karşılaşılan gerçek durumlardır. Bu sebeple araştırma kapsamında MST yöntemi ile tasarlanmış bir sınav ve bu sınavı alan adayların maddeleri işaretleme biçimleri araştırma soruları doğrultusunda simüle edildiğinde Açık

ve Uzaktan Öğrenme (AUÖ) sisteminde farklı yöntem ve farklı örneklem büyüklüklerinde uygulanabilir olup olmadığının tespiti araştırmanın temel amacı olarak belirlenmiştir.

1.3. Önem

Günümüz koşullarında hızla gelişen teknoloji ölçme ve değerlendirme alanının da niteliğini ve kapsamını zenginleştirerek dijital araçların bu alanda kullanımını yaygınlaştırmaktadır. Bu nedenle ölçme ve değerlendirme işleminin hangi yöntem aracılığı ile gerçekleştirileceği hususu öğrenme ortamlarının oluşumunda yaşamsal faktör konumunda bulunmaktadır. Son zamanlarda eğitimsel ve psikolojik test sonuçlarıyla ilgili risklerin artması, ölçüm hatalarının rolü ve yanlış sınıflandırmalar gibi konulara daha fazla önem verilmesi hususunu da beraberinde getirmektedir (Zenisky, Hambleton ve Luecht, 2009).

Tez araştırması olarak tasarlanan simülasyon çalışması KTK, IRT ve MST yöntemlerinin her birini farklı örneklem büyüklüklerinde ayrı ayrı ele alarak gerçek değere en yakın doğrulukta kestirim yapabilen yöntemi tespit etmeye odaklanması sebebiyle önem arz etmektedir. Açık ve uzaktan öğrenme sistemlerinde yaygın olarak KTK'ya dayalı puanlama stratejileri kullanılmaktadır. Bu stratejiler uzun süredir eleştiri altındadır (Arnold, 1996; Kline, 2005; Magno, 2009; Rusch, Lowry, Mair ve Treiblmaier, 2017). Klasik Test Kuramı'nın ölçülen özellik ile bireylerin performansları arasında doğrusal ilişki varsayması, elde edilen parametrelerinin örnekleme bağımlı olması, standart hatayı ve güvenilirliği testi alan tüm bireyler için eşit kabul etmesi önemli kısıtlılıklardır (Crocker ve Algina, 2008; Weis ve Yoes, 1990). Açık ve uzaktan öğrenmede birey materyal etkileşiminin barındırdığı transaksiyonel uzaklığa KTK'dan elde edilen puanların taşıdığı bu kısıtlılıklar eklendiğinde sorunun büyüklüğü daha net görünmektedir. IRT (Item Response Theory/Madde Tepki Kuramı) sıklıkla bilgisayarda bireye uyarlanmış testlerin geliştirilmesinde kullanılmaktadır. IRT puanlama yöntemini kullanan CAT test sunum yöntemi, KTK'nın sınırılıklarını büyük oranda çözmekle birlikte kendi içinde bazı dezavantajlar barındırmaktadır. Örneğin; CAT sisteminde sınav katılımcılarının sınav sırasında önceki sorulara verdikleri yanıtları gözden geçirmelerine izin verilmemesi (Han ve Guo, 2014; Rotou, 2007, s. 1-2) katılımcılar üzerinde strese sebep olmaktadır. Bilgisayarlı bireye uyarlanabilir test uygulamalarının büyük soru

bankalarına gereksinim duyması CAT sisteminin bir diğer dezavantajıdır. Çünkü IRT'nin KTK'ya göre yüksek oranlarda bilgi sağlayabilmesi yeterli bir örneklem büyüklüğünü gerektirmektedir (Cappelleri, 2014). Bilgisayarda bireye uyarlanabilir testlerin en önemli dezavantajlarından biri de maliyetli olmasıdır (Glas ve van der Linden, 2003; Yan, Lewis ve von Davier, 2014). CAT sisteminin madde bazlı yapısı ölçüm sırasında diğer yöntemlere oranla daha fazla miktarda bilgi sağlanmasına olanak tanımaktadır. Ancak fazla miktarda bilgi sağlanabilmesi nitelikli soru ile mümkün olmaktadır. Teknik nitelikleri yüksek soru yazmak ise emek, uzmanlık, maliyet ve zaman gerektirmektedir. CAT test sunum yönteminin bu dezavantajlarına bir çözüm olarak MST, IRT'den daha düşük maliyetle KTK'nın sınırlılıklarını da aşarak kullanılabilme potansiyeli vaad etmektedir. MST modül temelli yapısı gereği sınava giren bir kişiye soruların sadece bir kısmını göstermekte ve soruların tamamı tek seferde sunulmamaktadır. Ayrıca MST'nin modül bazlı soru sunma biçimi IRT (madde bazlı soru sunma biçimi kullanılmaktadır)'ye göre maliyet bakımından avantaj sağlarken soruların tamamını aynı anda herkese açıklamadan sınavı gerçekleştirebilmesi KTK'nın sınırlılıkları açısından önemli bir avantaj sağlamaktadır. Soruların hepsinin herkese aynı anda açıklanmak zorunda kalınması özellikle sınav güvenliği açısından sorun teşkil ederken beraberinde her sınav için hazırlanması gereken soru miktarının artması anlamına gelmektedir. Ekstra soru hazırlama maliyeti artırmanın yanı sıra soru kalitesine de etki edeceğinden soruların tamamını tek seferde sunulmadan sınav yapabilen test sunum yöntemleri ölçüm kalitesi açısından önem arz etmektedir.

MST'nin kitlesel eğitim hizmeti veren açık ve uzaktan öğrenme sistemlerinde uygulanabilirliği ve olası sonuçları hakkında fikir sahibi olabilmek amacıyla gerçek durumu temsil eden simülatif verilerin kullanılması düşük maliyet ile araştırma sorularına yanıt bulunmasını sağlayacaktır. Bu sayede MST yöntemi ile tasarlamış bir sınavın Açık ve Uzaktan Öğrenme (AUÖ) sisteminde ne tür avantajlar sağlayabileceği önceden kestirilerek optimum sınav koşullarının ortaya konulması konusunda önemli ölçüde veri sağlayacağı düşünülmektedir. Dolayısıyla bu araştırma sonuçlarından elde edilen bulguların özellikle geniş ölçekli sınav hizmeti veren kurumlara optimum test yöntemini belirlemede rehber niteliğinde veriler sunmayı hedeflemesi bakımından da yararlı olacağı düşünülmektedir.

1.4. Sınırlılıklar

Araştırmanın sınırlılıkları aşağıda maddeler halinde verilmiştir:

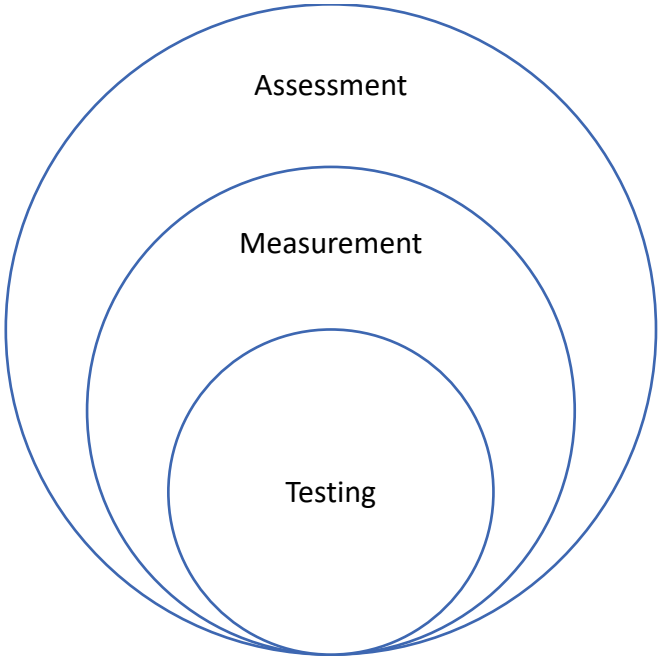
1. Araştırma binary (1-0) puanlama ile sınırlıdır.
2. IRT'nin 1, 2 ve 3 PL modelleri ile sınırladır.
3. Araştırmada kullanılan veriler tek faktörlüdür.
4. MST'nin MST-R ve MST-S yöntemleri ile sınırlıdır.

2. ALAN YAZIN

Bu bölümde tez çalışmasının odaklandığı konulara yönelik alan yazın taramasında elde edilen bulgular sunulmuştur.

2.1. "Test", "Ölçme" ve "Değerlendirme"

"Test", "Ölçme" ve "Değerlendirme" öğrenme sürecinin nasıl ilerlediğini ve öğrenenlerin nihai öğrenme çıktılarının nasıl değerlendirildiğini açıklamak için eğitimde kullanılan temel kavramlardır (Adom, Mensah ve Dake, 2020). Ölçme ve değerlendirme, psikolojiden tıp bilimine ve hatta iş dünyasında insan kaynakları yönetimine kadar daha pek çok alanda olduğu gibi öğrenme alanında da çeşitli formlarda önemi sıklıkla gündeme gelen konular arasında yer almaktadır (Nunnally ve Berstein, 1994). Ölçme ve değerlendirme hedefe ulaşmanın ya da ulaşamamanın bir göstergesi olup özellikle geniş kitlelere uygulanan sınavlarda en yaygın kullanılan ölçme aracı "test"lerdir (Adom, Mensah ve Dake, 2020; Masters, 2014). Eğitim uygulamalarında "Test", "Ölçme" ve "Değerlendirme" terimlerinin ön planda yer alan hâkim kavramlar olması sebebiyle birbirleri ile bağlantı noktaları/ilişkileri ana hatlarıyla bu başlık altında incelenmiştir. "Test", "Ölçme" ve "Değerlendirme" kavramları arasındaki söz konusu ilişki Şekil 2.1.'de görsel olarak belirtilmiş olup her bir kavram tezin sınırlılıkları kapsamında temsil ettiği anlamlar çerçevesinde açıklanmaya çalışılmıştır.



Kaynak: Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language testing, 18*(4), 351-372. Erişim adresi: https://journals.sagepub.com/doi/pdf/10.1177/026553220101800403

Şekil 2.1. *Test, ölçme ve değerlendirme*

Test, genellikle standartlaştırılmış koşullar altında bir veya daha fazla yeteneği ölçmek için kullanılan bir uygulama sürecini ifade etmektedir (Adom, Mensah ve Dake, 2020; Braun vd., 2006). Bu uygulama (ölçme) sürecinde belirli kriterlere göre yeterliliği belirlemek için ölçme aracı olarak testlerin kullanılması değerlendirme aşamasında yargıda bulunma sürecinin de kriterlere ve kanıtlara dayandığı anlamına gelmektedir (Overton, 2012). Bu nedenle Şekil 2.1.'de görsel olarak ifade edildiği üzere "Test", "Ölçme" ve "Değerlendirme" kavramları birbiriyle ilişkili olup "Ölçme" ve "Test", "Değerlendirme"nin bileşenleri olarak görülmekte ve "Değerlendirme" ölçüm ve test ile sınırlı kalmayan bir üst terim olarak tanımlanmaktadır (Lynch, 2001). Diğer bir ifadeyle bireyler hakkında karar vermek için topladığımız sistematik bilgiler testlerden veya diğer ölçüm yöntemlerinden gelmekle birlikte ölçme ve test dışında farklı bilgi kaynakları da bulunmaktadır. Kısaca eğitim sistemlerinin öğrenme sürecinde testlerin ve diğer ölçme araçlarının kullanılmasının özü, öğrenenlerin ders hedeflerine ulaşma yolundaki ilerlemelerini izlemek ve yönledirmektir (Eleje, Onah ve Abanobi, 2018).

Ölçme, alan yazında sayısız şekilde farklı tanımları bulunan bir kavramdır. Ölçmenin en klasik tanımı; "belli kurallara göre nesnelere veya olaylara sayıların atanması" olarak ifade edilmektedir (Stevens, 1946, s. 677). Genel olarak alan yazında sıklıkla karşılaşılan diğer *ölçme* tanımları ise: "Nitelikleri nicelemek; gözlem sonuçlarını sayısallaştırmak, bir niceliğin gözlenip, gözlem sonuçlarını sayı ve sembollerle göstermek, ilgilenilen bir nesneye bir dizi kurala göre bir nicelik atamak" şeklinde sıralanmaktadır (Kizlik, 2014; Masters, 2014; Murphy ve Davidshofer, 2005; Stevens, 1968; Stevens, 1946). Bu tanımlara göre *ölçme*, bir kavramın veya fiziksel nesnenin niteliklerine veya boyutlarına genellikle sayısal olan bir değerin atandığı süreçtir (Braun vd., 2006). Bu teknik tanımların yanı sıra ölçme kavramının kökeni tarihsel anlamda incelendiğinde 19. yüzyılın sonlarında Galton'un, insanların neden ölçülmesi gerektiği sorusunu ele aldığı çalışması normatif testlerle ilgili en eski çalışma örneğidir (Galton 1890). Galton çalışmasında insanın görme yeteneği hakkında şunları belirtmektedir: Ölçme, birey görme bozukluğunu kendisi fark etmeden çok önce ya da görme bozukluğu başkalarının dikaktini çekmeden önce görme yeteneğinin azaldığına dair bir gösterge vermektedir. Bu gösterge ölçmenin sayıların kurallara göre atanmasından (yani etiketlemeden) çok daha fazla anlam içeren değişkenin doğasını anlamaya yönelik bir süreç olduğunu ortaya koymaktadır. Bu durum eğitimsel ölçme bağlamında değerlendirildiğinde ise ölçümlerden elde edilen sonuçlar, eğitimsel değerlendirme için önemli kaynaklar sunmaktadır (Hori, Fukuhara ve Yamada, 2020). Bu sebeple ölçme süreci ölçüm işlemini gerçekleştirecek kişi ya da kurumların üzerinde titizlikle durması gereken en temel konular arasında yer almaktadır.

Değerlendirme, amaç ve hedeflere göre bilgilerin elde edildiği (Kizlik, 2014; Masters, 2014) ve ölçme sonuçlarının bir ölçütle karşılaştırılması sonucunda ölçülen nitelik hakkında bir değer yargısına (karara) varıldığı süreçtir (Braun vd., 2006; Doğan, 2019). Baehr (2005) değerlendirmeyi başarı ya da başarısızlık anlamında bir standardın karşılanıp karşılanmadığını belirlemek olarak tanımlarken; Lynch (2001), bireyler hakkında karar vermek veya yargıda bulunmak amacıyla sistematik bilgi toplama süreci olarak tanımlamaktadır. Overton'a (2012) göre değerlendirme; öğrencilerin akademik başarılarını izlemek, müfredat standartlarının başarısını ölçmek gibi öğrenci ilerlemesinin birçok değerlendirme türünü içermektedir. Tanımlardan da anlaşılacağı üzere değerlendirme yorum, yargı, karar gibi ifadeleri içeriğinde barındırması dolayısıyla

ölçmeyi içine alan ve ölçmeye göre çok daha kapsayıcı bir süreç olarak nitelendirilmektedir (Millman, Bishop ve Ebel, 1965). Değerlendirme, ölçme süreci sonucunda elde edilen verilerin anlam kazandığı önemli bir basamak konumundadır. Kearney'e (1983) göre geniş ölçekli bir değerlendirme; yerel, bölgesel veya ulusal düzeyde kamuya raporlama yapmayı, gereksinimleri belirleyerek kaynakları tahsis etmeyi ve terfi, mezuniyet gibi konular hakkında karar vermeyi amaçlamaktadır (s. 9-11). Bu amaçlar değerlendirmenin Şekil 2.1.'de sunulan kapsayıcı yönünü vurgulamaktadır.

2.2. Öğrenme Sürecinde Ölçme ve Değerlendirme

Öğrenme sürecinin sonucunu görebilmek için sürekli bir ölçme ve değerlendirme aşamasına gereksinim duyulduğundan "test", "ölçme" ve "değerlendirme" kavramları dünya çapında eğitim uygulamalarında hâkim kavramlar olmaya devam etmektedir (Adom, Mensah ve Dake, 2020; Masters, 2014). Bu açıdan bakıldığında öğrenme süreçlerinde ölçme ve değerlendirme aşaması (ister test ile ister diğer ölçme araçları aracılığıyla yapılsın) sürecin çıktısı ile ilgili olarak öğrenenin kazanımlarına yönelik veri elde edebilmenin (Braun vd., 2006; Masters, 2014) yanı sıra etkililiği ve verimliliği konusunda da önemli bilgiler sağlamaktadır. Bu doğrultuda ölçme ve değerlendirme sürecinin eğitim sisteminin yeniden yapılandırma ihtiyacı olan yönleri belirlenerek geleceğe dönük önlem alınabilmesi noktasında da dikkate değer katkısı bulunmaktadır. Bu sebeple ölçme ve değerlendirme öğrenciler hakkında eğitimsel kararlar vermek, öğrenciye ilerlemesi, güçlü ve zayıf yönleri hakkında geri bildirim vermek, öğretimin etkililiğini ve müfredat yeterliliğini yargılamak için kullanılan bilgileri elde etme süreci olarak eğitim sisteminin yaşamsal bileşeni konumundadır.

Aşağıdaki bölümde aşağıda belirtilen "Yetenek", "Şans Başarısı", "Güvenlik", "Maliyet" kavramları tez çalışması kapsamında ifade ettikleri anlamların sınırlılıklarında açıklanmaya çalışılmıştır.

2.2.1. Yetenek

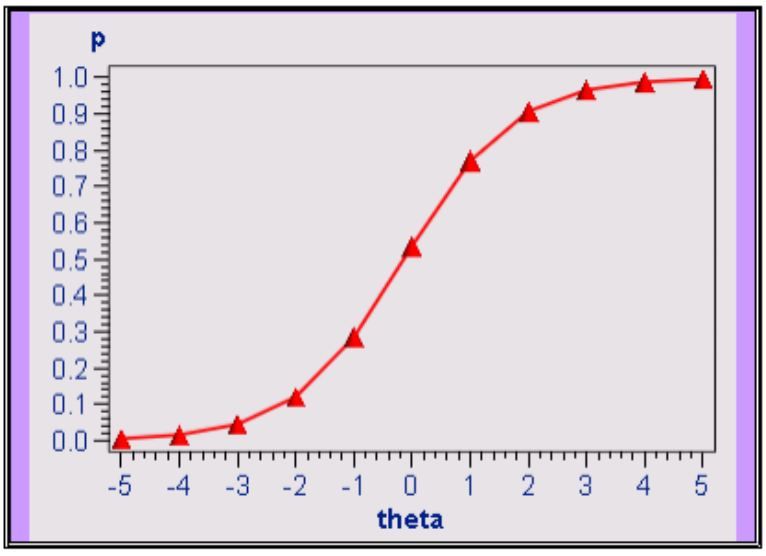
"Yetenek" kelimesi taşıdığı anlam itibariyle oldukça geniş bir alana hitap edebilen bir kavramdır. Gündelik yaşamda sıklıkla kullanılan/dile getirilen yetenekli birey, üstün yetenek, fiziksel (bedensel) yetenek, özel yetenek, yetenek havuzu vb. pek çok kavram insanlığın/toplumun "yetenek" ifadesi ile ne denli iç içe olduğunun önemli bir göstergesi

olarak kabul edilebilir. Bu durum yeteneği anlayabilmek ve anlamlandırabilmek adına yapılan tanımların çeşitlenmesine yol açmaktadır. Türk Dil Kurumu (TDK) tarafından yapılan yetenek tanımları söz konusu çeşitliliğin güzel bir örneği olabilir. TDK, yeteneği genel anlamda; "*Bir kimsenin bir şeyi anlama veya yapabilme niteliği, istidat, kabiliyet, kudret*" veya "*Bir duruma uyma konusunda organizmada bulunan ve doğuştan gelen güç, kapasite*" olarak tanımlarken eğitim bilimleri bağlamında "*Kişinin kalıtıma dayanan ve öğrenmesini çerçeveleyen sınır*" veya "*Dışarıdan gelen etkiyi alabilme gücü*" olarak tanımlamaktadır ("Türk Dil Kurumu", t.y.). TDK örneğinde olduğu gibi "yetenek" tanımı yapılırken birçok değişkenin dikkate alınmasının gerekliliği ve dönemin koşullarına göre çerçevesinin yeniden belirlenmesi gibi durumlardan dolayı herkes tarafından kabul gören genel geçer bir tanım yapabilmek mümkün görünmemektedir.

Yetenek kavramının bu çeşitliliği özellikle yetenek ölçümünü görev edinen (yetenek ölçümünden sorumlu olan) kişi ya da kurumları çok sayıda değişkeni göz önünde bulundurmaya/dikkate almaya sevk etmektedir. Yetenek kavramının geniş kapsamı sebebiyle sanat alanından tutunda tıp bilimlerine hatta dil becerilerine kadar daha pek çok farklı disiplinde "yetenek"ten bahsedilmektedir. Yeteneğin sürdürülebilir olması ve uzun vadede yetenekten söz edebilmek için yeteneğin özünde barındırdığı farklı disiplinlerin yanı sıra bireysellik, şeffaflık ve kültür gibi toplumun temel bileşenlerinin de yeteneği ölçme süreçlerine dahil edilmesi gerekmektedir/önem arz etmektedir. Bu durumun en güzel örneklerinden biri "Avrupa Dil Portfolyosu (European Language Portfolio: ELP)"nda dil yeteneği için beceri setleri hazırlanırken bireysellik ve kültürel boyutların da sürece dahil edilerek derinlemesine bir bakış açısı ile ele alınmış olmasıdır ("European Language Portfolio: ELP", t.y.). Bir diğer örnek ise insan kaynakları yönetiminde beceri setlerini temel alan yetenek yönetimine odaklanan stratejilerin uygulanması olarak verilebilir. Blass'a (2007) göre organizasyonlar açısından şeffaf bir yetenek yönetimi sistemine sahip olmanın anahtarı bu sistemin organizasyon içerisinde yerleşik hale gelmesine ve kültürel olarak kabul görmesine bağlıdır (s. 7-9).

Yukarıda yer alan açıklamalardan da anlaşılacağı üzere yetenek söz konusu olduğunda pek çok alan ve değişken sürece dahil olmaktadır. Bu nedenle "yetenek" kavramı tez çalışması sınırlılıklarında ölçme ve değerlendirme alanı için ifade ettiği anlam çerçevesinde açıklanmaya çalışılmıştır. Bu kapsamda "yetenek", "Theta: θ" değeri diğer adıyla "Öğrenci Yetenek Parametresi"dir.

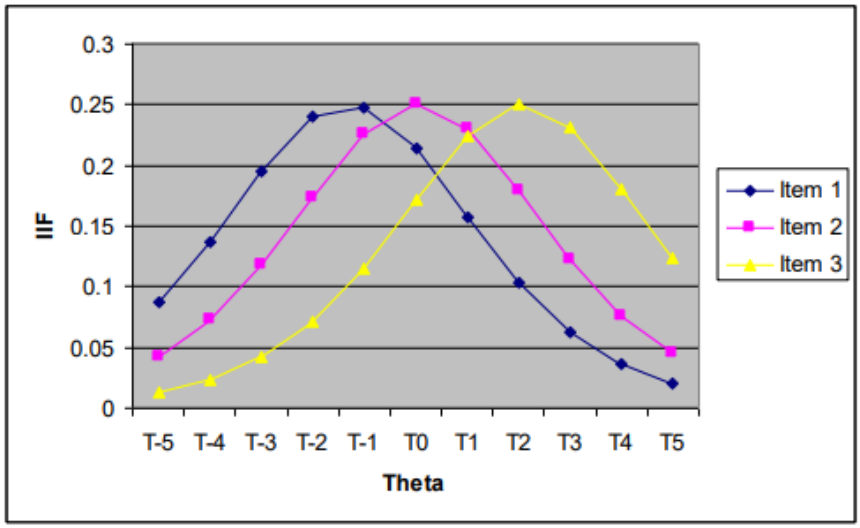
Ölçme ve değerlendirme alanında sınava giren kişinin yeteneği (zihinsel özelliği), genellikle bir grup test maddesine verdiği yanıtların fonksiyonu olan bir test puanı (her bir yanıtın 0 veya 1 olarak puanlandığı madde puanlarının toplamı) cinsinden ölçülmektedir (Lord, 1951). Bir diğer ifadeyle bireyin ölçme aracına yaptığı işaretlemelerine (doğru ve yanlış yanıtlanan maddelerden yola çıkarak) göre bir puan hesaplanmaktadır. Hesaplanan bu puan yanıtlayıcının yetenek düzeyi olarak nitelendirilmekte ve θ (theta) sembolü ile gösterilmektedir. Farklı θ seviyelerinde doğru cevabı verme olasılıkları elde edildikten sonra olasılıklar ile θ arasındaki ilişki (sınava giren kişinin bir maddeyi doğru yanıtlama olasılığı ile sınava giren kişinin yeteneği arasındaki ilişki) Şekil 2.2' de görüldüğü üzere Madde Karakteristik Eğrisi (ICC: Item Characteristic Curve) olarak grafik biçimde sunulmaktadır (Baker ve Kim, 2004, s. 1-22; Eleje, Onah ve Abanobi, 2018; Hori, Fukuhara ve Yamada, 2020; Yu, 2013). ICC'ler, test geliştiricilerin bir değerlendirmede belirli maddelerin nasıl çalıştığını anlamalarına yardımcı olmak için yaygın olarak kullanılmaktadır (Ockey, 2012, s. 340-341). ICC'leri yorumlayabilmek, belirli bir değerlendirme aracı için hangi öğelerin korunacağına, değiştirileceğine veya atılacağına karar verebilmek için önemlidir.



Kaynak: Yu, C. H. (2013). A simple guide to the item response theory (irt) and rasch modelling. *Published in http://www.creative-wisdom.com/computer/sas/IRT.pdf*.

Şekil 2.2. *Madde Karakteristik Eğrisi*

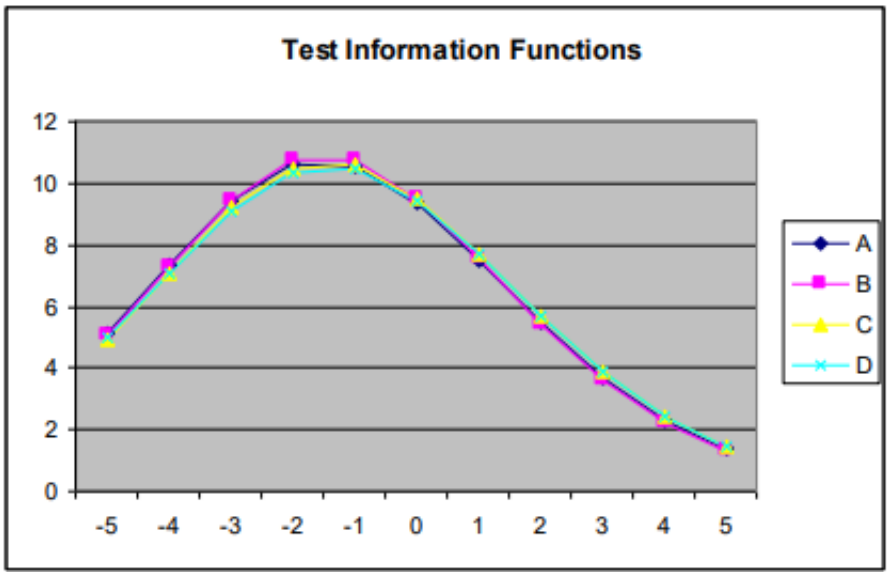
Yetenek ölçümü için test sunum sürecinde sıklıkla bahsi geçen önemli diğer iki grafik ise "Madde Bilgi Fonksiyonu (IIF: Item Information Functions)" ve "Test Bilgi Fonksiyonu (TIF: The Test Information Function)" olarak ifade edilmektedir. Söz konusu grafikler Şekil 2.3. ve Şekil 2.4'te sunulmuştur.



Kaynak: Yu, C. H. (2013). A simple guide to the item response theory (irt) and rasch modelling. *Published in http://www.creative-wisdom.com/computer/sas/IRT.pdf.*

Şekil 2.3. *Madde Bilgi Fonksiyonu*

Şekil 2.3'de bir testin üç maddesinin IIF (Item Information Functions)'leri diğer adıyla Madde Bilgi Fonksiyonları gösterilmektedir. Bu IIF'ler birbirinden farklıdır ve temsil ettikleri anlamlar şu şekilde ifade edilmektedir (Baker, 2001; Baker ve Kim, 2004, s. 23-59; Hori, Fukuhara ve Yamada, 2020; Yu, 2013): Mavi çizgi ile belirtilen Madde 1'de, θ (theta) seviyesinin -1 olduğu nokta "tepe" bilgisini vermektedir. θ (theta) seviyesi -5 olduğu noktada da 0,08 oranında bir miktar bilgi mevcuttur. Ancak θ (theta) seviyesi 5 olduğunda neredeyse hiç bilgi bulunmamaktadır. Pembe çizgi ile belirtilen Madde 2 ise çoğu bilgi θ (theta)'nın sıfır olduğu noktada ortalanırken, en düşük θ (theta)'daki bilgi miktarı ile en yüksek θ (theta)'daki bilgi miktarı aynı düzeyde konumlanmaktadır. Sarı çizgi ile belirtilen Madde 3 ise mavi çizgi ile belirtilen Madde 1'in tersi durumda konumlanmaktadır. Bir kişi üst düzey bir θ (theta) yakınlarında çok fazla bilgiye sahip olabilirken θ (theta) seviyesi alt uca yaklaştıkça bilgi önemli ölçüde düşmektedir.



Kaynak: Yu, C. H. (2013). A simple guide to the item response theory (irt) and rasch modelling. *Published in http://www.creative-wisdom.com/computer/sas/IRT.pdf.*

Şekil 2.4. *Test Bilgi Fonksiyonu*

Test Bilgi Fonksiyonu (TIF: The Test Information Function) ise kısaca testteki tüm IIF (Item Information Functions)'lerin toplamıdır. IIF belirli bir madde parametresine dair bilgi verirken TIF de aynı bilgiyi sınav düzeyinde vermektedir (Baker, 2001; Baker ve Kim, 2004, s. 78-80; Hori, Fukuhara ve Yamada, 2020; Yu, 2013).

2.2.2. Şans Başarısı

"Şans Başarısı" kavramı ilerleyen bölümlerde "Madde Tepki Kuramı/Item Response Theory (MTK/IRT)" ana başlığı altında "c parametresi" alt başlığı ile detaylı olarak ele alındığından bu bölümde ana hatları ile genel bir biçimde açıklanmıştır.

Tahmin parametresi olarak da bilinen şans başarısı IRT 3PL modelde "c parametresi" değeri ile temsil edilmektedir. "c parametresi" değeri en temel anlamıyla; yeteneği çok düşük olan bir kişinin bir maddeye verilen doğru yanıtı tahmin edebilmesi ve dolayısıyla doğru yanıtlama olasılığının sıfırdan büyük olması durumudur (Thorpe ve Favia, 2012; Yu, 2013, s. 7-11). Örneğin, çoktan seçmeli test tekniğinin kullanıldığı bir sınavda dört seçenekli maddelere rastgele yanıtlar veren bir kişi, bu maddelere yaklaşık 4 defada 1 doğru yanıt verebilir, yani doğru cevabı tahmin etme olasılığı (şansa dayalı hata miktarı) yaklaşık 0.25 civarındadır. Obinne (2012) çalışmasında "c parametresi" değerinin 0.20'nin üzerinde kalan maddelerin çok iyi maddeler olmadıklarını tespit

etmiştir. Ölçme ve değerlendirme alanında bu hata değerinin 0.20'nin altında olması beklenmekte ve bu oranın üzerindeki değerlerin ortaya çıktığı durumlarda ilgili maddelerin çok iyi olmadığı sonucuna varılmaktadır (Harris, 1989; Warm, 1978).

2.2.3. Güvenlik

"Güvenlik" olgusunun insanlık tarihi kadar eskiye dayanan bir kavram olması dolayısıyla tarih boyunca üstlendiği anlamlar itibariyle oldukça geniş bir kapsamı bulunmaktadır. Bu sebeple tek bir evrensel güvenlik tanımı yapmak mümkün görünmemektedir. Gündelik yaşamda güvenliğe dair sıklıkla kullanılan "Sosyal Güvenlik", "Uluslararası İlişkilerde Güvenlik", Ulusal Güvenlik", "Bölgesel Güvenlik", "Çevresel Güvenlik", "Sektörel Güvenlik", "Ekonomik Güvenlik", "İnsan Güvenliği", "Hasta Güvenliği", "Gıda Güvenliği", "İş güvenliği", "Bilgi Güvenliği", "Sınav Güvenliği", "Ölçüm Güvenliği", "Test Güvenliği" gibi kavramlar bu durumu net bir biçimde örneklendirmektedir.

Alan yazında güvenlik kavramı birçok araştırmacı tarafından farklı bakış açıları ile tanımlanmaktadır: Wolfers (1952) güvenliği öznel ve nesnel olmak üzere iki farklı çerçevede ele almaktadır. Wolfers, öznel güvenliği sahip olunan değer ya da değerlerin saldırıya uğrayacağı korkusunun yokluğu olarak tanımlarken nesnel güvenliği sahip olunan değer ya da değerlere yönelik herhangi bir tehditin yokluğu olarak tanımlamaktadır (s. 481-502). Baylis (2020) ise Wolfers'e atıfta bulunarak güvenliği temel değerlere karşı tehdidin yokluğu olarak tanımlamaktadır (s. 241-255). McSweeney (1999) geniş kapsamı nedeniyle güvenlik kavramının tanımlanmaya direnen bir terim olduğunu vurgularken (s. 13-21), Buzan (2008) güvenlik terimi tanımının muğlaklığına vurgu yaparak tartışmaya açık bir kavram olduğunu belirtmektedir. Balwin (1997) güvenlik kelimesinin anlamının yeterince açıklanamamış (yetersiz şekilde açıklanmış) olduğu görüşünü savunarak "Security for whom? (Kimin için güvenlik?)" ve "Security for which values? (Hangi değerler için güvenlik?)" soruları üzerinden tanımlanması gereği konusu üzerinde durmaktadır. Bu durum Wolfers'in (1952) öznel ve nesnel güvenlik tanımında yer alan sahip olunan değerler vurgusuyla örtüşmektedir.

Tez konusu kapsamında odaklanılan "güvenlik" kavramları ise eğitim bilimleri alanında sıklıkla gündeme gelen ve birbiriyle yüksek oranda ilişkili olan "sınav

güvenliği", "ölçüm güvenliği", "test güvenliği"dir. Eğitim bilimleri alanında "sınav güvenliği" kavramı kopya çekme sorununu temsil etmektedir (Jung ve Yeom, 2009; Turani, Alkhateeb ve Alsewari, 2020). Özellikle 2019 yılında başlayıp tüm dünyayı saran covid-19 pandemisi sebebiyle sınavların çevrimiçi ortamlarda gerçekleştirilmek durumunda kalınması sınav güvenliğine duyulan ihtiyacın önemini bir kez daha vurgulamıştır. Bu nedenle sınav güvenliği konusunda en çok gündemde olan konulardan bir tanesi kopyanın önüne nasıl geçilebileceği hususudur. Çevrimiçi sınav ortamlarında güvenlik sorunları olarak nitelendirilen; yanıtlar için internette arama yapmak, başkalarının yanıtlarını kopyalamak, daha önce sınava giren bireylerin sınava daha sonra giren katılımcılarla bilgi paylaşma olasığının bulunması, öğrencinin yerel bilgisayarında kayıtlı verileri ve yazılımı kullanması (McCabe, Trevino Butterfield, 2001; Rogers, 2006), soru, yanıt ve notların değiştirilmesi, sınav öncesi soru kağıtlarına yetkisiz erişim (Savulescu, Polkowski ve Alexandru, 2015), soruların sosyal platformlarda hızla yayılması, sınavı e-posta, telefon veya anlık mesajlaşma yoluyla tartışmak gibi farklı kopya çekme kalıpları bulunmaktadır/mevcuttur (Ko ve Cheng, 2004). Bilgisayar tarafından bir test penceresi aracılığıyla yönetilen çevrimiçi testler sınav güvenliği bakımından barındırdığı risklere rağmen geleneksel kâğıt ve kalem (P&P: Paper&Pencil) testleri ile karşılaştırıldığında daha yüksek derecede kontrol sağlamaya olanak tanımakta ve test güvenliği arttırma fırsatı sunmaktadır (Wang, Zheng ve Chang, 2014). Örneğin her öğrenciye farklı bir problem seti verme (McGough, vd. 2001) sınav odasını kısıtlama veya yanıt gönderim sayısını sınırlama (DePiero, 2001; Rogers, 2006) gibi yöntemler güvenlik (kopya) sorunu ile mücadele etmede kullanılmaktadır.

Eğitim bilimleri alanında güvenlik ile ilgili bir diğer konu ise "ölçüm güvenliği"dir. Çevrimiçi testlerin en yaygın uygulanma biçiminin çoktan seçmeli test tekniği olması sebebiyle çevrimiçi sınavlarda "ölçüm güvenliği"nin sağlanması öncelikle "test güvenliği"nin sağlanmasını gerektirmektedir. Çevrimiçi testler, web tabanlı teknolojilerin ortaya çıkmasıyla birlikte büyük ölçekli eğitim değerlendirmelerinde ana akım mod haline gelmiş durumdadır. Büyük ölçekli ve yüksek riskli sınavlarda bilgisayar tarafından yönetilen testlerin giderek daha fazla benimsenmesiyle test güvenlik kontrolü ölçme ve değerlendirme alanında son derece önemli konular arasında bulunmaktadır. Alan yazında test güvenliğini değerlendirmek için özellikle madde havuzlama konusunda gerçekleştirilmiş çeşitli çalışmalar bulunmaktadır (Chang, 2004; Chang ve Zhang, 2002;

Stocking, 1994, Way, 1998). Geçmişte yapılan araştırmalar çoklu madde havuzlarının kullanılmasının CAT gibi sürekli testlerde test güvenliğini artırmak için geçerli bir strateji olarak kabul edilebileceğini göstermektedir (Ariel, Veldkamp ve van der Linden, 2004; Barrada, Olea ve Abad, 2008; Davey ve Nering, 2002; Zhang, Chang ve Yi, 2012). Bu anlamda ölçüm kalitesi ve güvenlik kontrolünü sağlamanın bir yolu madde gruplarını bir araya getirerek havuzdaki maddelerin kullanım oranlarını yönetmektir (Wainer ve Kiely, 1987). Bu süreçte test güvenliği açısından madde kullanım oranı ne kadar düşükse test örtüşme(çakışma) oranınında (madde kullanım sıklığı) o kadar düşük olması beklenlenmekte ve test örtüşme(çakışma) oranlarının ortalamaları ve standart sapmaları aracılığıyla değerlendirilmektedir (Chen, Ankenmann ve Spray, 2003; Wang, Zheng ve Chang, 2014). Test güvenliği bağlamında bir diğer konu ise ayırt ediciliği yüksek maddelere sahip bir kısa testin ayırt ediciliği düşük uzun bir teste göre daha güvenilir bir θ (theta) değeri sunabiliyor olmasıdır. Örneğin, standart ölçüm hatası farklı θ (theta) seviyelerinde değişkenlik gösterdiğinden kısa testler uzun testlerden daha güvenilir olabilmektedir (Embretson ve Reise, 2000). Diğer bir ifadeyle belirli θ (theta) düzeylerinde oldukça ayırt edici olan maddeler kısa bir testte sunulabilmekte ve KTK'dakinden daha güvenilir bir θ (theta) tahmini sağlayabilmektedir. Farklı formların (madde kümelerinin) farklı θ (theta) seviyelerindeki yanıtlayıcılar için en iyi olduğunu belirten kurala (IRT) göre; yanıt verenler ne kadar farklı olursa, yanıtlayanın θ (theta) düzeyini daha iyi değerlendirmek için onlara sunulan madde kümeleri de o kadar farklı olmalıdır (Kline, 2005). Sınav katılımcılarının yetenek düzeylerine göre farklı soruları yanıtlaması ölçüm doğruluğu ve kalitesine katkı sağlarken diğer yandan da güvenlik problemine önemli bir katkı getirmektedir. Bu sebeple sınav katılımcısına hem nitelikli soru sormak hemde madde görüntülenme hızını (tüm soruların aynı anda açıklanmasını) sınırlandırarak sınav, test ve ölçüm güvenliğini sağlayabilmek ölçüm hassasiyetinin yüksek öneme sahip olduğu ölçme ve değerlendirme süreçlerinin ana hedefi olarak çözümü öncelikli konular arasında yer almaktadır.

2.2.4. Maliyet

Maliyet, TDK sözlüğünde en genel anlamıyla "*üretimde bir mal elde edilinceye değin harcanan değerlerin toplamı*" olarak tanımlanmaktadır ("Türk Dil Kurumu", t.y.). Dolayısıyla türü ne olursa olsun bir ürün ortaya koymak belirli bir miktar maliyet

gerektirmektedir. Eğitim bilimleri alanında da nihayetinde bir ürün ortaya koyan tüm süreçler gibi hedeflenen ürün elde edilinceye kadar harcanan değerler bir maliyeti temsil etmektedir. Eğitim bilimleri sürecinin önemli bir adımı olan ölçme ve değerlendirme aşamasında gerçekleştirilen sınavların çevrimiçi ya da yüz yüze olması farketmeksizin uygulanan test sunumunun bir maliyeti bulunmaktadır. Bu sebeple tez çalışmasında maliyet kavramı, bilgisayar uygulamalı testlerden en bilinen CAT test sunum yöntemi ve son zamanlarda alternatif yöntem (Luecht, 2005) olarak sıklıkla gündeme gelen MST test sunum yöntemi uygulanma maliyeti açısından sınırlandırılarak ele alınmıştır.

Bilgisayar tabanlı test uygulamaları kullanılabilir geniş bir madde havuzu gerektirmekte ve bu türden bir havuz oluşturmak ciddi anlamda mali alt yapıya gereksinim duymaktadır. Ayrıca sürekli olarak tek tip testler sağlamak için havuzdaki maddelerin de psikometrik ve kategorik açıdan sabit niceliksel nitelikleri olmalıdır. Kalite değişkendir ve bu koşulları sağlamak adına girişilen süreçler (yeni öğelerin yazılması, saklanması ve önceden test edilmesi vb.) kaynak gerektirdiğinden böyle bir uygulamanın sürekliliği için yüksek maliyetlere gereksinim duyulmaktadır (Ariel, Veldkamp ve Breithaupt, 2006; Glas ve van der Linden, 2003). Örneğin; soruların hazırlanması için uzman ekip, yazılımsal anlamda sistemin kurulması ve yönetilmesi için nitelikli kadro, tüm işlemler için gereksinim duyulan ekipmanlar vb. süreçlerin başlangıcı ve sürdürülebilirliği ciddi anlamda yüksek maliyetlere sahip işlemlerdir. Özellikle geniş kitlelere sınav uygulayan yapılar açısından bu derece yüksek maliyetlere sahip bir test sunum yönteminin sürdürülebilirliği güçleşmektedir. Araştırma sınırlılıkları çerçevesinde CAT ve MST yöntemleri maliyet bağlamında karşılaştırıldığında en önemli fark şu şekilde ifade edilebilir (Ariel, Veldkamp ve Breithaupt, 2006; Hendrickson, 2007; Wainer ve Kiely, 1987; Wang, Zheng ve Chang, 2014; Yan, Lewis ve Davier, 2014, s. 4-20);

- CAT madde bazlı soru havuzu yapısında sahip olmasından dolayı daha yüksek maliyetli bir seyir izlerken,
- MST modül bazlı yapısından kaynaklı olarak CAT yöntemine oranla mali anlamda daha avantajlı bir seyir izlemektedir.

IRT yöntemleri doğru parametre tahminleri elde etmek için önemli örneklem büyüklüklerine gereksinim duymakta olduğundan bilgisayar aracılığıyla yönetilmesi maliyetli bir yöntemdir (Kline, 2005, s. 163-166; Ockey, 2012. s. 347; Yan, Lewis ve von

Davier, 2014, s. 19). IRT yöntemlerinin bu yönü yüksek mali kaynak gerektiren ve gerçekleştirmesi zor bir süreç olması dolayısıyla önemli bir dezavantaja sahiptir (Jabrayilov, Emons ve Sijtsma, 2016). Bu durum MST gibi alternatif test sunum yöntemlerine gereksinim duyulmasının temel sebepleri arasında yer almaktadır. Bilindiği üzere mali değere sahip olan her türlü işlemde öncelikli amaç minimum maliyet ile maksimum fayda sağlamaktır. Ölçme ve değerlendirme alanında da yüksek maliyet gerektiren geniş kitlelere uygulanan büyük ölçekli sınavlarda düşük maliyetle ölçüm sağlayabilen test sunum yöntemleri ddiğerlerine göre avantajlı olarak değerlendirilmektedir.

2.3. Açık ve Uzaktan Öğrenmede Ölçme ve Değerlendirme

Ölçme ve değerlendirme süreci eğitim öğretim sisteminin tüm kademelerinde olduğu gibi açık ve uzaktan öğrenme (AUÖ) sisteminin de en temel yapı taşları arasında yer almaktadır. Açık ve uzaktan öğrenme (AUÖ) sistemleri geniş kitlelere hizmet vermeleri dolayısıyla ölçme ve değerlendirme işlemini sıklıkla çoktan seçmeli sorular (Doğan, 2019; Howel ve Hricko, 2005) aracılığı ile gerçekleştirmek durumunda kalmaktadırlar. Bu anlamda uygulanacak test tekniğinde en doğru yöntemi belirlemek öğrenen becerisini gerçek değere yakın oranda ölçmenin (Kizlik, 2014; Nunnally ve Berstein, 1994) yanı sıra zaman, emek ve maliyet açısından da dikkate değer öneme sahip konulardandır. Özellikle son zamanlarda çevrimiçi sınav uygulamalarının (Howel ve Hricko, 2005) artması sebebi ile öğrenenin bilgi düzeyini çoktan seçmeli sorularla maksimum güvenilirlikte ve doğru bir biçimde ölçebilmek (Howel ve Hricko, 2005; Keng, 2008) açık ve uzaktan öğrenme sistemlerinin önemle üzerinde durması gereken temel konular arasında yer almaktadır.

2.4. Ölçme ve Değerlendirmede Test Teorileri

Test teorileri; test puanları ve madde puanları gibi gözlemlenebilir değişkenleri gerçek puanlar ve yetenek puanları gibi gözlemlenemeyen değişkenlerle ilişkilendirerek genel bir çerçeve sağlayan yapılardır (Hambleton ve Jones, 1993, s. 38). Ölçme ve değerlendirme prosedürünün ana amacı geçerli ve güvenilir test maddeleri aracılığıyla sınava giren kişilere yönelik verilecek kararların dayandığı test puanları verebilecek bir ölçüm yaklaşımı benimsemektir (Chikezie ve Joseph, 2016). Ölçme ve değerlendirme alanında Klasik Test Teorisi (KTK) ve Madde Tepki Kuramı (IRT) olmak üzere iki tür

ana ölçüm teorisi bulunmaktadır. KTK ve bu teorinin geliştirilmiş versiyonu olan IRT ölçme teorisinin geliştirilmesinde bireyin özelliklerini tanımlamak ve bir konudaki yeteneklerini (gizil niteliklerini) analiz etmek için kullanılan iki temel puanlama yöntemidir (Eleje, Onah ve Abanobi, 2018). Bir test teorisi çerçevesinde formüle edilen test modellerinde bir dizi varsayımla birlikte kavramlar ve bunların arasındaki ilişkiler oldukça ayrıntılı bir şekilde belirtilmektedir (Hambleton ve Jones, 1993). Test teorileri ve ilgili modellemeleri ile bunların varsayımları bu bölümde alt başlıklarla genel bir çerçevede sunulmaktadır.

2.4.1. Klasik Test Kuramı (KTK)

Klasik Test Kuramı (KTK) diğer adıyla Classical Test Theory (CTT), ilk olarak yirminci yüzyılın başlarında geliştirilerek yüzyılın ortalarına kadar ölçüm çevrelerinde hâkim olan ve gözlenen test puanları ile bu puanları etkileyen faktörler arasındaki ilişkiler hakkında bir dizi varsayımdan oluşan ölçüm teorisidir (Brown, 2012, s. 323-335). Tüm ölçümlerin hatalar içerdiğinin kabulü KTK'nın tarihini Spearman'a (1904) kadar götürmektedir. En eski ve en yaygın ölçüm teorisi olarak kabul edilen KTK (Gulliksen, 1950), ölçüm hatalarının gözlenen puanları nasıl etkileyebileceğini açıklayan bir modeldir.

Klasik teori, çok sayıda ölçüm aracına temel teşkil eden ve son ölçüm yaklaşımları için bir referans noktası olarak hizmet eden bir dizi kavram ve ilgili tekniklerden oluşmakta olup "hata" ve "gerçek puan" kavramları KTK'nın merkezinde yer almaktadır (DeVellis, 2006). Bireylerin gerçek puanlarının X (elde edilen toplam puan/gözlenen puan) = T (gerçek puanı) + E (tesadüfi hata miktarı/hata bileşeni) eşitliğiyle (Bond ve Fox, 2015; Bond, Yan ve Heene, 2020; Lord ve Novick, 2008, s. 55-63; Lord, 1980; Weis ve Yoes, 1990) bulunabileceğini savunan Klasik Test Kuramı (KTK)'nın temelleri 1900'lü yılların başında Spearman tarafından atılmıştır (Brown, 2012; Crocker ve Algina, 2008; Spearman, 1904; Traub, 1997). Gerçek Puan Modeli (TSM: True Score Model) olarak da bilinenen teorinin arkasındaki temel fikir, gözlemlenen puanın (X) gerçek puan (T) ve hata puanı (E) olmak üzere iki bileşenden oluşmasıdır. X=T+E formülü ile temsil edilen KTK, bu üç değişken (X, T ve E) arasındaki ilişki ile ilgilenmektedir.

Klasik test kuramı temelinde hataya odaklanmakta ve hata ne kadar iyi tanımlanırsa ölçüm sonucunun da bir o kadar güvenilir olabileceği ve gerçeği yansıtabileceğini öne sürmektedir (Brown, 2012; Crocker ve Algina, 2008; Nunnally ve Berstein, 1994). Diğer bir ifadeyle ölçüm hatasını açıklama girişmi olan KTK'da gözlemlenen puanın hata ile ölçüldüğü varsayılmakta ve KTK'nın amacı bu hatayı en iyi şekilde tanımlamaktır (Courville, 2004, s. 2-30; McBridge, 2001).

KTK (Klasik Test Kuramı), açık ve uzaktan öğrenme sistemleri dahil eğitim-öğretim süreçlerinin ölçme ve değerlendirme alanında en yaygın kullanılan puanlama yöntemidir (Wallace ve Bailey, 2010). KTK'da, madde ortalamaları ve madde-test korelasyonlarının yanı sıra güvenilirlik katsayıları ve standart hatalar belirli popülasyonların özelliklerine bağlıdır. KTK'da gerçek puan ölçeği, tek bir ölçü üzerinde belirli bir dizi madde tarafından tanımlanmaktadır (Reise, Ainsworth ve Haviland, 2005, s. 96-97). Maddelerin güçlük (*pj*) ve ayırt edicilik (*rjx*) indeksleri grubun tüm verisi elde edildikten sonra belirlenebilmektedir (Crocker ve Algina, 2008). Çünkü KTK'da madde istatistikleri gruba bağımlıdır ve testin uygulandığı gruptan elde edilen sonuçlar toplam puana bakılarak değerlendirilmektedir (Cappelleri, Lundy ve Hays, 2014; Stage, 1998). KTK'da yapılan ölçümlerde tüm bireyler için standart hata aynı kabul edilmekte ve puanlama ona göre yapılmaktadır. Örneğin; Puanı 40 olan birisi için de 75 olan birisi için de ölçmenin standart hatası aynı olmaktadır. Oysa bugün bilmekteyiz ki ölçme araçları farklı puana sahip bireyler için farklı miktarda hata yapmaktadır.

Psikolojik veya eğitimsel ölçümler kapsamında uygulanan bir test genellikle sayısal bir puan elde edebilmek için araç olarak kullanılmaktadır (Lord, 1980, s. 3-10). Wallace ve Bailey (2010) tarafından da belirtildiği gibi psikometrik analizlerdeki yaygın uygulama; gözlemlenen ve gerçek katılımcı puanındaki farkı (gerçek puan varyasyonuna karşı gözlemlenen puan varyasyonunu) karşılaştıran klasik test teorisinin (KTK) kullanılmasıdır. Klasik Test Kuramı'nın ölçülen özellik ile bireylerin performansları arasında doğrusal ilişki varsayması, elde edilen parametrelerinin örnekleme bağımlı olması, standart hatayı ve güvenilirliği testi alan tüm bireyler için eşit kabul etmesi önemli kısıtlılıklardır (Crocker ve Algina, 2008; DeVellis, 2006; Hambleton, Swaminathan ve Roger, 1991; Lord, 1980; Weis ve Yoes, 1990) ve bu stratejiler özellikle test maddelerinin ölçüm kalitesini yeterince yansıtmamaları nedeniyle (Cappelleri, Lundy ve Hays, 2014) uzun süredir eleştiri altındadır (Arnold, 1996; Kline, 2005; Magno, 2009; Rusch, Lowry,

Mair ve Treiblmaier, 2017). Ancak, madde puanlarının (ağırlıklı) toplamının bir test puanı (örneğin, doğru yanıtların sayısı, madde puanlarının ortalaması) olarak kullanıldığı KTK'ya dayalı olarak bu sorunları çözmek mümkün görünmemektedir (Elejei, Onah ve Abanobi, 2018). Söz konusu sınırlılıklar ve eleştiriler gerek açık ve uzaktan öğrenme sistemlerinde gerekse geniş kitlelere merkezi sınav uygulayan diğer sistemlerde birey bazlı ölçüme olanak tanıyan ölçüm hassasiyeti yüksek modern puanlama ve test sunum yöntemlerine duyulan ihtiyacı net bir biçimde ortaya koymaktadır. Bu nedenle madde zorluğunun katılımcı yeteneklerinden bağımsız olarak belirlendiği ve böylece madde ve ölçek bilgilerinin istatistiksel güvenirliğe olan bağımlılığının ortadan kaldırıldığı "Madde Tepki Kuramı (MTK)" yöntemi bir alternatif olarak kullanılmaktadır. Hattie, Jaeger ve Bond (1999) çalışmalarında MTK'dan klasik test teorisinin neredeyse tüm eksikliklerini ortadan kaldıran hassas ve güçlü bir test performansı modeli olarak bahsetmektedirler. Aşağıda metnin genelinde "IRT" kısaltması ile kullanılmış olan "Madde Tepki Kuramı" yöntemi tüm model ve parametreleri ile birlikte detaylı biçimde sunulmuştur.

2.4.2. Madde Tepki Kuramı/Item Response Theory (MTK/IRT)

Modern test teorisi veya gizli özellik teorisi olarak da adlandırılan (Bond ve Fox, 2015; Bond, Yan ve Heene, 2020) Madde Tepki Kuramı'nın tarihçesi 1900'lü yılların başlarında bağımsız bir değişkene karşı performans düzeylerini çizen ve grafikleri test geliştirmede kullanan Binet ve Simon'a uzanmaktadır (Goldstein ve Wood, 1989; Hambleton ve Swaminathan, 1985; Linden ve Hambleton, 1997). Thurston (1925) ise çocukların zihinsel başarılarını yaş dereceli bir ölçekte değerlendirmek için tasarlanmış maddelerin nasıl yerleştirileceğini açıkladığı "*A Method of Scaling Psychological and Educational Tests*" başlıklı makalesinde IRT'ye dair önemli sinyaller vermektedir (Bock, 1997). İlerleyen yıllarda Richardson (1936), IRT parametreleri ile klasik madde parametreleri arasında IRT parametre tahminlerini elde etmek için bir başlangıç yolu sağlayan ilişkiler türetirken Lawley (1943) parametre tahmini için bazı yeni prosedürler ortaya koymuştur. Bu gelişmeler doğrultusunda tarihsel süreçte 1940'lı yılların başlangıcının IRT modellerinin geliştirilmesine tanıklık ettğinden bahsedilmektedir (Ferguson, 1942; Finney, 1944; Lawley, 1943; Tucker, 1946). IRT'nin tarihinde bu yönde somut anlamda yapılan ilk çalışmalar ise 1952 yılında Lord ile beraber başlamıştır. Lord, bu çalışmasında iki parametreli normal ogive modelini, model parametre tahminlerini ve

modelin dikkate alınan uygulamalarını tanımlamıştır (Courville, 2004; Hambleton ve Swaminathan, 1985; Linden ve Hambleton, 1997). Bu sayede teorinin varsayımlarının ana hatlarıyla belirlenmesi ve ayrıntılı modeller sağlanmış olması bir ölçüm teorisi olarak IRT'nin temellerinin Lord tarafından atıldığının göstergesidir (Ockey, 2012, s. 337) Birnbaum ise 1950'lerin sonlarına doğru normal ogive modellerinin yerine daha izlenebilir lojistik modelleri koydu ve bu yeni modeller için istatistiksel temeller geliştirdi (Courville, 2004, s. 11-49; Birnbaum, 2008, s. 365-379). 1960 yılına gelindiğinde ise Rasch daha sonra kendi adı ile anılacak olan ilk IRT modelini geliştirmiştir (Anderson, Kearney ve Everett, 1968; Engelhard ve Wang, 2020; Lord ve Novick, 2008; Owen, 1969). Rasch'ın IRT parametreleri konusundaki çalışmaları ABD'de Wright'ı ve Avrupa'da Andersen ve Fischer gibi psikologları etkileyerek dikkatlerini çekmeyi başarmıştır. Örneğin; Wright, 1970'ler boyunca ETS ve AERA gibi önemli organizasyonlarda gerçekleştirmiş olduğu sunularda Rasch modelinin birçok araştırmacı tarafından anlaşılmasına önemli ölçüde katkıda bulunmuştur. Diğer taraftan Lord ve Novick'in editörlüğünde hazırlanan "*Statistical Theories of Mental Test Score*" başlıklı kitapta gizli özellikler teorisi üzerine yoğunlaşılmış olması ve yazarların IRT'ye odaklanması kuramın gelişiminde teşvik edici rol oynamıştır (Lord ve Novick, 2008). Bu kitap hem yazarlarının hemde onlardan önceki yıllarda bu alanda çalışan araştırmacıların geliştirdikleri IRT ilkelerini özetlemektedir (Ockey, 2012). Hambleton ve Cook, Wright, Marco, Rentz ve Bahaw, Urry gibi araştırmacıların katılımıyla IRT uygulamalarında birçok ölçüm atılımının tanımlandığı "*Journal of Educational Measurement*"in 1977 yılında yayınlanan özel sayısı madde tepki kuramının gelişimine katkıda bulunan adımlar arasında sayılabilmektedir (JEM, 1977). Lord'un (1980) "*Applications of Item Response Theory to Practical Testing Problems*" ismini verdiği kitabında madde tepki kuramı üç parametreli modelin teorik gelişmeleri ve uygulamalarının güncel bir incelemesi bulunmaktadır. Tarihsel süreç boyunca atılan tüm bu adımlar madde tepki kuramının gelişimine ve günümüz ölçümlerinde modern puanlama yöntemi olarak kabul görmesine önemli katkılar sağlamıştır.

Madde tepki kuramı (IRT), psikolojik ölçümlerin geliştirilmesi, iyileştirilmesi, ölçeklerin uygulanması ve bireysel farklılıkların ölçeklendirilmesi için kullanılan bir dizi psikometrik modelden oluşmaktadır (Embretson ve Reise 2000). Madde tepki kuramı (IRT), eğitsel ve psikolojik testler (örneğin, standartlaştırılmış testler, kişilik testleri,

lisans testleri ve sertifikalandırma testleri vb.) geliştirmek için kullanılan bir gizli değişken modelleri sınıfıdır (Hori, Fukuhara ve Yamada, (2020). IRT kapsamında ölçülmesi hedeflenen özellik gözlemlenemez olduğundan alan yazında bu teori gizli özellik teorisi olarak da adlandırılmaktadır (Ockey, 2012, s. 336). Teori, bireylerin gözlemlenen performanslarını gözlemlenemeyen özelliğin altında yatan bir konumla ilişkilendirmeyi amaçlamaktadır. IRT yöntemi bu gözlemlenemeyen gizli özelliği sınava giren kişinin bazı gözlemlenebilir özelliklerine özellikle de testi oluşturan bir dizi soruya doğru yanıt verme yetilerine bağlamaya izin veren modeller önermektir (Bichi vd., 2015).

Madde Tepki Kuramı, doğrusal olmayan bir fonksiyon kullanarak sınava giren katılımcıların gizli bir değişken üzerindeki seviyeleri ile bir maddeye verdikleri belirli bir yanıtın olasılığı arasındaki ilişkiyi (sınava giren kişinin bir madde üzerindeki yeteneği ile performansı arasındaki ilişki) özel olarak tanımlayan bir matematiksel modeller ailesidir (Ayala, 2009; Cappelleri, Lundy ve Hays, 2014; Embretson ve Reise, 2000; Hays vd., 2009; Lord vd., 2008; Progar, Sočan ve Peč, 2008; Reise ve Waller, 2009; Reise, Widaman ve Pugh, 1993). Bu ilişkiyi ifade eden temel Madde Tepki Kuramı denklemi, Şekil 2.2.'de görsel olarak sunulan madde karakteristik fonksiyonu veya madde karakteristik eğrisi (ICC: item characteristic curve) olarak ifade edilmektedir (Baker ve Kim, 2004). IRT üretmeye çalıştığı eğrinin herhangi bir noktasını bulduğunda/kavradığında eğrinin geri kalan bölümünü tamamlayabilmektedir (Lord, 1980). Madde Tepki Kuramı'nda, madde karakteristik eğrisi (ICC: item characteristic curve); yaygın olarak her madde için bir yanıt eğrisi (belirli bir yeteneğe sahip bir öğrencinin soruyu doğru yanıtlama olasılığı) oluşturmak, bilinenlere dayalı olarak tüm test hakkında ölçekli bir puan elde etmek (Adams ve Wieman, 2011; Fotaris ve Mastoras, 2014), maddelere verilen yanıtları modellemek ve eğitsel testlerin puanlanması için kullanılmaktadır. Bu kapsamda IRT, sınav katılımcılarının yeteneklerinin ölçülmesinde, test maddelerinin seçiminde ve testleri eşitlemede kullanılan güçlü bir araçtır (Adedoyin ve Mokobi, 2013). Uygun IRT modeli aracılığıyla sınava giren bir kişinin yetenek düzeyi, bu yeteneği ölçen herhangi bir (alt) madde seti ile doğru bir şekilde tahmin edilebilmektedir.

IRT modellerinin özellikleri, ilk olarak Hambleton ve Swaminathan (1985) tarafından şu şekilde özetlenmiştir;

- Öncelikle bir IRT modeli, gözlemlenen tepki ile altta yatan gözlemlenemeyen yapı arasındaki ilişkiyi belirtmelidir.
- İkinci olarak, model yetenek puanlarını tahmin etmek için bir yol sağlamalıdır.
- Üçüncüsü, sınava giren kişinin puanları, altta yatan yapının tahmini için temel oluşturmalıdır.
- Son olarak, bir IRT modeli, sınava giren bir kişinin performansının bir veya daha fazla yetenekle tamamen tahmin edilebileceğini veya açıklanabileceğini varsaymaktadır.

Nering ve Ostini (2010), madde tepki kuramını "Gizli Özellik Teorisi", "Güçlü Gerçek Puan Teorisi" veya "Modern Zihinsel Test Teorisi" olarak adlandırmakta ve yetenekleri, tutumları veya diğer özellikleri ölçen testlerin, anketlerin ve benzer araçların tasarımı, analizi ve puanlaması için bir paradigma olarak görmektedir. Dolayısıyla söz konusu modellerden türetilen geçerlilik ve güvenilirlik tahminleri önemli etkilere sahiptir. Bu açıdan bakıldığında IRT puanlama yöntemleri ölçüm anlamında birçok yönden KTK'ya göre çok daha fazla avantaj barındıran özelliklere sahiptir. Reise, Ainsworth ve Haviland (2005) çalışmalarında IRT'nin avantajlarını; nitel çeşitliliğin ele alınması, bireysel farklılıkların ölçeklendirilmesi ve psikometrik analizler şeklinde üç örnek üzerinden açıklamaya çalışmışlardır. Bu kapsamda bir IRT çerçevesi altında, bilginin kesinliği bir bireyin özellik aralığı boyunca nereye düştüğüne bağlı olarak değişebilmektedir; buna karşılık KTK'da ise ölçek güvenilirliği (kesinliği) ham puan seviyelerine bakılmaksızın tüm bireyler için aynı kabul edilmektedir (Reise, Ainsworth ve Haviland, 2005, s. 95-96). Kavramsal olarak KTK'dan üstün olan IRT, test geliştiricileri için daha esnek ve daha çok bilgi barındıran zengin bir araç seti sağlama kabiliyetine sahiptir (Bichi vd., 2015; Eleje, Onah ve Abanobi, 2018). Çünkü IRT göstergelerin bir yapıyla olan ilişkisini ortaya koyarken klasik test teorisi "etki" gösterge modelinden daha iyi karşılık veren bir "nedensel" gösterge modeli sunmaktadır (Bollen ve Lennox, 1991). IRT modellerinden elde edilen tahminler, yeterli örneklem büyüklüğü sağlandığı taktirde (Cappelleri, 2014) klasik test teorisinin üzerinde ve ötesinde bilgi sağlayabilmektedir (Ferrando ve Chico, 2007).

IRT modelleri büyük ölçekli yetenek testlerinin uygulanma ve puanlanma biçimini derinden değiştirmiştir (Embretson ve Reise 2000). IRT puanlama, maddeler ile ilgilenilen yapılar arasındaki ilişkinin gücünü ve gizil yapı genelinde mevcut olan bilgiyi değerlendirmek için kullanılmaktadır (Hill vd., 2007, s. 39-47). IRT'de madde güçlüğü (madde zorluğu) ve sınava giren kişinin yeteneği ayrı ayrı tahmin edilerek bu parametreler aynı ölçekte değerlendirilebilmektedir (Hori, Fukuhara ve Yamada, 2020). IRT'nin bu özelliği hem öğrenenin yeteneği hem de madde güçlüğü (madde zorluğu) hakkıda bağımsız olarak bilgi edinilmesine (maddeden bağımsız kişi ölçümü ve kişiden bağımsız test kalibrasyonu) olanak sunmaktadır (Embreston ve Reise, 2000; Hambleton ve Swaminathan, 1985; Rupp ve Zumbo, 2006; Wright, 1968). Ayrıca IRT, test sonuçlarının genellenebilirliği, çeşitli madde analizleri, test yanlılığı ve diferansiyel madde işleyişini inceleme, test formlarını eşitleme, yapı parametrelerini tahmin etme, alan puanlaması ve uyarlanabilir test gibi birçok pratik test problemine uygulanabilmektedir (Ani, 2014). IRT aynı zamanda iki katılımcı grup arasındaki ayrımcılığın kapsamını, yani "farklı özellik seviyelerindeki bireyler arasında ayrım yapma"yı da açıklayabilmektedir (Morizot, Ainsworth ve Reise, 2007).

Madde Tepki Kuramı (IRT) yöntemleri, büyük test firmaları, devlet kurumları ve okul bölgeleri tarafından en önemli yetenek, başarı, yeterlilik, giriş ve profesyonel lisans sınavlarını oluşturmak, analiz etmek ve puanlamak için kullanılmaktadır (Resie ve Henson, 2003). Yakın zamanlarda IRT, klinik araştırmalarda sağlık sonuçları araştırmacıları tarafından da benimsenmiştir (Orlando-Edelen ve Reeve, 2007). IRT uygulamaları gelişim sürecini çoğunlukla büyük ölçekli bilişsel testler bağlamında (test düzenleme, analiz etme ve uygulamadaki pratik sorunları çözmeye hizmet etmeye yönelik olarak) sürdürmekle birlikte son zamanlarda kişilik, psikopatoloji ve sağlık sonuçlarının bireysel anlamda değerlendirilmesinde uygulanabilirliğini araştıran çalışmalar mevcuttur (Jabrayilov, Emons ve Sijtsma, 2016; Reise ve Haviland, 2005; Reise ve Waller, 2009).

Thorpe ve Favia (2012) ise IRT'yi bir ölçekte gözlemlenen madde yanıtları ile temeldeki bir yapı arasındaki bağlantıyı açıklamaya çalışan bir ölçüm modelleri topluluğu olarak tanımlamaktadır. IRT; 1PL (Rash Model), 2PL, 3PL ve 4PL modelleri (An ve Yung, 2014; Embretson ve Reise, 2000; Hambelton, 1994; Nunnally ve Berstein, 1994) aracılığıyla madde *ayırt ediciliği=a* (maddenin ölçülmeye çalışılan beceriye sahip

olanlarla olmayanları ayırt etme gücü), *madde güçlüğü=b* (bir maddenin uygulandığı grupta doğru yanıtlanma olasılığı), ve *şans başarısı=c* (sorunun şansla doğru yanıtlanma olasılığı) gibi parametrelere dair ölçümlere olanak tanımaktadır (Adedoyin ve Mokobi, 2013; Hambleton ve Linden, 1997; Hambleton, Swaminathan ve Rogers, 1991). Bir ölçüm teorisi olarak IRT, ölçme aracı geliştirme ve puanlamaya yardımcı olmakta ve parametre olarak adlandırılan yukarıda verilen a, b ve c parametre istatistiklerini tahmin etmeyi ve yorumlamayı mümkün kılmaktadır (Ani, 2014). Madde Tepki Kuramı, bir maddeden elde edilecek puana karar verirken çeşitli parametreleri (1PL = Rash Model, 2PL, 3PL ve 4PL) matematiksel olarak dikkate alan bir puanlama yöntemidir. Model, bağımsız değişkenlerin bağımlı değişkenleri en iyi şekilde tahmin etmek için birleştirildiği matematiksel bir denklemdir (Embretson ve Reise, 2000). Bu yöntemde bireyin performansına, sorunun ayırt ediciliğine, madde güçlüğüne ve şansla doğru yanıtlanma durumuna göre her soru başka puan almaktadır. Bu bölümde söz konusu parametrelerin (1PL: Rasch Model, 2PL, 3PL ve 4PL) her biri alt başlıklar halinde kısaca açıklanmaya çalışılmıştır.

2.4.2.1. 1PL (Rasch Model) IRT model

1PL IRT model diğer adıyla "Rasch Model", maddelerin ne kadar zor (madde güçlüğü) olduklarına göre ayırt edildiği en basit IRT modellemesidir (Keng, 2008). Rasch model veya 1-parametre (1PL) lojistik model; sınava giren kişinin yeteneğini maddelerin zorluğuyla (madde güçlüğüyle) ilişkilendirerek bir maddeye doğru yanıt verme olasılığını matematiksel olarak modellemeyi mümkün kılan bir IRT modellemesidir (Ockey, 2012, s. 339-340). Bu modelde her bir maddenin madde zorluk parametresi olarak bilinen yalnızca bir madde parametresi bulunmakta olup "b parametresi" değeri ile temsil edilmektedir. Zorluk parametresi (*b*), maddenin göreceli zorluğunu veya kolaylığını göstermektedir (Hambleton, Swaminathan ve Rogers, 1991). Madde zorluğu ya da güçlüğünü temsil eden (*b*) parametresi değeri, Şekil 2.2.'de de görsel olarak sunulduğu üzere madde karakteristik eğrisi (ICC) için bükülme noktasına karşılık gelen (yani eğimin maksimum olduğu nokta) yetenek ölçeği (θ: teta) değerine eşittir. Bu değer madde karakteristik eğrisi (ICC) yetenek ölçeğine göre konumunu belirlediği için bir konum parametresi olarak bilinmektedir. Bu nedenle, daha zor olan maddeler daha büyük "*b"* değerlerine sahiptir ve madde karakteristik eğrisi (ICC) yetenek ölçeğinde daha sağda yer

almaktadır. Zorluk parametresinin aralığı -∞ ile + ∞ arasındadır, ancak çoğu madde için "*b*" değerleri genel olarak -3 ile +3 arasında alınmaktadır (Baker ve Kim, 2004; Embretson ve Reise, 2000; Hambelton, 1994).

2.4.2.2. 2PL IRT model

2PL IRT model, maddelerin yalnızca zorluk açısından değil, aynı zamanda ayrım gücü (*a*: ayırt edicilik) açısından da değişmesine izin veren IRT modellemesidir (Hambleton, Swaminathan ve Rogers, 1991; Keng, 2008). Bu moldelleme madde zorluk parametresi (*b*)'ye ek olarak, madde ayırt ediciliğini belirtmek için bir "*a*" parametresi değeri içermektedir. Ayırt edicilik parametresi (*a*), bir maddenin ölçülen özellik için yetkin olanları (yani *θ: teta* değerleri yüksek olanları) daha az yetkin olanlardan ne kadar iyi ayırt edebileceğini tanımlamaktadır (Embretson ve Reise, 2000; Hambleton, Swaminathan ve Rogers, 1991). Değeri yüksek olan bir madde daha ayırt edicidir ve sınava girenleri farklı yetenek seviyelerine ayırmak için daha kullanışlıdır. Madde karakteristik eğrisi (ICC) açısından, *a*'nın değeri ICC'nin bükülme noktasındaki eğimi ile orantılıdır. Eğimi daha dik olan maddeler daha ayırt edicidir. Ayırt edicilik parametresinin teorik aralığı -∞ ile + ∞ aralığındadır (Obinne, 2008, s. 36-40). Bununla birlikte, negatif bir ayırt edicilik değeri genellikle maddeyle ilgili bir sorun olduğunu göstermekte ve ilgili maddenin atılması gerekmektedir. Bu nedenle, çoğu madde için a değerleri pozitiftir ve genellikle +2'den düşüktür (Hambelton, 1994; Hambleton, Swaminathan ve Rogers, 1991; Keng, 2008).

2.4.2.3. 3PL IRT model

3PL IRT modeli, madde ayırt ediciliği (*a*) ve zorluğa (*b*) ek olarak soruların şansla doğru yanıtlanma olasılığına karşın bir tahmin parametresi (*c*) de içeren modelleme türüdür. "c parametresi" değeri bilgisiz olarak sınava giren kişinin bile şans eseri bir maddeyi tahmin edebileceğini dolayısıyla ilgili soruyu doğru yanıtlayabileceğini kabul etmekte ve kendinden önceki modellemeleri (1PL ve 2PL) daha da genişletmektedir. Bu modelleme içeriğinde yer alan "a", "b" ve "c" parametrelerine yönelik açıklamalar aşağıda alt başlıklar halinde sunulmuştur.

2.5.2.3.1. a parametresi

a parametresi, madde karakteristik eğrisinin (ICC) dikliği ile grafik olarak ifade edilebilen bir ölçüdür. Madde ayırt ediciliğini (eğim) temsil eden "a parametresi", bir maddenin madde konumunun solundaki yeteneklere sahip katılımcıların madde konumunun sağındaki yeteneklere sahip olanlardan ne kadar iyi farklılaştığını göstermektedir (Adedoyin ve Mokobi, 2013; Thorpe ve Favia, 2012). Diğer bir ifadeyle "a parametresi", bir maddenin farklı yetenek seviyelerine sahip bireyler arasında ne kadar iyi farklılaşabileceğini ifade etmektedir. Yüksek ayırt edicilik düzeyi, maddenin düşük ve yüksek vasıflı bireyler arasında iyi bir ayrım yaptığını göstermektedir. a parametresi, ICC'nin dikliği ile grafiksel olarak ifade edilebilen bir ölçüdür. Bir ICC'nin eğimi ne kadar dik olursa bir maddenin ayırt edici değeri o kadar yüksek olmaktadır.

Madde ayırt ediciliği için değerlerin aralığını ve yorumunu Baker (2001, s. 42-44) şu şekilde ifade etmektedir;

- Hiç Yok (None): 0
- Çok Düşük (Very Low): .01 - .34
- Düşük (Low): .35 - .64
- Orta (Moderate): .65 - 1.34
- Yüksek (High): 1,35 - 1,69 ve
- Çok yüksek (Very High): 1.70 ve üzeri.
- Mükemmel (Perfect): +∞ (infinity)

2.5.2.3.2. b parametresi

Madde zorluğu (eşik) olarak da bilinen "b parametresi"; x ekseni boyunca bize bir maddenin ne kadar kolay veya ne kadar zor olduğunu söyleyen bir konum indeksidir. Bir maddenin konumunun indeksi, eğrinin y eksenindeki 0,5 olasılık değerini kestiği x ekseni üzerindeki noktadır. Negatif güçlük tahminleri maddelerin kolay olduğunu, pozitif güçlük tahminleri ise maddelerin zor olduğunu göstermektedir (Baker, 2001; Thorpe ve Favia, 2012). 1'den büyük "b" değerleri çok zor maddeleri ifade ederken -1'in altındaki "b" parametresi değerleri ise kolay maddeleri ifade etmektedir. "b" parametresi değerleri -0,5 ile 0,5 arasında olduğunda bu tür güçlük indekslerine sahip test maddeleri orta güçlükte maddeler olarak kabul edilmektedir (Adedoyin ve Mokobi'ye, 2013). Bir maddenin

zorluğu, "S" şeklindeki eğrinin en dik eğime sahip olduğu noktadır. Baker, (2001) pratikte zorluk değerlerinin genellikle -3 ile + 3 arasında olduğuna dikkat çekmiştir. Bir madde ne kadar zorsa, o maddeye doğru yanıt verebilmek için sınava giren kişinin yeteneğinin o kadar yüksek olması gerekmektedir. Bu noktada yüksek "b" parametresi değerine sahip maddeler zor maddelerdir, yani 1'den büyük "b" parametresi değerleri çok zor bir maddeyi temsil etmekte olup düşük yetenekli adayların bu maddeleri doğru yanıtlamaları olası değildir. Madde zorluğu ve öğrenci yeteneği arasında hassas bir eşleşme (Zenisky, Hambleton ve Luecht, 2009), verimli ölçüm sağlamakta ve öğrencilerin çok kolay veya çok zor olan test maddelerinden cesaretlerinin kırılmasını ya da sıkılmalarının önüne geçmektedir.

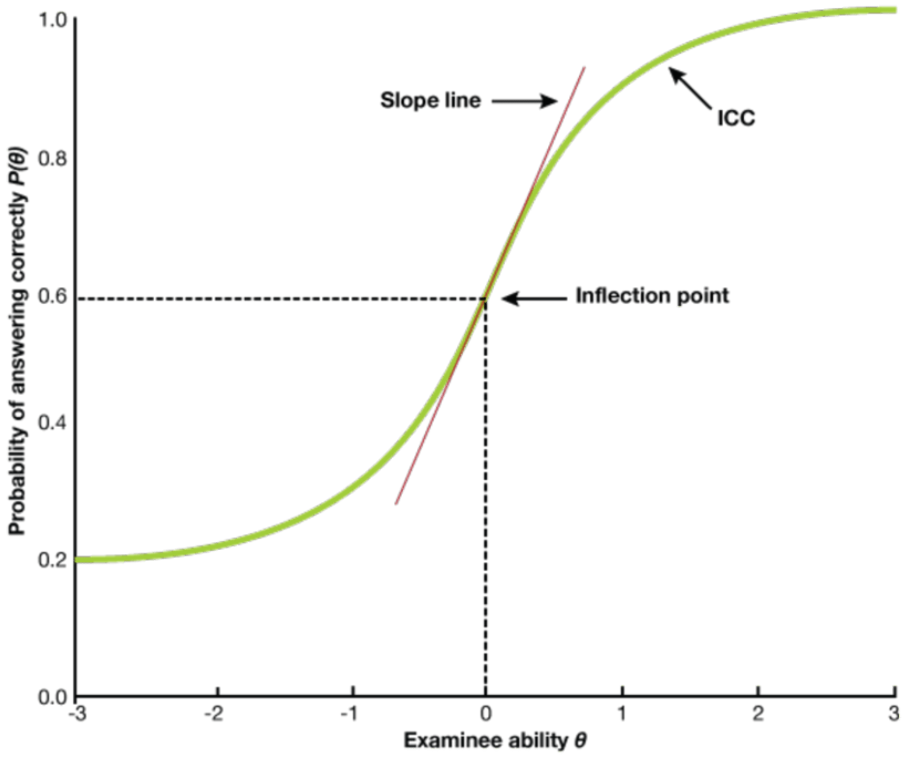
2.5.2.3.3. c parametresi

"c parametresi" değeri önceki bölümlerde "Şans Başarısı" başlığı altında kısaca açıklanmış olup bu bölümde detaylı olarak ele alınmıştır.

"Şans Başarısı" başlığı altında da ifade edildiği üzere tahmin parametresi olarak da bilinen şans başarısı IRT 3PL ve 4PL modelde "c parametresi" değeri ile temsil edilmektedir. Bu modeller; çok düşük yeteneğe sahip bir kişinin bir maddede doğru yanıtı tahmin edebilme olasılığını (Obinne, 2012) ve bu nedenle sıfırdan büyük bir doğru yanıtlama olasılığına sahip olma durumunu ifade eden "şans başarısı" olarak adlandırılan bir tahmin parametresi (c) içermektedir. Madde tahmin parametresi "c", bir ICC'nin ulaştığı en düşük değer olarak nitelendirilmektedir (Adedoyin ve Mokobi, 2013; Fotaris ve Mastoras, 2014).

Şans tahmin parametresi (*c*), diğer adıyla şans başarısı parametresi yetenek ölçeğinin alt ucunda yer alan sınav katılımcılarının performansını açıklamaktadır (Hambleton, Swaminathan ve Rogers, 1991; Keng, 2008). Bu tür katılımcılar sınava girdiklerinde maddeyi doğru yanıtlayacak kadar yetkin olmasalar bile tahmin yoluyla veya tesadüfen doğru maddeye ulaşabilmektedirler. Maddeler tahmin edilmelerinin ne kadar kolay olduğuna göre değişiklik göstermektedir. "*c*" parametresinin değeri 0 (tahmin mümkün değil) ila + 1 arasında değişebilmektedir. Burada 0 tahmini mümkün olmayan maddeleri ifade ederken +1 tahmin olasılığının yüzde yüz olduğu anlamında gelmektedir. Bu sebeple ilgili maddelerde söz konusu değerin 0'a yakın olması beklenmektedir. Şekil 2.4.'de de gösterildiği gibi "*c*" parametresinin değeri madde

karakteristik eğrisi (ICC)'nin düşük asimptotuna karşılık gelmektedir (Hambelton, 1994; Hambleton, Swaminathan ve Rogers, 1991). “a”, “b” ve “c” parametrelerinin her biri IRT 3PL model bağlamında madde karakteristik eğrisi ile ilişkili olarak şu şekilde tanımlanmaktadır: Şekil 2.4.’te belirtilen madde için zorluk parametresi (*b: difficulty*) = 0'a eşittir; ayırt edicilik parametresi (*a: discrimination*) = 1'e eşittir ve şans tahmin parametresi (*c: guessing*) = 0.2'dir.



Kaynak: Fotaris, P. ve Mastoras, T. (2014). LMS assessment: using IRT analysis to detect defective multiple-choice test items. *International Journal of Technology Enhanced Learning*, *6*(4), 281-296. Erişim adresi: https://www.academia.edu/download/38461842/IJTEL60401_Fotaris___Mastoras.pdf

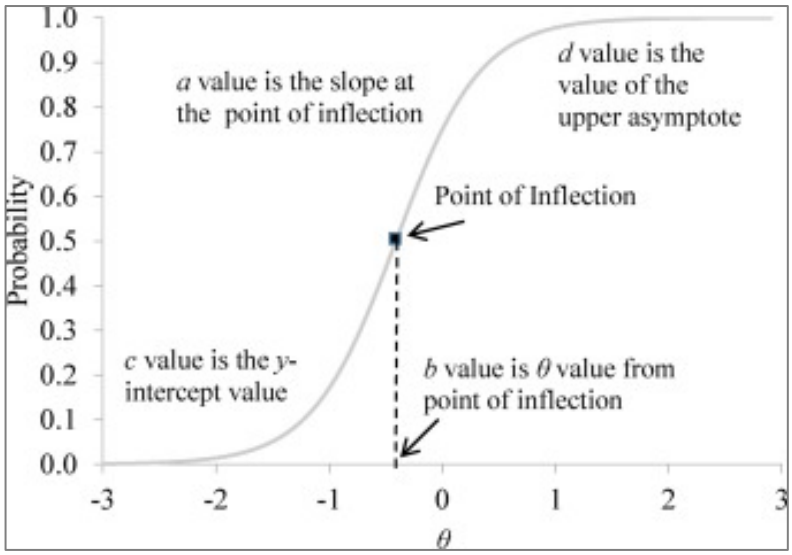
Şekil 2.5. *Madde Tepki Kuramı 3PL IRT model*

Lojistik denklemler Şekil 2.5.’de de görüldüğü gibi grafik haline getirildiğinde ICC olarak adlandırılan grafikler üretmektedir. Madde karakteristik eğrileri (ICC) çizildiğinde sınava giren kişinin yeteneği x ekseninde teta (θ) ile gösterilirken sınava giren bir kişinin soruyu doğru yanıtlama olasılığı y ekseninde P(θ) ile gösterilmekte ve ICC'ler tipik olarak S şeklinde bir eğrinin şeklini almaktadır. Yeteneğin en düşük seviyelerinde doğru yanıt

olasılığı 0 (sıfır)'a daha yakınken en yüksek seviyelerinde doğru yanıt olasılığı 1(bir)'e daha yakındır (Fotaris ve Mastoras, 2014; Hambleton vd., 1991). ICC'yi tanımlamak için madde güçlük ve madde ayırt edicilik değerleri olmak üzere iki teknik özellik kullanılmıştır. "b" ile gösterilen madde güçlük değeri, bir sınava girenin maddeyi %50 doğru bulma şansına sahip olması için gereken yeteneğe göre madde özellikleri eğrisinin konumunu gösteren bir konum parametresidir. "a" ile gösterilen madde ayrıt ediciliği ise bir maddenin yüksek ve düşük yetenek seviyelerine sahip kişileri ne kadar iyi ayırdığına dair bilgi sağlamaktadır.

2.4.2.4. 4PL IRT model

Barton ve Lord, (1981) küçük harf "*d"* ile ifade edilen bir üst asimptot parametresini 3PL modeline dahil ederek 4PL modelini ortaya çıkarmıştır. 4PL modelin tanımlayıcı özelliği, daha yaygın olan üç parametreli lojistik (3PL) modeline dahil edilenlere ek olarak "d" değeri ile temsil edilen bir üst asimptot parametresinin modele dahil edilmesidir (Du ve Kern, 2020). Bu model, kaygı ve dikkatsizlik nedeniyle yüksek yetenek düzeyine sahip adayların beklenmedik yanlış ya da eksik yanıtlarını açıklamaktadır (Barnard-Brak, Lan ve Yang, 2018; Kalkan ve Çuhadar, 2020). "*d"* parametresi değeri, en yüksek beceri seviyesine sahip olan katılımcıların yanıtlayamadığı soruları ve gizil beceri düzeyini tespit etmek için kullanılmaktadır. 4PL IRT modelinin temel özelliği, yüksek yetenekli yanıtlayıcılar için maddeye yanlış yanıt verme olasılığının sıfır (0) olmayan bir olasılığa izin vermesidir (Magis, 2013).



Kaynak: Barnard-Brak, L., Lan, W. Y. ve Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: A 4PL item response theory examination. *Studies in Educational Evaluation, 56*, 1-7. https://doi.org/10.1016/j.stueduc.2017.11.002

Şekil 2.6. *Madde Tepki Kuramı 4PL IRT model*

Yukarıdaki açıklamalardan da anlaşılacağı üzere IRT; maddeleri ve ölçekleri analiz etmek, psikolojik ölçümler oluşturmak ve yönetmek aynı zamanda psikolojik yapılar üzerindeki bireyleri ölçmek için kullanılan matematiksel modeller ve istatistiksel yöntemlerin bir koleksiyonudur. Bu çerçevede IRT'nin temel öğeleri madde yanıt fonksiyonları, bilgi fonksiyonları ve değişmezliktir (Reise, Ainsworth ve Haviland, 2005). Madde Tepki Kuramı'nda her bir maddeye verilecek puanlar KTK ile karşılaştırıldığında çok daha karmaşık bir matematik altyapısı kullanılarak hesaplandığı görülmektedir (Lord, 1980, s. 11-26). Madde Tepki Kuramı'nın bu yapısı yüksek oranda bilgi içermesi dolayısıyla düşük hata değerlerinde gerçeğe daha yakın hesaplamalar yapılmasına olanak tanımaktadır. Bu anlamda öğrenci başarısı gibi gizil değişkenler barındıran ölçülemeyen yapılar da IRT aracılığıyla hatadan arındırılarak daha düşük bir standart hatayla hesaplanabilmektedir (Betz ve Turner, 2011; Bock, 1997). IRT ve KTK arasındaki en önemli fark madde parametre tahminlerinin yorumlanmasından daha çok ölçüm hatasının kavramsallaştırılmasından kaynaklanmaktadır (Reise ve Henson, 2003). Daha önce belirtildiği üzere KTK tüm sınav katılımcıları için sabit bir standart hata verirken IRT ölçümün özelliklerine bağlı olarak gizli değişken sürekliliği boyunca ölçüm hatasının değişmesine izin vermektedir. Bu nedenle KTK'nın aksine IRT'de standart hatalar gruptan/örneklemden bağımsızdır ve birey bazlı değişmektedir (Yu, 2013).

IRT yöntemleri, çoğu diğer mevcut yöntemlerle karşılaştırıldığında istatistiksel hassasiyeti daha kuvvetlidir (An ve Yung, 2014; Kean ve Reilly, 2014; Nunnally ve Berstein, 1994). IRT (item response theory) tabanlı istatistiksel ölçümler, kesinlikte ve etkilerin tespit edilmesinde (Hambleton, Swaminathan ve Rogers, 1991) önemli avantajlar sunduğundan özellikle başarı testleri denemelerinin tasarlanmasında ve uygulanmasında birçok avantaj sağlayabilmektedir. Uyarlanabilir testin sıklıkla bahsedilen yararlarından biri yeteneği incelemek için testleri hedefleyerek (böylece test süresini kısaltarak) her bir sınava giren kişiye sunulan madde sayısı açısından testleri kısaltma fırsatıdır, ancak alan kapsamı ve ölçüm hassasiyeti ile ilgili hususlar yine de dengelenmelidir (Zenisky, Hambleton ve Luecht, 2009, s. 358-361). Dolayısıyla bir test bazı bireylerde daha çok hata verirken bazılarında daha az hata vermekte ve bu sayede az sayıda soru ile testi sonlandırmaya olanak tanımaktadır.

Ölçme ve değerlendirme alanında yaygın olarak bilinen PISA (Programme for International Student Assessment) ve NEAP (National Assessment of Educational Progress) gibi büyük ölçekli değerlendirmelerde IRT puanlama yöntemleri kullanılmaktadır (Hori, Fukuhara ve Yamada, 2020). Öğrenenlerin çeşitli konulardaki yeteneklerini ölçen ve izleyen bu büyük ölçekli sınavlarda IRT yönteminin tercih ediliyor olması modern ölçme ve değerlendirmenin (modern eğitim ölçümünün) önemli bir unsuru olduğunun göstergesidir.

IRT puanlama yöntemi CAT (Computer Adaptive Testing) ya da MST (Multistage Testing) gibi modern test sunum yöntemleri aracılığıyla uygulanabilmektdir. Ölçüm hassasiyetleri bakımından sınav uygulayıcıları tarafından sıklıkla tercih edilen CAT ve MST test sunum yöntemlerinin işleyişindeki temel algoritma şu şekildedir: CAT test sunum yönteminde maddeler, her bir sınav katılımcısının önceki maddelere verdiği yanıtlara göre altta yatan gizil yeteneğinin tahminindeki kesinliği hedefleyerek bu kesinliği (doğruluğu) en üst düzeye çıkaracak şekilde seçilmektedir (Rotou, 2007). MST test sunum yönteminde de maddelerin seçimi benzer algoritma ile gerçekleşmekte olup tek fark seçimlerin CAT'teki gibi madde bazlı değil modül bazlı ilerlemesidir. Tez sınırlılıkları kapsamında aşağıda MST (Multistage Testing) yöntemi detaylı biçimde sunulmuştur.

2.5. MST (Multistage Testing): Çok Aşamalı Test Tasarımları

Bilgisayar tarafından uygulanan testlerin "Bilgisayar Tabanlı Testler (CBT)", "Bilgisayarlı Uyarlamalı Testler (CAT)" ve "Çok Aşamalı Testler (MST)" olmak üzere üç ana tasarımı bulunmaktadır (Luecht, 2005; Zheng ve Chang, 2015). CBT, adaptasyon algoritması olmayan bilgisayar tarafından uygulanan lineer testleri ifade ederken (Mills vd., 2002), CAT ve MST'nin her ikisi de adaptif test sunum yöntemleridir. Bu bölümde tez çalışmasının ana temasını oluşturması sebebiyle MST test sunum yöntemine yönelik detaylı bilgiler sunulmuştur.

Çok aşamalı testler ilk olarak 1950'li ve 1960'lı yıllarda klasik test teorisi çerçevesinde öğrenenleri sıralamaktan çok sınıflama yönünde tasarlanmış bir test sunum yöntemidir (Angoff ve Huddleston, 1958; Cronbach ve Gleser, 1965; Linn, Rock ve Cleary, 1968). Tarihsel süreçte MST'nin bir tür bilgisayar olmadan geliştirilmiş versiyonunun uygulanması CAT (bireye uyarlanmış testlerden) uygulamalarından çok daha öncesine dayanmaktadır. Multistage yöntem soruları karma olarak "kolay", "orta" ve "zor" aşamalarda sunan bir test yapısı ile test sunumunda esnekliğe olanak tanımaktadır (Lord, 1971; Lord, 1980, s. 114-127). Bu yöntem MST test sunum tekniği kullanılarak test kitapçıkları aracılığıyla farklı bireylere aynı zorluk düzeyinde farklı soruların sunulabildiği sınavların bilgisayar olmadan uygulanmasıdır. Ancak bilgisayarların ve Madde Tepki Kuramı'nın CAT temelli kullanımının yaygınlaşması öncekilerden daha kısa ve daha fazla güvenirliğe sahip testler oluşturma potansiyelini arttırarak MST'yi gölgede bırakmış (Mead, 2006) bilgisayar uyarlamalı test (CAT) hem araştırma hem de uygulamada ön plana çıkmıştır. Bu sebeple MST yöntemi CAT kadar yaygın operasyonel kullanımda değildir (Macken-Ruiz, 2008). CAT, değerlendirmeyi daha verimli hale getirme potansiyelini yerine getirmiş olmakla beraber kullanıldıkça pratikte bazı eksiklikleri ortaya çıkmıştır. Özellikle test yapımı ve güvenlik sorunları, birçok kişinin CAT'in yararlarını yeniden düşünmesine yol açmıştır (Carlson, 2000, Chang ve Ying, 2007; Keng, 2008; Zheng ve Chang, 2015).

Uygulamada sınav, test ve ölçüm güvenliği açısından ölçüm hassasiyetinin yüksek olduğu kabul edilen en popüler test sunum yöntemi CAT olmakla birlikte son zamanlarda CAT'in eksikliklerini göz önünde bulunduran ve modern bir test sunum türü olan MST alternatif yöntem olarak sıklıkla gündeme gelmektedir (Hambleton ve Xing, 2006; Jodoin, Zenisky ve Hambleton, 2006; Luecht, 2005). MST'nin gündeme gelmesinde

katkı sağlayan en önemli konulardan biri ETS (Educational Testing Service) tarafından gerçekleştirilen GRE (Graduate Record Examination) ve GMAT (Graduate Management Admission Test) gibi büyük ölçekli sınavlarda CAT sisteminin eksikliklerinin tespit edilmiş olmasıdır (Carlson, 2000). Hali hazırda uygulanan çoğu CAT'in uluslararası nitelikte yüksek riskli sınavlar olması sebebiyle söz konusu sınavların güvenilirliklerinin artırılması aciliyet kazanmış ve bu durum MST'nin alternatif bir yöntem olarak gündeme gelmesine yol açmıştır. ETS gibi büyük ölçekli sınav uygulayan sistemlerin MST'ye geçiş yaparak bu test sunum yöntemini benimsemesi MST'nin popülerliğine ciddi anlamda katkı sağlamıştır (Zheng ve Chang, 2014, s. 21-22). Çok aşamalı test (MST), CAT'in eksikliklerini giderdiği iddia edilen alternatif bir uyarlanabilir test tasarımı (Luecht, 2005) olmasına rağmen MST üzerinde önemli ölçüde daha az araştırma yapılmıştır (Keng, 2008).

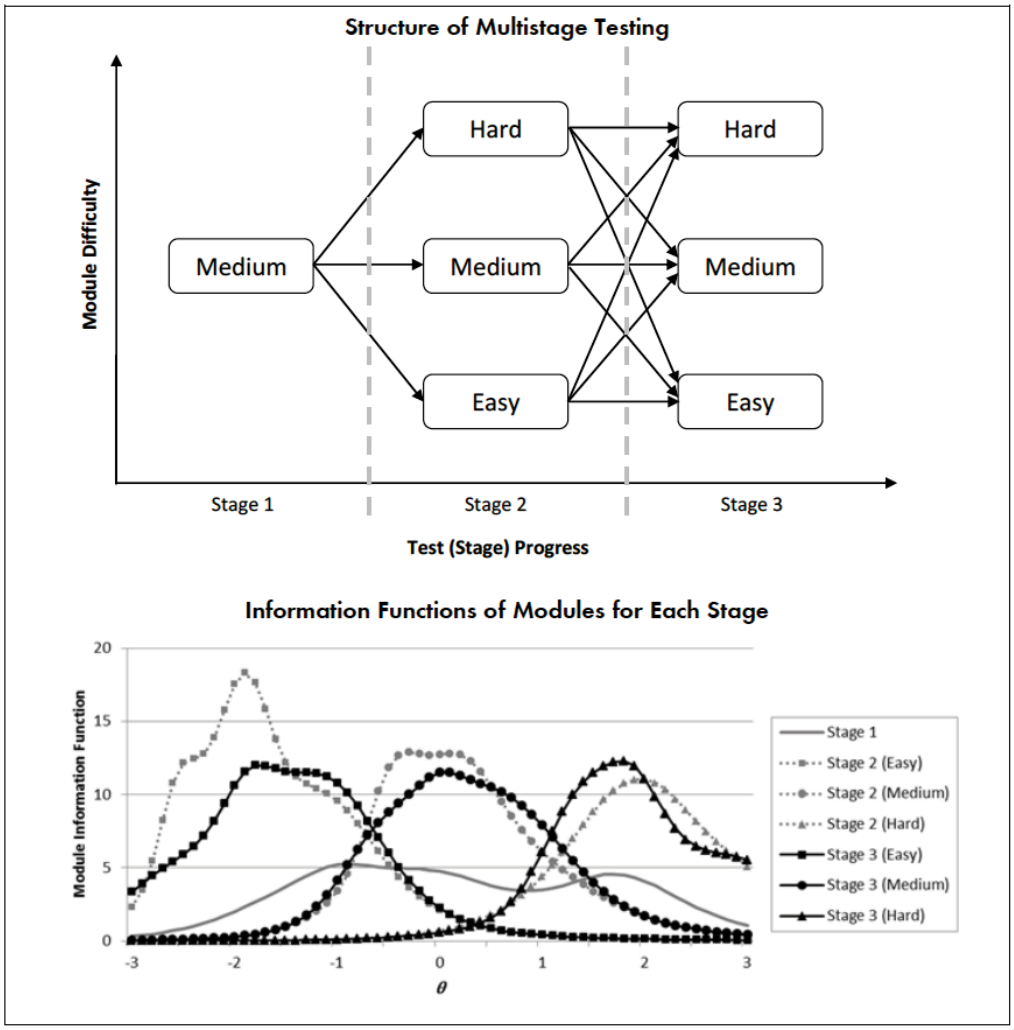
Çok aşamalı test veya MST, yukarda yer alan açıklamalarda da belirtildiği üzere madde setleri (yani modüller) düzeyinde bir test yönetmenin tercih edildiği uygulamalar için bilgisayarlı uyarlamalı teste (CAT) alternatif olarak geliştirilmiştir (Han, 2013; Luecht, 2005). MST test sunum yöntemi, özel bir bilgisayarlı uyarlamalı test (CAT) durumu gibi davranır ve önceden uygulanan test öğelerindeki performansına bağlı olarak her bir test katılımcısını önceden birleştirilmiş birkaç öğe grubundan birine uyarlamalı olarak yönlendirir. Aynı şekilde CAT yöntemide her aşamanın tek bir maddeden oluştuğu ve maddelerin belirli bir aşamaya bağlı olmadığı özel bir MST durumu olarak ifade edilebilir (Han ve Guo, 2014, s. 119-133). MST diğer adıyla çok aşamalı testler 2011 yılında "ETS" ve "GRE" sınavlarında kullanılmasının ardından popüler hale gelen bilgisayar uyarlamalı bir test sunma yöntemidir (Yan, von Davier ve Lewis, 2014; Sarı, 2020). Son yıllarda giderek daha popüler hale gelen MST, testin zorluğunun test katılımcısının yeterlilik düzeyine uyarlanmasına izin veren hassasiyette özel bir ölçme ve değerlendirme tasarımıdır (Magis, Yan ve von Davier, 2017; Sarı ve Huggins-Manley, 2017; Yang ve Reckase, 2020, s. 955-958). MST, CAT ve doğrusal testin bir karması olarak her iki tasarımın da özelliklerini içermesi sebebiyle hibrit teknolojiye sahip bir test sunma yöntemi sınıfında görülmektedir (Magis, Yan ve von Davier, 2017, s. 113-122). MST, teknoloji ilerledikçe ve ince ayarlara izin verdikçe test endüstrisinde talebi artan uygulamalar arasında yerini almaktadır. MST'nin göreceli yararlarının bireysel test programlarının özelliklerine, gereksinimlerine ve hedeflerine büyük ölçüde bağlı olduğu

farklı tasarım değişkenleri arasındaki ilişkisel etkiler tasvir edildikçe; MST'nin çeşitli alanlarda önemli değerlendirme göreviyle ilgilenen test ajansları için uygulanabilir bir alternatif olarak giderek daha önemli rol üstlenme potansiyeli mevcuttur (Luecht, 2005; Zenisky, Hambleton ve Luecht, 2009, s. 368-369).

MST yapısının test üretme yöntemi ilerleyen bölümlerde detaylı olarak verileceğinden genel hatları ile şu şekilde ifade edilebilir: MST'de tüm sınav katılımcılarına yönlendirme modülü veya birinci aşama testi olarak bilinen ortak bir madde seti uygulanır. Sınav katılımcısının performansına bağlı olarak sınav katılımcısı her biri sabit bir dizi maddeden oluşan ve ortalama zorluk derecesi farklı olan birkaç alternatif ikinci aşama testinden birine yönlendirilir. Sınav katılımcısı ikinci aşama testindeki performansına bağlı olarak, birkaç alternatif üçüncü aşama testinden birine yönlendirilir. Bu süreç, MST prosedüründeki aşama sayısına bağlı olarak devam etmektedir (Ariel, Veldkamp ve Breithaupt, 2006, s. 204-207; Han, Dimitrov ve Al-Mashary, 2019, s. 991-993; Mead, 2006, s. 185-187; Patsula,1999). Aşama sayısı ve aşama başına düşen modül sayısı, diğer faktörlerin yanı sıra, MST kullanan farklı test programları arasında değişiklik göstermekte olup istenen içerik ve ölçüm hassasiyeti kapsamından etkilenen bir test tasarımı kararıdır (Rotou, 2007; Zenisky, Hambleton ve Luecht, 2009).

2.5.1. MST (Multistage Testing) örüntüsü

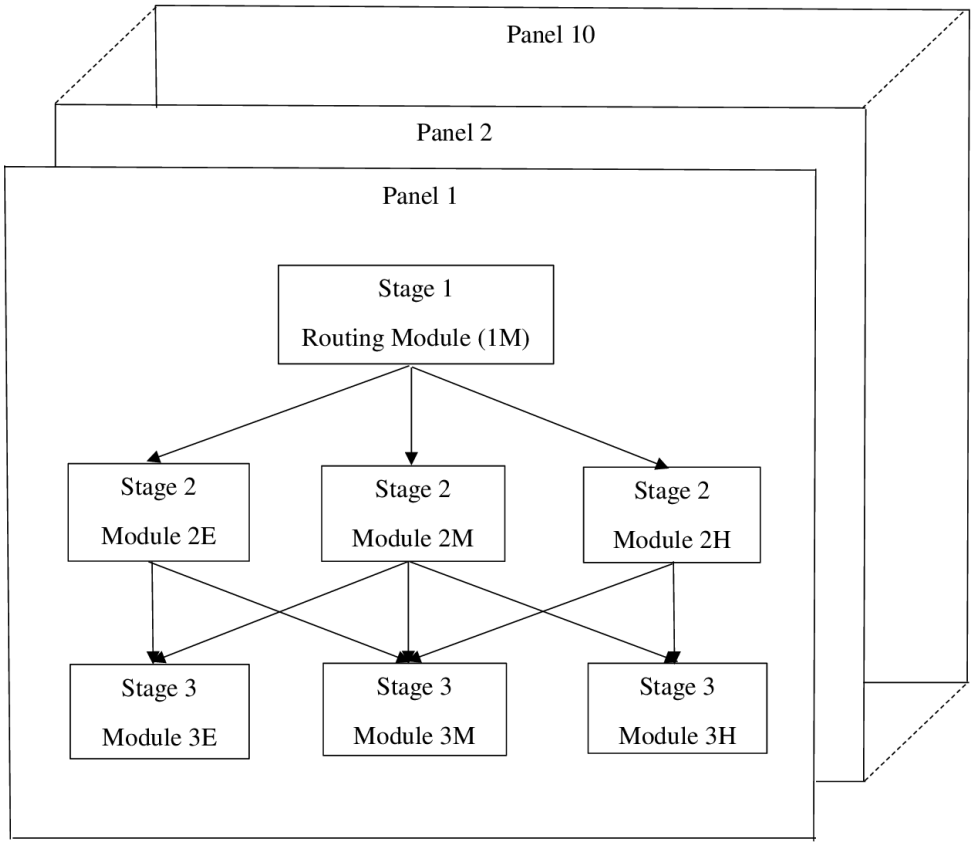
Alan yazında MST (Multistage Testing) olarak adlandırılan çok aşamalı testlerin yapısı "çeşitli potansiyel amaçlar için bağımsız, otomatik uyarlanabilir birimleri tasarlamanın, birleştirmenin ve yönetmenin modern bir yolu" olarak ifade edilmektedir (Han ve Guo, 2013; Yan, von Davier ve Lewis, 2014, s. 21-27). MST yapısı sıralı uygulama için CAT uygulamasında olduğu gibi tek bir madde yerine madde kümeleri üzerinde uyarlamalı seçimi içermektedir (Ariel, Veldkamp ve Breithaupt, 2006; Berger vd., 2019; Gierl, Lai ve Li, 2011; Han, 2013).



Kaynak: Han, K. T. ve Guo, F. (2013). An approach to assembling optimal multistage testing modules on the fly. GMAC Research Reports RR-13-01.

Şekil 2.7. *MST (Multistage Testing) modül örüntüsü tasarımı*

Şekil 2.7'de yer alan yapı MST örüntüsünün temel işleyiş mantığı görsel olarak genel hatları ile ifade etmekte olup panel sayılarını uygulayıcının tercihine göre arttırılıp azaltılabilme esnekliği sağlamaktadır. Testteki toplam madde sayısı, testteki toplam aşama sayısı ve aşama başına kaç madde dahil edileceği, çok aşamalı bir test geliştirme sürecinde ortaya çıkan ilk hususlardandır (Luecht ve Nungester, 1998; Zenisky, Hambleton ve Luecht, 2009, s. 358-361).



Kaynak: Wang, K. (2017). A fair comparison of the performance of computerized adaptive testing snd multistage adaptive testing (Doktora Tezi). ProQuest Dissertations & Theses Global veri tabanından erişildi (Order No. 10273809). Erişim adresi: https://www.proquest.com/dissertations-theses/fair-comparison-performance-computerized-adaptive/docview/1901897901/se-2

Şekil 2.8. *MST (Multistage Testing) panel yapısı*

MST uyarlanabilir olarak seçilen ve tek tek uygulanan maddelerden oluşan bir testten ziyade sınava girenlere madde kümelerini uyarlanabilir şekilde uygulayan önceden oluşturulmuş bir test panelidir (Wang, 2017). Şekil 2.8.'de söz konusu panellerin MST yapısı içerisindeki tasarım biçimi ana hatları ile görsel olarak ifade edilmeye çalışılmıştır. MST taksonomisi bir panel (test), bir panel içindeki birkaç aşama, bir aşama içindeki birkaç modül (madde seti) ve bir modül içindeki birkaç maddeden oluşmaktadır. Herhangi bir testten önce, yeteneği tahmin etmek için eşdeğer verimliliğe sahip çoklu paneller oluşturulmaktadır. Panel içinde, sırasıyla her biri tipik olarak iki veya daha fazla modül veya madde kümesi içeren iki veya daha fazla aşama bulunmaktadır. Her aşamada, MST modülleri genellikle kolaydan zora doğru genel bir zorluk sürekliliğini temsil edecek şekilde tasarlanmaktadır (Macken-Ruiz, 2008; Xu vd., 2021). Test süreci rastgele bir test paneli seçilerek başlamaktadır. Panel seçildikten sonra yetenek tahmini ve modül

(ürün seti) seçim süreçleri CAT'deki gibidir. Uygulanan ilk modül genellikle orta zorlukta olacak şekilde tasarlanmakta ve sınava girenleri varsayılan yetenek dağılımının ortalamasına göre konumlandırmaktadır. MST yönteminde yetenek tahmini CAT'den farklı olarak tüm modül uygulandıktan sonra tahmin gerçekleştirilmektedir. İlk aşama modülünün tamamlanmasından sonra yetenek tahminine bağlı olarak testin bir sonraki aşamasında uygun modül uyarlamalı olarak seçilmektedir (Macken-Ruiz, 2008). Multistage yöndemde panel sayısının artırılması, özellikle yüksek riskli testlerde test güvenliği ve ölçüm hassasiyeti amaçları için dikkate değer önemde bir konudur (Luecht, Brumfield ve Breithaupt, 2006; Yan, von Davier ve Lewis, 2014). MST, panellerde daha yüksek sayıda aşama ve aşamalı modül sayısı (madde seti) ile daha doğru yeterlilik diğer bir ifade ile daha doğru yetenek seviyesi tahminleri sağlanabilmektedir (Patsula, 1999; Zenisky, 2004).

2.5.2. MST (Multistage Testing) yapısı ve test üretme yöntemi

MST yapısının test üretme aşamasında sırasıyla şu işlem adımları takip edilmektedir; ilk aşamada katılımcıların yetenek seviyesini belirlemek için hazırlanmış olan "yönlendirme modülü" bulunmaktadır. Devamında ise her aşama "kolay", "orta" ve "zor" olmak üzere farklı yetenek seviyelerine ayrılmış modüllerden oluşmaktadır (Berger vd., 2019; Han ve Guo, 2013; Magis, Yan ve von Davier, 2017; Mead, 2006; Sarı, 2020; Yan, von Davier ve Lewis, 2014). Her aşamada bazı modüller düşük yetenek grubundaki katılımcılara daha uygunken bazı modüller yüksek yetenek grubundaki katılımcılara daha uygundur. Sınava giren kişi her bir madde setini tamamladıktan sonra yetenek tahmini, sınava giren kişinin yeteneği hakkında elde edilen yeni ölçüm bilgilerini yansıtacak şekilde güncellenir ve bir sonraki modül, bu hesaplanmış yetenek düzeyindeki birey için optimum düzeyde ölçüm bilgisi sağlamak üzere seçilmektedir (Zenisky, Hambleton ve Luecht, 2009). Bu sayede katılımcılar her aşamada yetenek düzeylerine en uygun modüle yönlendirilmektedir. Sınava giren bireylerin yetenek seviyelerini ölçmede her bir aşamada yer alan modüller anahtar konumundadır (Berger vd., 2019; Sarı, 2020; Sarı veHuggins-Manley, 2017). MST örüntülerinde ölçümler bireysel olarak sınava giren kişinin yetenek seviyesine odaklanarak ortalama beceriye sahip olanlar ve ölçüm ölçeğinin sonlarına yakın olanlar da dahil olmak üzere tüm sınav katılımcıları için hassas ölçümler sağlanmaktadır (Magis, Yan ve von Davier, 2017, s. 3-5). Bu sistemde testin yönünü öğrenenin verdiği yanıtlar belirlemekle birlikte puanını da yine öğrenenin verdiği

yanıtlar belirlemektedir. Ayrıca sınava giren kişiler madde havuzundan yetenek tahminleri ile eşleştirilen aynı beceriyi ölçen farklı türde sorular almaktadırlar (Sarı, 2020; Yan, von Davier ve Lewis, 2014, s. 87-94).

Çok aşamalı bir test; ikinci ve üçüncü aşamaların her birinde zorluk derecesine göre değişen (belirli sayılarda soru barındıran) iki veya üç modülle temsil edilmektedir. Bu anlamda Lord (1980) tarafından tartışılan konu sınava girenlere ilk aşama uygulandıktan sonra modüllerin her birinde ve sonraki her aşamada soruların sayısı ve göreceli zorluğudur. Bu modüller için birinciyi takip eden aşamalardaki tasarım süreci, test programı tarafından istenen yönlendirme hassasiyeti seviyesi, ürün bankasının derinliği, genişliği ve bu modüllerin ne kadar ayrık ya da örtüşebilir olması gerektiği dahil olmak üzere birkaç noktaya bağlı bulunmaktadır (Zenisky, Hambleton ve Luecht, 2009). Modül montajı için istatistiksel özellikler ise madde yanıt teorisi (IRT) parametreleri kullanılarak ifade edilmektedir (Ariel, Veldkamp ve Breithaupt, 2006).

Multistage yöntem "Routing" ve "Shaping" olmak üzere iki farklı yapıda test üretme yöntemi bulunmaktadır. Dolayısıyla MST katılımcılara iki farklı şekilde uygulanabilmekteidr. Bu yöntemlerin işleyiş biçimleri aşağıda sunulmuştur.

2.5.3. MST-R (Multistage Testing by Routing)

MST-R (Multistage Testing by Routing), test modüllerinin önceden monte edildiği ve kullanıcılar tarafından amaçlanan/hedeflenen aşamalara atandığı klasik MST test sunum yöntemini temsil etmektedir (Han, 2013; Jodoin, Zenisky ve Hambleton, 2006; Luecht, Brumfield ve Breithaupt, 2006; Luecht ve Nungester, 1998; Yan, von Davier ve Lewis, 2014, s. 411-420). Bu mod, sınava giren katılımcıların vermiş oldukları yanıtlara göre önceden birleştirilmiş birkaç test modülünden birine yönlendirildiği tipik geleneksel MST diğer adıyla çok aşamalı test yöntemidir (Luecht ve Nungester, 1998). MST-R yönteminde her aşamanın modülü önceden monte edilmiş olup bireyin performansına göre hazır yapı içerisinden sunulmaktadır (Han ve Guo, 2013; Luecht ve Nungester, 1998; Yan, von Davier ve Lewis, 2014). Bu yöntemde ilk aşama uygulandıktan sonra katılımcının yetenek düzeyine göre ara yeterlilik puanları hesaplanır ve katılımcılar ara puan tahminlerine göre bir sonraki aşama için önceden belirlenmiş test modüllerinden birine yönlendirilir (Han, 2013; Han ve Guo, 2013; Jodoin, Zenisky ve Hambleton, 2006).

Araç aynı zamanda kullanıcılara test maddesi maruziyet kontrolü için birden fazla paralel modül seçenekleri sunmaktadır (Luecht, Brumfield ve Breithaupt, 2006; Yan, von Davier ve Lewis, 2014, s. 21-38).

2.5.4. MST-S (Multistage Testing by Shaping)

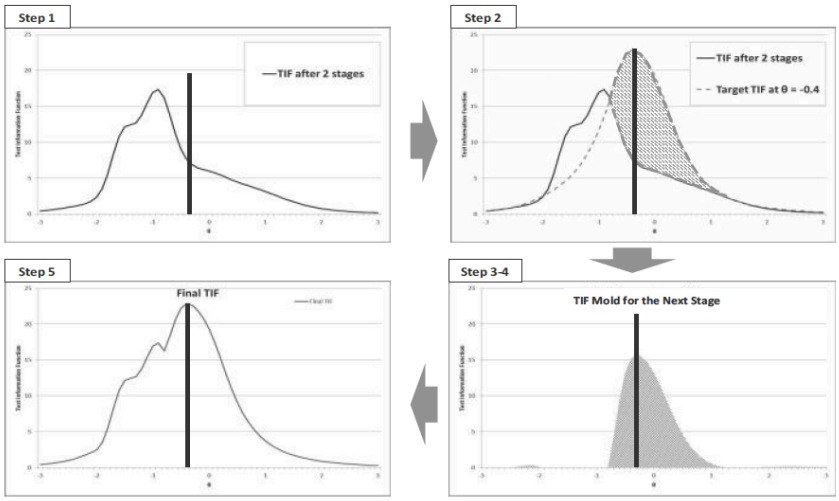
MST-S (Multistage Testing by Shaping), test modülünü testin her aşamasından sonra anlık olarak monte edilen yeni tasarlanmış bir test modülüyle değiştiren yeni bir uyarlamalı çok aşamalı test yöntemidir (Han ve Guo, 2014, s.119-133). MST-S (Multistage Testing by Shaping), MST-R (Multistage Testing by Routing)'nin aksine test modüllerinin montajının anlık olarak katılımcının performansına göre şekillendirildiği MST test sunum yöntemini temsil etmektedir (Han, 2013; Yan, von Davier ve Lewis, 2014, s. 411-420). Han (2013) tarafından önerilen MST-S yaklaşımına göre her aşama için bir madde modülü otomatik olarak hedef test bilgi fonksiyonuna (TIF: The Target Test Information Function) göre anında şekillendirilerek monte edilmektedir. Örneğin; birtakım komutları öncesinde verip seçimler yapması sağlandıktan sonra devamındaki seçimler sistemin kendisine bırakılabilmektedir. Çünkü MST-S yaklaşımı, modül şekillendirme süreci sırasında aynı zamanda madde görüntülenme kontrolü ve kapsam dengeleme işlemlerini gerçekleştirebilmektedir (Han ve Guo, 2013). MST-S yazılımı farklı madde işleyişini ve madde parametresi kayma (Item Parameter Drift "IPD") koşullarını simüle edebilir ve ön test madderini yönetebilir durumdadır (Han, 2013; Yan, von Davier ve Lewis 2014, s. 412).

MST-S yöntemi önceden birleştirilmiş test modülleri seçmek yerine bir sonraki aşama için yeni madde modülünü birleştirmek üzere aşağıda belirtilen adımları kullanarak her aşamadan sonra anlık şekillenen bir test modülü oluşturmaktadır (Han ve Guo, 2014, s. 122-125):

- İlk olarak ara bir θ (theta) tahmin edilir ve uygulanan maddelere dayalı olarak TIF değerlendirilir.
- Ara θ'daki bir sonraki aşama için mevcut TIF ile hedef TIF (test geliştiricileri tarafından önceden belirlenir) arasındaki fark değerlendirilir.
- 2. adıma dayalı olarak bir sonraki öğe modülünün (önceden uygulanan modüller hariç) bilgi işlevi için ideal bir şekli tanımlayan ve yeni bir terim olan bir TIF kalıbı oluşturulur.

- 3. adımdaki kalıba göre bir öğe modülü şekillendirilir.
- Adım 4'te şekillendirilen öğe modülü yönetilir
- Son aşama bitene kadar 1-5 arası adımlar tekrarlanır.

Aşağıda yer alan Şekil 2.9'da bu beş adım ana hatları ile görsel olarak ifade edilmektedir.



Kaynak: Han, K. T. ve Guo, F. (2014). Multistage testing by shaping modules on the fly. D. Yan, A. A. von Davier ve C. Lewis (Ed.), *Computerized multistage testing: Theory and applications* (s. 119-133) içinde. Boca Raton/London/New York: CRC Press Taylor&Francis Group.

Şekil 2. 9. *MST-S (Multistage Testing by Shaping) işleyiş şeması*

Bu yapının en dikkate değer özelliği modüllerin bireyin performansına göre anlık olarak sınav esnasında oluşturuluyor olmasıdır (Han ve Guo, 2013; Yan, von Davier ve Lewis, 2014). MST-S test sunum yönteminde sorular bireyin yetenek düzeyine göre anlık olarak atanması durumu aynı yetenek düzeyine sahip bireylerin bile farklı sorularla karşılaşması anlamına gelmektedir. MST-S'nin bu yönü testin ölçüm hassasiyetini artırmasının yanı sıra ölçme ve değerlendirme alanının önde gelen sorunsallarından biri olan test güvenirliğine önemli derecede katkı sağlamaktadır. Buradaki bir diğer önemli husus ise "Shaping" yönteminin bireyin yeteneğine göre anlık soru atama özelliğinin (Zheng ve Chang, 2014) CAT ile MST arasındaki sınav rotasının oluştuturulmasına yönelik bu kapsamdaki temel farklılığı ortadan kaldırmasıdır.

2.5.5. MST (Multistage Testing)'nin diğer bilgisayarlı testlerden farkı

MST test sunum yönteminin diğer bilgisayarlı test sunum yöntemlerinden farkı bu başlık atında CAT örneği üzerinden açıklanmaya çalışılmıştır.

Çok aşamalı test (MST) tasarımı, otomatikleştirilmiş (çevrimiçi) testlerin uygulanması için alternatif bir tasarımdır (Macken-Ruiz, 2008). MST diğer adıyla çok aşamalı testler madde setleri düzeyinde bir test yönetmenin tercih edildiği uygulamalar için CAT'lere alternatif olarak geliştirilmiştir. MST, CAT formunda adaptif testlere benzer yapıda olmakla birlikte CAT'in yaptığı gibi tek tek maddeleri sunmak yerine modül bazında bir dizi maddeyi sunmaktadır (Sarı, 2020; Sarı, Yahsı-Sarı ve Huggins-Manley, 2016; Yan, von Davier ve Lewis, 2014). Diğer bir ifadeyle CAT tek maddeler arasında uyum sağlarken MST yalnızca aşamalar (yani madde blokları) arasında uyum sağlamaktadır (Zheng ve Chang, 2014, 21-22). MST'de önceden oluşturulmuş madde setleri (ya da modüller) aşamalara yerleştirildikten sonra panellere yerleştirilmektedir (Han, 2013; Sarı 2016; Yan, von Davier ve Lewis, 2014). MST örüntüsü içinde aynı anda uygulanan sınav katılımcılarının sayısı, zaman aralığı ve bir test sunucusu ile istemci bilgisayarlar arasındaki iletişim sıklığı (yani terminaller) test geliştiricilerin tercihine göre koşullandırılabilmektedir (Han, 2013, s. 666-668). Hedeflenen test tasarımları, öğrencileri eşleşen test formlarına atamak için yeteneklerle ilgili arka plan değişkenlerini de göz önünde bulundurarak işlem yapmaktadır. Ayrıca çok aşamalı test tasarımları, öğrencileri en bilgilendirici modüllere yönlendirmek için sınav sırasında öğrencilerin performansını dikkate almaktadır (Berger vd., 2019). MST algoritması ile hazırlanmış bir test sunum yönteminde sınava giren her bir katılımcı tarafından aşama başına yalnızca bir test parçası görülebilmekte ve madde havuzundaki tüm sorular tek seferde sunulmadan güvenli bir biçimde test sunumu gerçekleştirilmektedir. Çok aşamalı test tasarımı yüksek başarılı ve düşük başarılı alt grupların geleneksel doğrusal formlardan daha doğru bir şekilde ölçülmesine yardımcı olması sebebiyle güçlü bir test sunumu modeli olarak görülmektedir (Yan, von Davier ve Lewis, 2014). Çok aşamalı bir test, ölçüm hassasiyeti ve doğruluğu açısından bir CAT'e daha çok benzeyecek şekilde oluşturulabilirse, tasarımın uyarlanabilirlik, pratiklik, ölçüm doğruluğu ve test formları üzerindeki kontrol arasında bir denge kurması nedeniyle tercih edilebilmektedir (Zenisky, Hambleton ve Luecht, 2009).

Multistage yöntemin CAT'e alternetif bir test sunumu olarak gündeme gelmiş olması sebebiyle her iki yöntemin özellikle avantaj ve dezavantaj bağlamında karşılaştırıldığı durumlar araştırmalara konu olmaktadır:

CAT'lerin ölçüm verimliliği, daha kısa test uzunluğu, sınava girenler için esnek sınav programı, test kopyalamayı önleme gibi önemli avantajlarının yanı sıra uygulamasının karmaşıklığı, kalibrasyonunun büyük bir veri seti gerektirmesi, test geliştirme için büyük çabalara gereksinim duyması, madde teşhir kontrolünün daha zor olması, bilgisayar aracılığıyla yönetiminin maliyetli olması ve güvenlik endişeleri gibi bir takım dezavantajlarının gündeme gelmesiyle MST yöntemlerine odaklanılmaya başlanmıştır (Ockey, 2012, s. 347; Rotou, 2007; Weiss ve Kingsbury, 1984; Yan, von Davier ve Lewis, 2014, s. 19). MST'nin özellikle kısa uzunluktaki testlerle ölçüm verimliliğini ve test optimizasyonunu hassasiyetle iyileştirmesi pratik bir avantaj olarak değerlendirilmektedir (Han, 2020, s. 87-96). MST'nin bilgisayar tabanlı doğası, yeni madde biçimleri, ölçülebilen yeni beceri türleri, daha kolay ve daha hızlı veri analizi, madde yanıt süresi gibi zengin davranış verilerini toplamada da birçok avantaj sağlamaktadır (Wang, Zheng ve Chang, 2014). MST'nin bir diğer avantajı da doğrusal testlerin gerektirdiğinden daha az madde kullanarak daha doğru yetenek (θ: theta) tahminleri sağlayabilmesidir (Wang, Chen ve Jiang, 2020). Reese, Schnipke ve Luebke (1999)'in testlerin optimal bir şekilde birleştirilmesi için stratejilere odaklanan çalışmaları; dikkatlice oluşturulmuş ve içerik dengeli iki aşamalı bir testin, CAT ve mevcut kağıt ve kalem testinden daha iyi performans gösterdiğini ortaya koymuştur. MST'de bir test uygulamasında ("Testlet (sets of test question): MST" yaklaşımı) madde bazlı ölçüm yerine maddeler kümeler halinde birleştirilerek modüller oluşturulur ve yönetilir (Ariel, Veldkamp ve Breithaupt, 2006). Bu yapıda hazırlanmış bir test sunumu kâğıt ve kalem testlerinin bazı avantajları (örneğin; içerik özelliklerine göre madde oluşturma ve madde kullanımını kontrol etme fırsatı vb.) ile CAT yönteminin bazı avantajlarını bünyesinde barındırmaktadır. Bu sayede kültürel anlamda daha anlaşılır bir test sunumu ile test içeriğinin güvenliğini kontrol altına almaya olanak tanıyabilen bir sınav uygulaması gerçekleştirmek mümkün görünmektedir.

Bilgisayarlı uyarlamalı testler (CAT), 1970'li yılların başından beri bilgisayar üzerinden gerçekleştirilen test uygulamalarına hâkim yöntem olsa da MST yöntemlerine olan ilgi artma eğilimindedir. Test durumuna bağlı olarak bir grup sabit öğeyi ayrı ayrı yönetmek yerine tek seferde yönetmenin test yönetiminde bazı avantajları bulunmaktadır (Han ve Guo, 2014, s. 119-122; Hendrickson, 2007; Luecht, 2003; Macken-Ruiz, 2008; Raborn ve Sarı, 2021; Yan, Lewis ve Davier, 2017, s. 3-20):

- Uygulanmasının ve montajının daha kolay olması
- Test geliştirmek için CAT yöntemine göre daha düşük düzeyde çaba gerektirmesi
- Test maddeleri modül setlerinden oluşmaktadır.
- MST'de aşamaların yapısı, modüllerin yerleşimi ve her modül içindeki öğelerin bileşimi test uygulamasından önce belirlendiğinden MST test spesifikasyonlarının ve özelliklerinin ayrıntıları üzerinde CAT'den daha fazla kontrole izin vermektedir.
- MST'nin CAT'e göre bir diğer avantajı da madde seçim sürecinde istemci bilgisayarlara daha az yük getirmesidir. MST ile, bir istemci bilgisayarın her madde yerine her aşamadan sonra yalnızca geçici yeterlilik tahminlerini hesaplaması gerekir. Yüzlerce ayrı madde arasından seçim yapmak yerine yalnızca bir avuç madde modülünü dikkate aldığından dolayı seçim algoritması için hesaplama iş yükü MST'de çok daha basittir.
- Daha da önemlisi, sınava girenler genellikle MST'yi tercih ederler çünkü her zaman olmasa da genellikle, sınava girenlerin yanıtlarını gönderdikten sonra geri gitmelerini engelleyen CAT'in aksine, MST sınav katılımcılarının maddeler arasında ileri geri hareket etmelerine ve her bir modül içinde ilk yanıtlarını değiştirmelerine izin verir. Yani bir modül içinde sınava giren kişiler test öğelerinde ileri ve geri atlayabilir ya da daha önce yanıtlanan maddelerde değişiklik yapabilir. Bir aşama tamamlandığında, kişinin bir önceki aşamaya dönmesi engellenmektedir.

MST’nin CAT yöntemi ile eş değer avantajları; ölçüm verimliliği, sınava girenler için esnek sınav programı ve test kopyalamayı azaltması gibi özellikleri sıralanırken MST’nin dezavantajları; model varsayımlarına bağlı olması, CAT'den daha uzun ancak lineer (doğrusal) testten daha kısa testlerle uygulanabilmesi, CAT'e benzer madde teşhir endişeleri ve bilgisayar yoluyla yönetiminin maliyetli olması (CAT'den fazla değil) hususları belirtilmektedir (Yan, Lewis ve Davier, 2014).

3. YÖNTEM

Bu bölümde araştırma deseni, simülasyon modeli, verilerin üretilmesi ve verilerin analizi aşamaları ele alınmaktadır.

3.1. Araştırma Deseni

Bu tez çalışmasında nicel araştırma yöntemi kullanılarak MSTGen yazılımı aracılığıyla elde edilen veriler KTK ve IRT puanlama yöntemleri ile MST test sunma tekniklerinin her biri için farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) ayrı ayrı sınanarak en az hata ile gerçek değere en yakın doğrulukta kestirim sağlayan yöntem araştırma soruları dahilinde simülatif olarak tespit edilmeye çalışılmıştır. Simülasyon yöntemi hem basit rastgele örneklemlerde hem de daha gerçekçi çok aşamalı örneklemlerde tahmin edicilerin tutarlılığını destekler nitelikte bir uygulama (Cohen vd., 2008) olması sebebiyle tercih edilmiştir. Araştırma kapsamında simülatif bir ortamda farklı açılardan test edilen yöntem ve teknikler aracılığıyla açık ve uzaktan öğrenme sistemi sınavlarında/çevrimiçi sınavlarda uygulanabilecek optimum algoritmanın minimum emek, zaman ve maliyet ile keşfedilmesi hedeflenmektedir.

3.2. Simülasyon Modeli

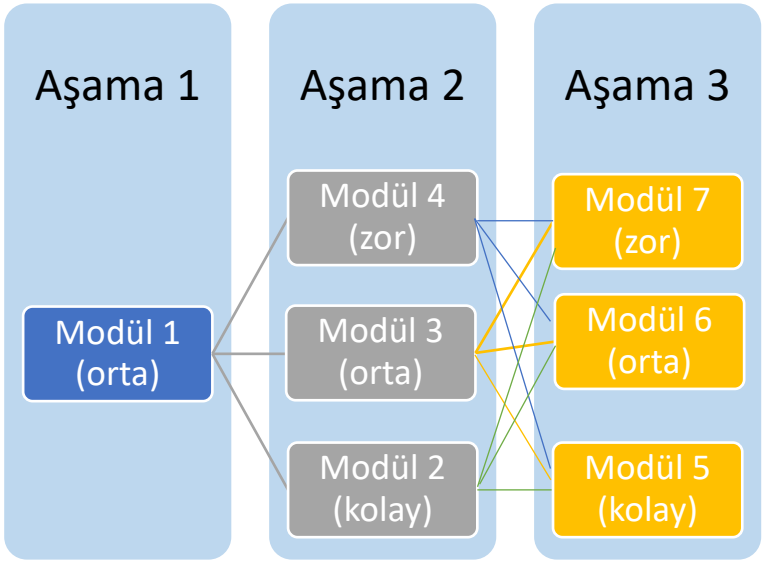
Araştırmanın bu bölümünde simülatif verilerin hangi yöntemle, hangi yazılım aracılığıyla ve hangi koşullar altında üretildiği sunulmuştur.

3.2.1. MST (Multistage Testing) test sunum yapısının simülatif modül ve panel montajı

"MST (Multistage Testing) test sunum yapısının simülatif modül ve panel montajı" başlığı altında modüller ve panellerin hangi sistematik çerçevede oluşturulduğu/monte edildiği açıklanmaktadır:

MST test sunum yöntemi temel alınarak MSTGen yazılımı aracılığıyla 3 panel ve 7 modülden oluşan 70 maddelik simülatif bir soru havuzu oluşturulmuştur. Panel montajları ilk panelde orta (medium) zorlukta maddeleri barındıracak tek bir modül, ikinci ve üçüncü panellerde ise kolay (easy), orta (medium) ve zor (hard) düzeylerde maddeleri barındıracak ve üçer modül olacak şekilde tasarlanarak test montajları gerçekleştirilmiştir.

MST test sunum yapısının oluşturulmasında 1-3-3 çok aşamalı test deseni kullanılmıştır. 1-3-3 deseni sırasıyla; ilk aşamada bir modül, ikinci aşamada üç modül ve üçüncü aşamada üç modül olarak test montajının gerçekleştirildiğini ifade etmektedir (Berger vd., 2019; Hambleton ve Xing, 2006; Jodoini Zenisky ve Hambleton, 2006; Keng, 2008; Luecht ve Nungester, 1998; Luecht, Brumfield ve Breithaupt, 2006; Patsula, 1999; Yan, von Davier ve Lewis, 2014).



Şekil 3. 1. *MST (Multistage Testing) test sunum yapısının simülatif modül ve panel montajı*

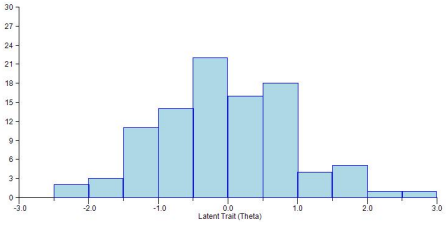
Şekil 3.1.'de gösterildiği gibi MST test sunumu öncelikle orta zorluktaki maddeleri içeren ve uygulama modülü olarak adlandırılan "Modül 1" ile başlamaktadır. İlk aşamada göstermiş olduğu performansa göre bireyin ikinci aşamadaki test rotası belirlenerek "Modül 2", "Modül 3" ve "Modül 4"ten birine yönlendirilmektedir. Bu süreç üçüncü aşama içinde aynı şekilde "Modül 5", "Modül 6" ve "Modül 7"den birine yönlendirilerek devam etmektedir. Her bir veri seti bu aşamalar takip edilerek hem MST-R hem de MST-S yöntemi ile Şekil 3.1.'deki test sunum yapısı çerçevesinde oluşturulmuştur.

3.3. Araştırma Verilerinin Üretilmesi

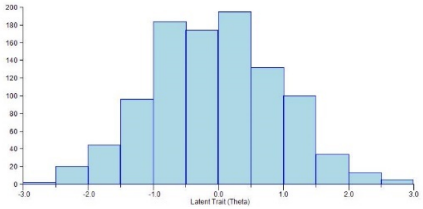
Günümüz koşullarında bilgisayar teknolojilerinde yaşanan ilerlemeler pek çok disiplinde olduğu gibi ölçme ve değerlendirme alanında da simülasyon uygulamalarını problemleri tespit etme ve çözmeye ilişkin olarak formal ve popüler bir araştırma yöntemi

konumuna getirmiştir (Wang, Zheng ve Chang, 2014; Yang ve Reckase, 2020). Araştırma verileri farklı algoritma biçimlerinde veri üretme konusunda alan yazında önemli bir yere sahip MSTGen (Han, 2013) yazılımı kullanılarak elde edilmiştir. Araştırma verilerinin üretilmesinde MSTGen yazılımının kullanılmasının sebebi; farklı koşullara göre gerçek veri üretebilmesi ve elde edilen verinin pek çok program için betik (syntax) dosyası görevi görebilmesinin yanı sıra çıktıların bu programlarda okunmasına olanak sağlayan ve kullanım kolaylığı bulunan bir yapıda olmasıdır. "MSTGen" yazılımı aracılığıyla MST koşullarına uygun formda üretilen veriler ile farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) ve farklı yöntemlerle (KTK, IRT, MST) bir simülasyon çalışması gerçekleştirilmiştir.

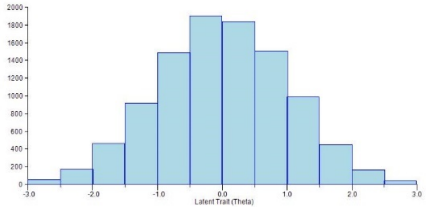
Araştırma kapsamında tasarlanan MST simülasyonları 100, 1000 ve 10000 kişilik farklı örneklem gruplarına ve farklı dağılımlara sahip olmak üzere toplamda 18 ayrı veri seti olarak gerçekleştirilmiştir. Bu veri setlerinden 9 tanesi MST-R yöntemi ile diğer 9'u ise MST-S yöntemi ile farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) oluşturulmuştur.



100 Kişilik Örneklem Grubu Dağılımı



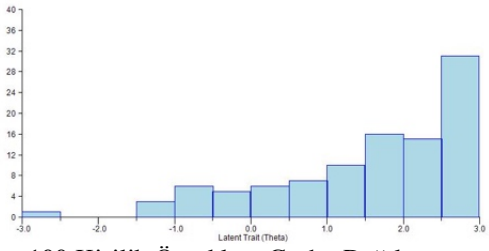
1000 Kişilik Örneklem Grubu Dağılımı



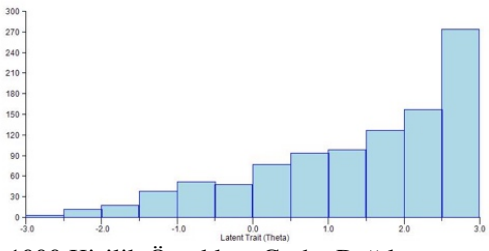
10000 Kişilik Örneklem Grubu Dağılımı

Şekil 3.2. *Normal dağılma sahip örneklem grupları*

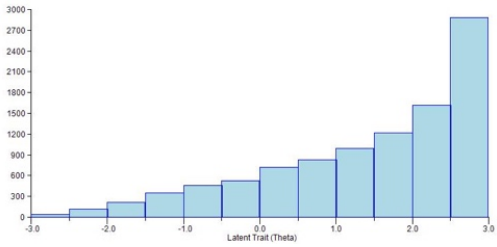
Şekil 3.2.'de örnek olarak görüleceği gibi MST-R ve MST-S test sunum yöntemleri kullanılarak her bir test sunum yöntemi için normal dağılıma sahip 100, 1000 ve 10000 kişilik örneklem grubundan oluşan veri setleri oluşturulmuştur.



100 Kişilik Örneklem Grubu Dağılımı



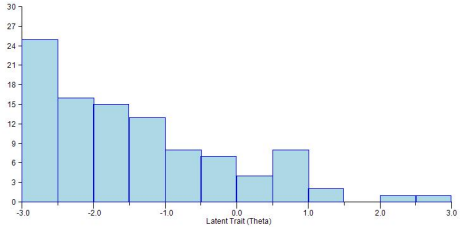
1000 Kişilik Örneklem Grubu Dağılımı



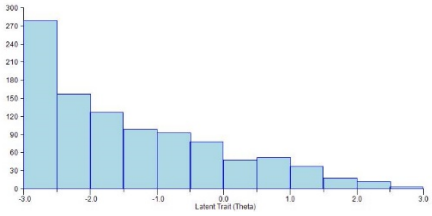
10000 Kişilik Örneklem Grubu Dağılımı

Şekil 3.3. *Normal olmayan (sola çarpık) dağılma sahip örneklem grupları*

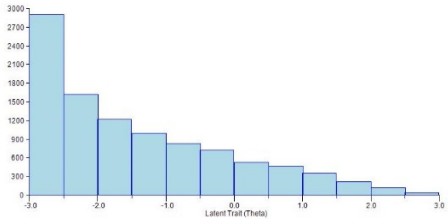
Şekil 3.3.' te de MST-R ve MST-S test sunum yöntemleri kullanılarak her bir test sunum yöntemi için normal olmayan (sola çarpık) dağılıma sahip 100, 1000 ve 10000 kişilik örneklem grubundan oluşan veri setleri oluşturulmuştur.



100 Kişilik Örneklem Grubu Dağılımı



1000 Kişilik Örneklem Grubu Dağılımı



10000 Kişilik Örneklem Grubu Dağılımı

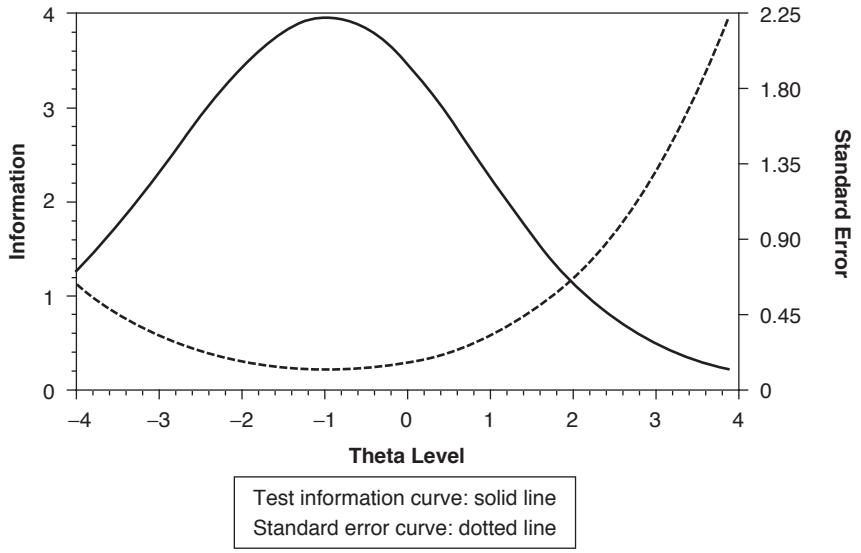
Şekil 3.4. *Normal olmayan (sağa çarpık) dağılma sahip örneklem grupları*

Şekil 3.4. ise MST-R ve MST-S test sunum yöntemleri kullanılarak her bir test sunum yöntemi için normal olmayan (sağa çarpık) dağılıma sahip 100, 1000 ve 10000 kişilik örneklem grubundan oluşan veri setleri oluşturulmuştur. Sonuç olarak MST koşullarına uygun formda ve farklı örneklem koşullarını (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) sağlayan toplamda 18 ayrı veri seti üretilmiştir.

3.4. Verilerin Analizi

MST yöntemi ile simülatif olarak elde edilen verilerin tamamı öncelikle KTK ve IRT kuramlarının yapılarına göre uyarlanarak analiz edilebilir forma getirilmiştir. Her bir kuramın barındırdığı farklı yapılar nedeniyle bir kurama göre verinin sahip olması gereken formatla diğer kurama göre sahip olması gereken format ve bu formatların analizini yapabilen programlar farklılaşmaktadır. Bu sebeple verilerin tamamı öncelikle ilgili formlara uygun hale getirilmiştir. Bireylerin KTK'ya uygun yetenek kestirimleri MST yönteminin yapısı gereği çok sayıda "NaN (Not a Number)" değer diğer adıyla "Eksik" veri barındırmasından kaynaklı olarak Excel programı aracılığıyla her bir veri seti için manuel olarak hesaplanmıştır. IRT'ye uygun yetenek kestirimleri ise R programında "mirt" paketi kullanılarak elde edilmiştir. MST, KTK ve IRT'ye dayalı olarak üretilen verilerin standart hatalar hesaplanarak bu standart hatalar uygun istatistiksel analizlerle birbirleriyle karşılaştırılmıştır. Analizlerde standart hata

değerlerine odaklanılmasının sebebi; ölçüm doğruluğu tahminlerinin standart hata değerleri aracılığıyla elde edilebilir olması ve bu durumun MST ölçümleri için de geçerli olmasından kaynaklanmaktadır (Park vd., 2017). Gardner'a (1989) göre her test puanı bir ölçüm hatası içermektedir. Genel olarak, bir parametrenin istatistik tahminindeki hata miktarının bir ölçüsü olarak ifade edilen "Standart Hata", tahmin ettiği parametreye göre bir tahmincinin değişkenliğinin bir indeksidir (Ayala, 2009, s. 27-31). Standart ölçüm hatası, incelenen belirli örnekten bağımsızdır ve yetenek sürekliliğinin farklı noktalarında yetenek tahminindeki hata miktarının bir göstergesidir (Lord ve Norvick, 2008). Her madde, testteki diğer maddelerden bağımsız olarak standart hataya katkıda bulunmaktadır. Standart hatalar bilgi fonkisyonu aracılığıyla elde edildiğinden dolayı ölçüm sonucunda ulaşılan bilgi miktarı ne kadar yüksekse ölçümün standart hatası da o kadar düşüktür (DeMars, 2018, s. 67-69; Reise ve Haviland, 2005). Bu kapsamda ölçek bilgisi arttıkça hata değeri (SEM: Standard Error of Measurement) azalmakta olup (Reise ve Henson, 2003) belirli bir yetenek düzeyinde bir test tarafından sağlanan daha fazla bilgi, yetenek tahmini ile ilişkili hataların daha küçük olduğu anlamına gelmektedir (Hambleton ve Jones, 1993, s. 42). Bir diğer ifadeyle standart hatanın değeri ne kadar büyük olursa, hata o kadar büyük olur ve parametrenin değerinden o kadar az emin oluruz. Bu nedenle standart hata değerlerine yönelik olarak yapılan hesaplamalarda büyük standart hata değerine sahip araçlar küçük hata değerine sahip araçlara göre daha az tercih edilmektedir (Magno, 2009, s. 1-5; Overton, 2012, s. 116-123). Biirbaum, (2008) standart ölçüm hatasını ölçülen özellik yeteneğinin her seviyesinde bir test öğesinin ölçüm etkinliği olarak tanımlamaktadır. Yetenek tahminleri ve her maddenin geçerliliği gizli özellik modelindeki standart hata ile sağlanmaktadır. Standart hata, ölçümün kesinliğinin ve dolayısıyla testin gerçek θ (Theta)'yı ne kadar iyi tahmin ettiğinin bir göstergesidir (Macken-Ruiz, 2008). Şekil 3.5.'te görsel olarak ifade edildiği üzere bilgi miktarı ile stardart hata arasında ters yönlü bir ilişki bulunmaktadır (Kline, 2005, s. 122; Petrillo vd., 2015, s. 31).



Kaynak: Kline, T. J. B. (2005). Psychological testing: A practical approach to design and evaluation, Thousand Oaks, CA: Sage. Publications, Inc., https://www.doi.org/10.4135/9781483385693

Şekil 3.5. *Bilgi miktarı ile standart hata değeri arasındaki ilişki*

Elde edilen yetenek ölçüleri arasındaki korelasyon katsayıları ise JASP programında analiz edilerek "Kendall's Tau-b" korelasyon katsayısı ile karşılaştırılmıştır. Bu aşamada "Kendall's Tau-b" korelasyon katsayısının kullanılmasının sebebi veri setlerinin toplam puanların sıralama amacıyla kullanılma ihtimallerinin daha yüksek olmasıdır. Korelasyon analizlerinde hangi testin uygulanacağı değişkenin tipi ve dağılım özelliği (sıralı, sürekli, vb.) koşullarına göre değişebilmektedir. Örneğin; Verilerin rastgele dağılım esasına uyduğunun kabul edildiği parametrik bir dağılımda sürekli değişkenlerin birbirleriyle doğrusal ilişkisinin derecesine "*Pearson*" korelasyon testi uygulanarak bakılmaktadır (Chen ve Popovich, 2002; Dawson ve Trapp, 2004; Pagano ve Gauvreau, 2018). Parametrik olmayan (Nonparametric) sürekli değişkenler arası doğrusal ilişkinin derecesi ise "*Spearman*" korelasyon testi aracılığıyla analiz edilmektedir (Dawson ve Trapp, 2004; Pagano ve Gauvreau, 2018). Değişkenlerden birinin ya da her ikisinin de sıralı değişken olması durumunda kullanılan test ise "*Kendall'ın tau-b (τb)*" korelasyon analizidir (Arndt, Turvey ve Andreasen, 1999; Chen ve Popovich, 2002). Tez çalışması kapsamında incelenen KTK, IRT ve MST temelli puanların sıralama amacıyla kullanılması ihtimallerinin yüksek olması sebebiyle *Kendall'ın tau-b (τb)* katsayısının kullanımı tercih edilmiştir.

AIC = 2k – 2ln(L) formülü ile ifade edilen AIC (Akaike Information Criterion) değerleri uygun istatistiksel analizler aracılığıyla R Studio programında hesaplanarak model ölçümü yapılmış ve optimum model tespit edilmeye çalışılmıştır. Bilindiği üzere istatistiksel modellemedeki en göz korkutucu zorluklardan biri temel verileri karakterize etmek için bir aday koleksiyonundan uygun bir model seçmektir. Bu anlamda istatistik alanında ilk model seçim kriteri olan AIC (Akaike Information Criterion) aynı zamanda istatistiksel uygulamalarda en yaygın bilinen ve kullanılan model seçim araçlarından biri olmaya devam etmektedir (Cavanaugh ve Neath, 2019; Portet, 2020). AIC (Akaike Information Criterion) model seçim kriterleri çok basit ya da gereksiz yere karmaşık olan aday modelleri diskalifiye ederek bir aday koleksiyonu arasında uygun yapı ve boyutta bir modelin belirlenmesinde yararlı bir araç sağlamaktadır (Cavanaugh ve Neath, 2019). Kriterin amacı, Kullback-Leibler bilgisi anlamında üretici modele en yakın ve en uygun olan aday modeli belirlemektir. Akaike tarafından türetilen AIC beklenen göreli Kullback-Leibler (KL) sapmasının bir tahmini olup aday model ile gerçek model arasındaki mesafeyi ölçerek mesafe ne kadar yakınsa aday modelin gerçeğe o kadar benzer kabul edilmektedir (Vrieze, 2012). AIC, özellikle tahmine dayalı modelleme uygulamaları için çok uygun bir model seçim aracı olarak nitelendirilmektedir (Cavanaugh ve Neath, 2019).

Son olarak farklı kuramlara (KTK, IRT ve MST) göre elde edilen puan sıralama farklarının tespiti yönünde analizler gerçekleştirilmiştir. KTK, IRT ve MST yöntemlerine göre elde edilen puanların sıralama farkları uygun istatistiksel analizler aracılığıyla R Studio programında hesaplanarak her bir kuram puan sıralama farkları açısından birbiri ile karşılaştırılmıştır. Puan sıralama farklarının analizine gereksinim duyulmasının sebebi önceki analizlerde tespit edilen "*Kendall'ın tau-b (τb)*" korelasyon katsayılarının oransal değerleri vermesi sıralama anlamında bu yönde bir bilgiyi içermemesidir. Açık ve uzaktan öğrenme sistemleri dahil büyük ölçekli merkezi sınav gerçekleştiren tüm sistemlerde başarılı/başarısız, geçti/kaldı veya işe alımlarda puan sıralamasına göre sonuca karar verilmektedir. Puan sıralamasının bu denli önemli olduğu bir sınavda puanların maksimum kaç kişiye kadar yanılarak hesaplanmış olabileceğinin tespiti yapılmaya çalışılmıştır. Örneğin; KTK'nın 1. sıraya koyduğu öğrenciyi IRT hangi sırada konumlandırmıştır? Benzer şekilde MST'ye göre 1500. sırada yer alan öğreci IRT ya da KTK'ya göre kaçıncı sıradadır? Verilerin analizi tüm bu soruların cevabını içerecek

biçimde her üç kurama (KTK, IRT ve MST) göre farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) puan sıralamalarının göstermiş olduğu değişimleri birbirleri ile karşılaştırılarak gerçekleştirilmiştir.

4. BULGULAR

Bu bölümde araştırma soruları çerçevesinde MSTGen yazılımı aracılığıyla farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) ve farklı yöntemlerle (KTK, IRT, MST) bir simülasyon çalışması sonucunda üretilen 18 ayrı veri setine ait bulgulara yer verilmiştir. Söz konusu bulgular "Standart Hata Değerleri", "Farklı Yöntemlere Göre Elde Edilen Yetenek Ölçülerinin Korelasyon Katsayıları", "Farklı Yöntemlerin Verilere Ne Kadar Uyuştuğunu Gösteren AIC Değerleri" ve "Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları" alt başlıkları ile ifade edilmeye çalışılmıştır.

4.1. Standart Hata Değerleri

Araştırmada "MST test sunum yöntemi kullanılarak elde edilen yetenek kestirimleri (puanlar) ile aynı verilerden elde edilen KTK ve IRT'ye dayalı yetenek kestirimleri arasında simülatif bir ortamda MST lehine anlamlı bir farklılık var mıdır?" ana sorusu bağlamında aşağıda belirtilen alt sorulara yanıt aranmıştır.

Araştırma Soruları:

- **MST yöntemi ile elde edilen puanların standart hataları geleneksel KTK yöntemi ile elde edilen puanların standart hatalarından farklılık göstermekte midir?**
- **MST yöntemi ile elde edilen puanların standart hataları IRT yöntemi ile elde edilen puanların standart hatalarından farklılık göstermekte midir?**

Yukarıda yer alan ilk iki araştırma sorusunu yanıtlamak amacıyla "Standart Hata Değerleri" aşağıdaki formüller aracılığıyla her bir yöntem bağlamında ayrı ayrı hesaplanmıştır:

KTK için Standart Hata Değeri Formülü (van Rijn, 2016, s. 253);

$$\sigma_E = \sqrt{(1 - \rho_{XT}^2)}\sigma_X \qquad (4.1)$$

IRT için Standart Hata Değeri Formülü (Ayala, 2009, s. 160);

$$SEE(\hat{\theta}_i) = \sqrt{\sum_{j=1}^{L} \left\{ \frac{p_j(1-\chi_j)^2}{\alpha_j^2(1-p_j)(p_j-\chi_j)^2} \right\}} \quad (4.2)$$

MST için Standart Hata Değeri Formülü (Folk ve Smith, 2002, s. 45);

$$CSEM = \left[\sum_{i=1}^{n} I(\theta, u_i) \right]^{-1/2} = \left[\sum_{i=1}^{n} \frac{P_i'^2}{P_i(1-P_i)} \right]^{-1/2} \quad (4.3)$$

Yapılan analizler sonucunda ulaşılan bulgular şu şekildedir:

Tablo 4.1. *Normal dağılma sahip veri setlerinde KTK, IRT 3PL ve MST-R'ye göre standart hata değerleri*

	KTK_se	IRT_3PL_se	MST_R_se
100 Kişi	0,87013152	0,1810704	0,28203
1000 Kişi	0,84187527	0,2739308	**0,26777**
10000 Kişi	0,90876838	0,22273809	0,26452

Tablo 4.2. *Normal dağılma sahip veri setlerinde KTK, IRT 3PL ve MST-S'ye göre standart hata değerleri*

	KTK_se	IRT_3PL_se	MST_S_se
100 Kişi	0,7404971	0,1984967	0,2972
1000 Kişi	0,94392453	0,27393077	0,29626
10000 Kişi	0,87392764	0,3388127	**0,29014**

Tablo 4.1. ve Tablo 4.2.'de normal dağılıma sahip veri setlerinden elde edilen bulgulara hem MST-R hem de MST-S temelinde yer verilmiştir. İlgili tablolarda MST, KTK ve IRT 3PL kuramlarına ait hesaplanan standart hata değerleri karşılaştırılmıştır. Elde edilen bulgulara göre en az standart hata değerine sahip yöntem IRT 3PL'dir. MST test sunum yönteminin ise KTK'ya göre önemli derecede daha az hata ile kestirim yapabildiği tespit edilmiştir. Aynı zamanda MST yöntemi IRT 3PL'ye oldukça yakın hata değerleri ile kestirim yapabilmesinin yanı sıra 1000 kişilik veri setinde MST-R yönteminin 10000 kişilik veri setinde ise MST-S yönteminin IRT 3PL'den daha az hata ile kestirim yapabildiği sonucuna ulaşılmıştır.

Tablo 4.3. *Normal olmayan dağılma (sola çarpık) sahip veri setlerinde KTK, IRT 3PL ve MST-R'ye göre standart hata değerleri*

	KTK_se	IRT_3PL_se	MST_R_se
100 Kişi	0,71965076	0,2930234	0,33582
1000 Kişi	0,7086569	0,2813307	0,33733
10000 Kişi	0,62082535	0,2675895	0,31973

Tablo 4.4. *Normal olmayan dağılma (sola çarpık) sahip veri setlerinde KTK, IRT 3PL ve MST-S'ye göre standart hata değerleri*

	KTK_se	IRT_3PL_se	MST_S_se
100 Kişi	0,5805668	0,2255257	0,34198
1000 Kişi	0,65976709	0,2827763	0,3347
10000 Kişi	0,64587738	0,3102667	0,35355

Tablo 4.3. ve Tablo 4.4.'te normal olmayan (sola çarpık) dağılıma sahip veri setlerinden elde edilen bulgulara hem MST-R hem de MST-S temelinde yer verilmiştir. İlgili tablolarda MST, KTK ve IRT 3PL kuramlarına ait hesaplanan standart hata değerleri karşılaştırılmıştır. Elde edilen bulgulara göre en az standart hata değerine sahip yöntem IRT 3PL olarak bulunmuştur. Analiz sonuçlarına göre MST test sunum yönteminin ise KTK'ya göre önemli derecede daha az hata ile kestirim yapabildiği tespit edilmiştir.

Tablo 4.5. *Normal olmayan dağılma (sağa çarpık) sahip veri setlerinde KTK, IRT 3PL ve MST-R'ye göre standart hata değerleri*

	KTK_se	IRT_3PL_se	MST_R_se
100 Kişi	0,60913817	0,2283728	0,39721
1000 Kişi	0,59352599	0,2977713	0,38019
10000 Kişi	0,64604923	0,2931815	0,3701

Tablo 4.6. *Normal olmayan dağılma (sağa çarpık) sahip veri setlerinde KTK, IRT 3PL ve MST-S'ye göre standart hata değerleri*

	KTK_se	IRT_3PL_se	MST_S_se
100 Kişi	0,54247368	0,2004066	0,40867
1000 Kişi	0,6828017	0,3390708	0,35256
10000 Kişi	0,67000614	0,3620553	0,38371

Tablo 4.5. ve Tablo 4.6.'da ise normal olmayan (sağa çarpık) dağılıma sahip veri setlerinden elde edilen bulgulara hem MST-R hem de MST-S temelinde yer verilmiştir. İlgili tablolarda MST, KTK ve IRT 3PL kuramlarına ait hesaplanan standart hata değerleri karşılaştırılmıştır. Elde edilen bulgulara göre en az standart hata değerine sahip yöntem IRT 3PL olarak bulunmuştur.

Tablo 4.5.'te 100 kişilik IRT 3 PL standart hata değerini hesaplamak için yapılan analizlerde kişi sayısı az olmasından ve MST yönteminin yapısı gereği çok sayıda NaN değer barındırmasından kaynaklı olarak R Studio yazılımı hata vermiştir. Bu sebeple neredeyse hiçbir bireyin almadığı sorular (23,43,45 ve 55. maddeler) çıkarılarak IRT 3PL hata değeri hesaplaması yapılmıştır.

Tablo 4.5.'te IRT 3 PL 10000 kişilik veri setinde 1 maddenin (m13) ve Tablo 4.6.'da IRT 3 PL 100 kişilik veri setinde ise 2 maddenin (m39 ve m66) alternatif kategori (yani 1-0 matrisli veri setinde tüm bireylerin bu maddelerdeki soruyu doğru yanıtlamasından kaynaklı olarak sıfır değere sahip veri bulunmaması durumu) barındırmamasından dolayı IRT 3 PL hata değeri hesaplamalarında R Studio analizlerinde sağlıklı kestirimler yapılamamıştır.

- **MST yöntemi ile elde edilen puanların standart hataları MST-R ve MST-S yöntemi ile elde edilme durumuna göre farklılık göstermekte midir?**

Yukarıda yer alan üçüncü araştırma sorusu çerçevesinde ulaşılan sonuçlar ise Tablo 4.7.'de ifade edilmektedir.

Tablo 4.7. *MST-R ve MST-S yöntemlerinin standart hata değerlerinin karşılaştırılması*

Örneklem Dağılımı	Kişi Sayısı	MST_R_se	MST_S_se
Normal Dağılım	100	0,28203	0,2972
	1000	0,26777	0,29626
	10000	0,26452	0,29014
Normal Olmayan (Sola Çarpık) Dağılım	100	0,33582	0,34198
	1000	0,33733	0,3347
	10000	0,31973	0,35355
Normal Olmayan (Sağa Çarpık) Dağılım	100	0,39721	0,40867
	1000	0,38019	0,35256
	10000	0,3701	0,38371

MST test sunum yöntemi kendi içerisinde MST-R ve MST-S olmak üzere iki farklı şekilde uygulanabilmektedir. Bu iki yöntemde kendi içerisinde standart hata değerleri temel alınarak karşılaştırılmış ve MST-R yönteminin çok büyük farklar olmamak kaydıyla MST-S yöntemine göre daha az hata ile kestirim yapabildiği sonucuna ulaşılmıştır.

Aşağıda dördüncü ve son araştırma sorusu bulguları yer almaktadır.

- **KTK, IRT ve MST yöntemlerinden hangisi en az standart hata ile gerçeğe en yakın kestirimi yapabilmektedir?**

MST, KTK ve IRT 3PL hata değerleri karşılaştırıldığında elde edilen bulgulara göre en az standart hata değerine sahip yöntem IRT 3PL olarak bulunmuştur. MST test sunum yönteminin ise KTK'ya göre önemli derecede daha az hata ile kestirim yapabildiği tespit edilmiştir.

4.2. Farklı Yöntemlere Göre Elde Edilen Yetenek Ölçülerinin Korelasyon Katsayıları

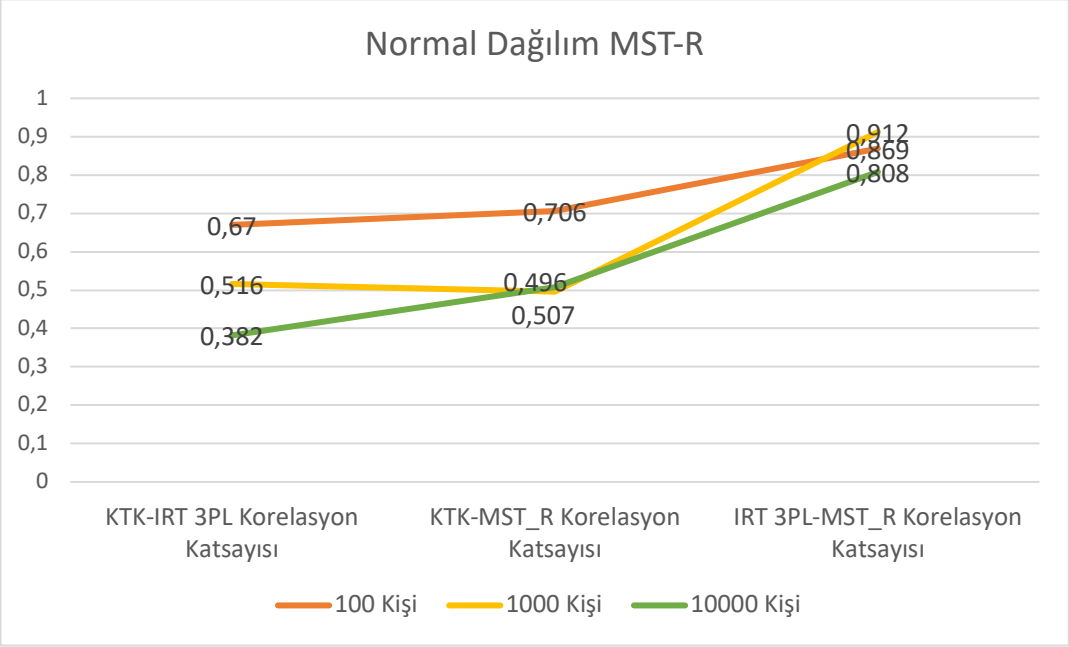
Analizlerin bir sonraki aşamasıda KTK, IRT ve MST yöntemleri veri setlerinin 1-0 matrisli sıralı değişken olmasından dolayı elde edilen yetenek ölçüleri arasındaki korelasyon katsayıları ise JASP programında analiz edilerek "Kendall's Tau-b" korelasyon katsayısı ile karşılaştırılmış ve aşağıda Tablo 4.8.'de belirtilen bulgulara ulaşılmıştır:

Tablo 4.8. *KTK, IRT 3PL ve MST'ye (MST-R ve MST-S) göre elde edilen yetenek ölçülerinin birbirleri ile korelasyonu*

Kendall's Tau-b	**Normal Dağılım**	**Kişi Sayısı**	**KTK-IRT 3PL Korelasyon Katsayısı**	**KTK-MST_R Korelasyon Katsayısı**	**IRT 3PL-MST_R Korelasyon Katsayısı**	**Kişi Sayısı**	**KTK-IRT 3PL Korelasyon Katsayısı**	**KTK-MST_S Korelasyon Katsayısı**	**IRT 3PL-MST_S Korelasyon Katsayısı**
		100	**0.670**	**0.706**	0.869	**100**	0.843	0.855	0.865
		1000	**0.516**	**0.496**	0.912	**1000**	**0.745**	**0.737**	0.934
		10000	**0.382**	**0.507**	0.808	**10000**	0.803	0.802	0.952
	Normal Olmayan (Sola Çarpık) Dağılım	**Kişi Sayısı**	**KTK-IRT 3PL Korelasyon Katsayısı**	**KTK-MST_R Korelasyon Katsayısı**	**IRT 3PL-MST_R Korelasyon Katsayısı**	**Kişi Sayısı**	**KTK-IRT 3PL Korelasyon Katsayısı**	**KTK-MST_S Korelasyon Katsayısı**	**IRT 3PL-MST_S Korelasyon Katsayısı**
		100	0.809	0.830	0.884	**100**	0.871	0.889	0.908
		1000	0.808	0.826	0.944	**1000**	0.840	0.839	0.941
		10000	0.895	0.899	0.976	**10000**	0.861	0.852	0.941
	Normal Olmayan (Sağa Çarpık) Dağılım	**Kişi Sayısı**	**KTK-IRT 3PL Korelasyon Katsayısı**	**KTK-MST_R Korelasyon Katsayısı**	**IRT 3PL-MST_R Korelasyon Katsayısı**	**Kişi Sayısı**	**KTK-IRT 3PL Korelasyon Katsayısı**	**KTK-MST_S Korelasyon Katsayısı**	**IRT 3PL-MST_S Korelasyon Katsayısı**
		100	**0.624**	0.805	**0.645**	**100**	**0.782**	0.885	0.850
		1000	0.854	0.899	0.900	**1000**	0.855	0.863	0.917
		10000	0.824	0.824	0.957	**10000**	0.880	0.884	0.942

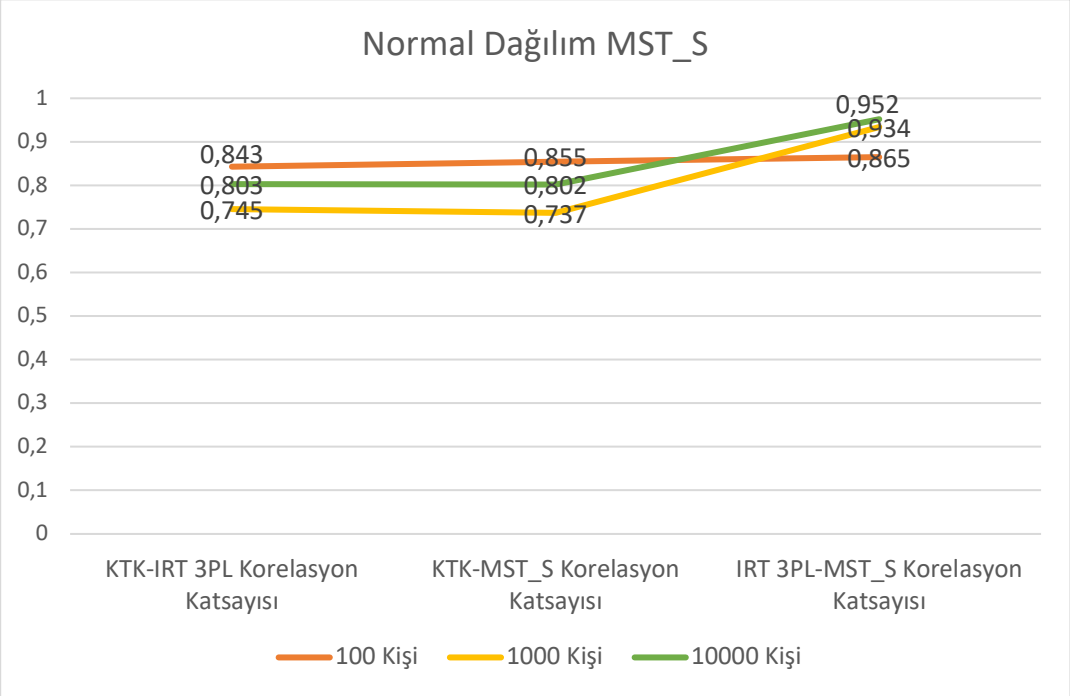
Farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) gerçekleştirilen analiz sonuçlarından elde edilen "Kendall's Tau-b" korelasyon katsayısı bulgularına göre MST ve IRT 3PL yöntemleri birbirleri ile büyük oranda uyum gösterirken KTK ile uyumsuzlaşmakta ve dikkate değer derecede büyük oranlarla birbirinden farklı puanlama yapmaktadır.

Yukarıda Tablo 4.8.'de verilen korelasyon analizi sonuçları dağılım şekillerine göre her bir kuram özelinde ayrı ayrı ele alınarak aşağıda yer alan grafikler aracılığıyla detaylandırılarak sunulmuştur:



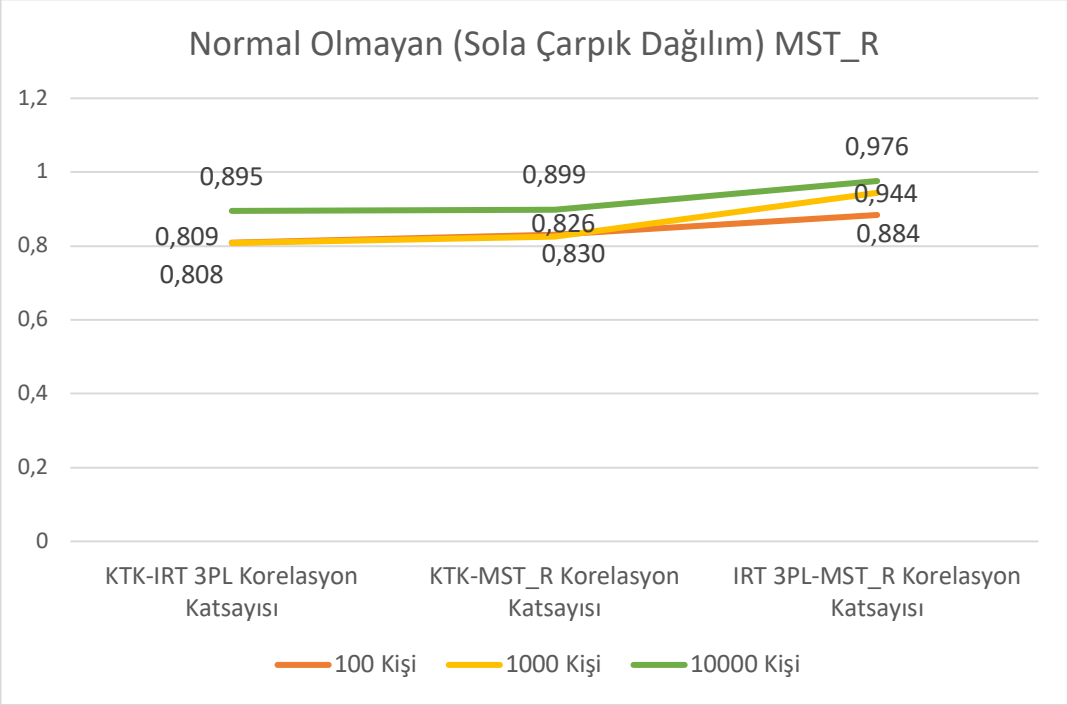
Grafik 4.1. *"Routing" yöntemine göre normal dağılım gösteren simülatif verilerin korelasyonları*

Grafik 4.1.'de görüldüğü üzere MST'nin "Routing" yöntemi ile elde edilen normal dağılıma sahip veri setine uygulanan korelasyon analizi sonuçlarına göre KTK'nın puanlama yaparken diğer kuramlarla (MST-R ve IRT 3PL) önemli derecede farklılaştığı tespit edilmiştir. MST-R ve IRT 3PL yöntemlerinin ise birbirleri ile kayda değer biçimde uyum gösterdiği sonucuna ulaşılmıştır.



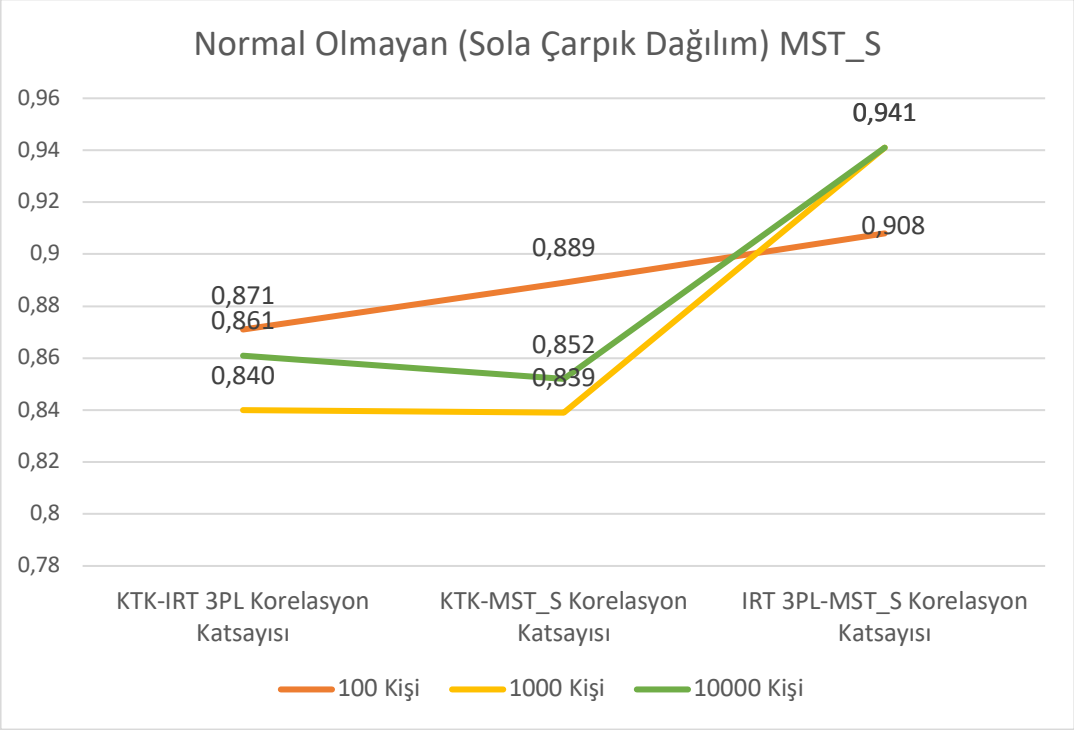
Grafik 4.2. *"Shaping" yöntemine göre normal dağılım gösteren simülatif verilerin korelasyonları*

Grafik 4.2.'de görüldüğü üzere MST'nin "Shaping" yöntemi ile elde edilen normal dağılıma sahip veri setine uygulanan korelasyon analizi sonuçlarına göre ise KTK'nın puanlama yaparken diğer kuramlarla (MST-S ve IRT 3PL) uyumu artmakla birlikte MST-S ve IRT 3PL yöntemlerinin birbirleri ile uyumunun daha ileri seviyede olduğu sonucuna ulaşılmıştır.



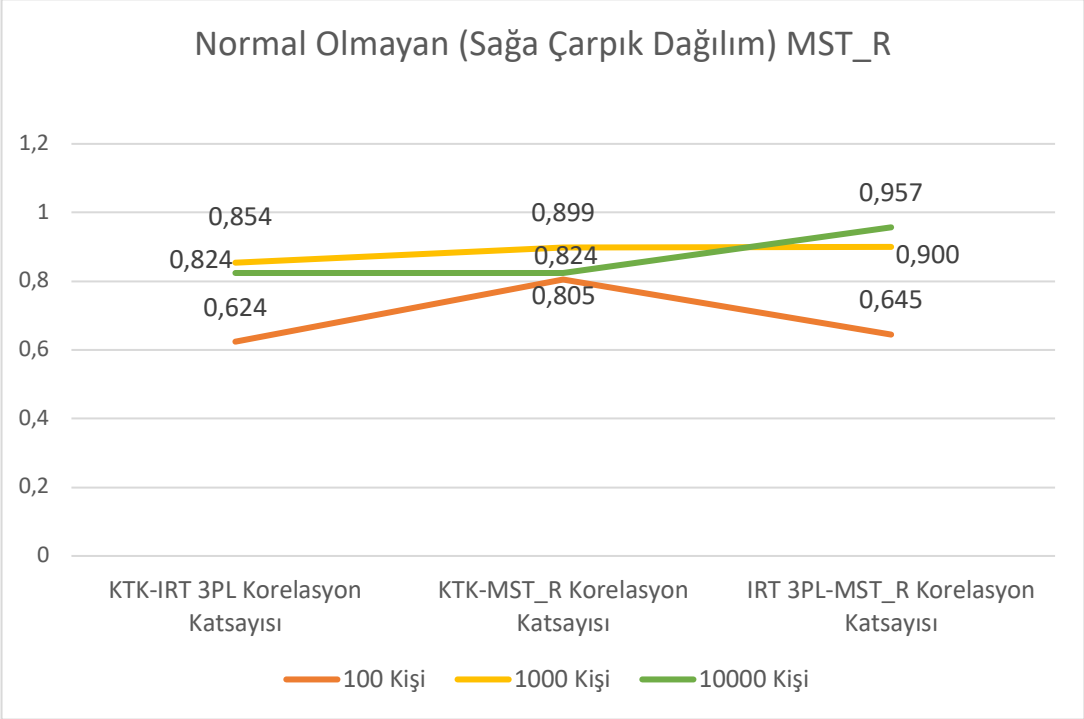
Grafik 4.3. *"Routing" yöntemine göre normal olmayan dağılım (sola çarpık) gösteren simülatif verilerin korelasyonları*

Grafik 4.3.'de görüldüğü üzere MST'nin "Routing" yöntemi ile elde edilen normal olmayan dağılıma (sola çarpık dağılım) sahip veri setine uygulanan korelasyon analizi sonuçları KTK'nın puanlama yaparken diğer kuramlarla (MST-R ve IRT 3PL) uyumu artmakla birlikte MST-R ve IRT 3PL yöntemlerinin birbirleri ile uyumunun daha ileri seviyede olduğunu göstermektedir.



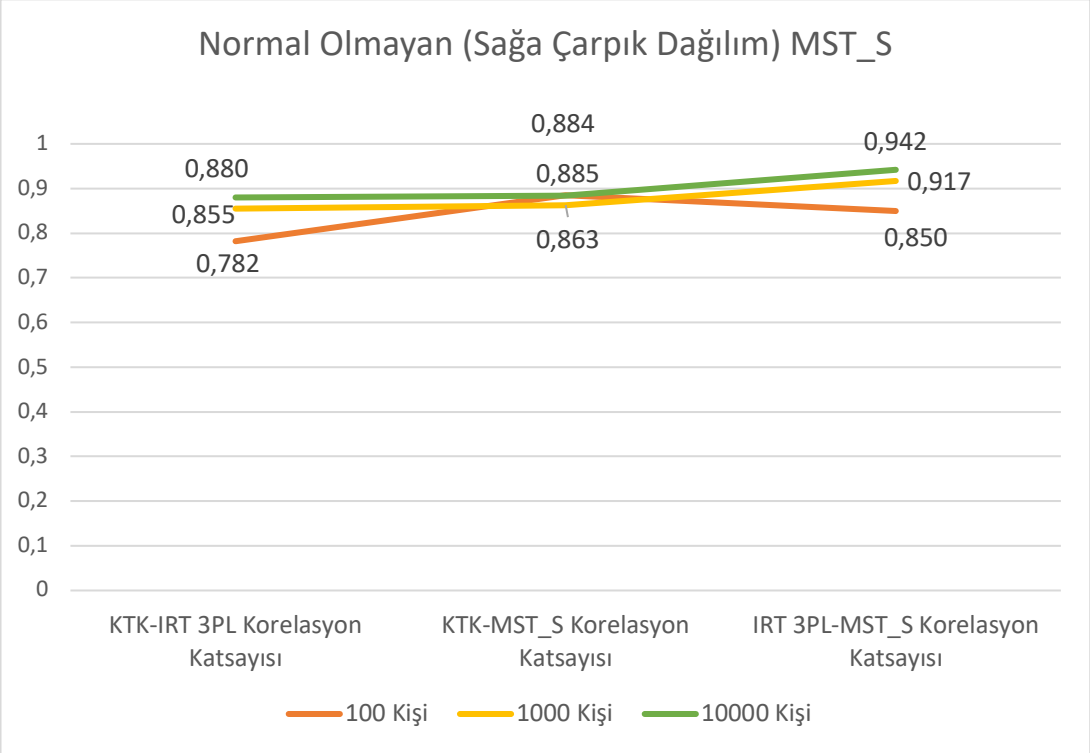
Grafik 4.4. *"Shaping" yöntemine göre normal olmayan dağılım (sola çarpık dağılım) gösteren simülatif verilerin korelasyonları*

Grafik 4.4.'de görüldüğü üzere MST'nin "Shaping" yöntemi ile elde edilen normal olmayan dağılıma (sola çarpık dağılım) sahip veri setine uygulanan korelasyon analizi sonuçları KTK'nın puanlama yaparken diğer kuramlarla (MST-S ve IRT 3PL) uyumunun artış yönünde bir seyir izlediğini göstermekle birlikte MST-S ve IRT 3PL yöntemlerinin birbirleri ile uyumunun daha ileri seviyede olduğu tespit edilmiştir.



Grafik 4.5. *"Routing"* yöntemine *göre normal olmayan dağılım (sağa çarpık) gösteren simülatif verilerin korelasyonları*

Grafik 4.5.'de görüldüğü üzere MST'nin "Routing" yöntemi ile elde edilen normal olmayan dağılıma (sağa çarpık dağılım) sahip veri setine uygulanan korelasyon analizi sonuçları göre KTK'nın puanlama yaparken diğer kuramlarla (MST-R ve IRT 3PL) uyumunun artış yönünde bir seyir izlediği görülmektedir. MST-R ve IRT 3PL yöntemlerinin ise özellikle 100 kişilik veri setinde birbirleri ile uyumunun diğer dağılım türlerine oranla daha düşük bir seviyede olduğu sonucuna ulaşılmıştır.



Grafik 4.6. *"Shaping" yöntemine göre normal olmayan dağılım (sağa çarpık) gösteren simülatif verilerin korelasyonları*

Grafik 4.6.'de görüldüğü üzere MST'nin "Shaping" yöntemi ile elde edilen normal olmayan dağılıma (sağa çarpık dağılım) sahip veri setine uygulanan korelasyon analizi sonuçları göre KTK'nın puanlama yaparken diğer kuramlarla (MST-S ve IRT 3PL) uyumunun artış yönünde bir seyir izlediği görülmektedir. MST-S ve IRT 3PL yöntemlerinin ise bu veri setinde de birbiri ile uyumu daha yüksek oranlara sahip olmakla birlikte özellikle 100 kişilik veri setinde birbirleri ile uyumunun diğer dağılım türlerine oranla daha düşük bir seviyede olduğu sonucuna ulaşılmıştır.

4.3. Farklı Yöntemlerin Verilere Ne Kadar Uyuştuğunu Gösteren AIC Değerleri

KTK, IRT ve MST yöntemleri için elde edilen yetenek kestirimlerinden yola çıkarak AIC değerleri hesaplanmış ve optimum model tespit edilmeye çalışılmıştır (Bu aşamada tüm maddelerinde "Eksik" veri barındırmasından kaynaklı olarak "Shaping (MST-S)" yöntemi için AIC değeri hesaplanamamıştır. "Routing (MST-R)" yöntemi için ise sadece tüm katılımcıların yanıtlamış olduğu "Eksik" veri barındırmayan maddeler için AIC değeri hesaplanabilmiştir.). Bu sınırlılıklar dahilinde elde edilen bulgular şu şekildedir:

Tablo 4.9. *Normal dağılma sahip veri setlerinde KTK, IRT 3PL ve MST-R'ye göre AIC (Akaike Information Criterion) değerleri*

	100 Kişi	1000 Kişi	10000 Kişi
KTK	235,2481	2843,3	28283,85
IRT 3PL	145,1369	2265,982	18387,68
MST_R	**143,8551**	**1473,117**	**14249,21**

Normal dağılıma sahip veri setlerinde tüm örneklem grupları için en uygun model "Routing (MST-R)" yöntemi olarak tepit edilmiştir. KTK ise veriler ile daha az uyumlu olan model olarak görünmektedir.

Tablo 4.10. *Normal dağılma sahip olmayan (sola çarpık) veri setlerinde KTK, IRT 3PL ve MST-R'ye göre AIC (Akaike Information Criterion) değerleri*

	100 Kişi	1000 Kişi	10000 Kişi
KTK	215,6836	1752,674	15744,92
IRT 3PL	220,2056	**1611,832**	**13870,8**
MST_R	**200,3982**	1873,512	16120,17

Normal olmayan (sola çarpık) veri setlerinde ise 100 kişilik örneklem grubunda en uygun model "Routing (MST-R)" olarak tespit edilirken 1000 ve 10000 kişilik örneklem gruplarında en uygun model IRT 3PL olarak tespit edilmiştir. KTK ise veriler ile daha az uyumlu olan model olarak görünmektedir.

Tablo 4.11. *Normal dağılma sahip olmayan (sağa çarpık) veri setlerinde KTK, IRT 3PL ve MST-R'ye göre AIC (Akaike Information Criterion) değerleri*

	100 Kişi	1000 Kişi	10000 Kişi
KTK	163,3668	1427,549	16518,93
IRT 3PL	**137,1622**	**818,2261**	**7731,794**
MST_R	148,468	1185,248	11541,99

Normal olmayan (sağa çarpık) veri setlerinde tüm örneklem gruplarında en uygun model IRT 3PL olarak tespit edilmiştir. KTK ise tercih edilmemesi gereken model sınıfında yer almaktadır.

4.4. Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları

Araştırma sorusu;

- **KTK, IRT ve MST yöntemleri ile elde edilen puanlarla yapılan sıralamalar farklılık göstermekte midir?**

Yukarıda ifade edilen araştırma sorusu temelinde üç kurama (KTK, IRT 3PL ve MST) göre farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) gerçekleştirilen analizler sonucunda elde edilen bulgular aşağıda yer alan tablo ve grafikler aracılığıyla sunulmuştur:

Tablo 4.12. *Normal dağılıma sahip veri setlerinin kuramlara göre puan sıraları farkı (MST-R)*

MST_R	KTK_IRT 3PL		KTK_MST-R		IRT 3PL_MST-R	
	min.	max.	min.	max.	min.	max.
100 Kişi	-43	42	-40	38	-13	32
1000 Kişi	**-566**	444	**-559**	415	-139	137
10000 Kişi	**-5788**	**6384**	**-5465**	**5102**	-3347	3773

Normal dağılıma sahip "MST Routing" veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırılarak minimum ve maksimum değerleri hesaplanmıştır. Tablo 4.12.'den de görüleceği üzere KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarında 100, 1000 ve 10000 kişilik veri setlerinde puan sıralamalarındaki farklar (100 kişide: 43 sıra, 1000 kişide: 566 sıra ve 10000 kişide: 6384 sıraya kadar sınav katılımcılarının puan sıraları farklı sıralanabilmektedir.) önemli derecede artarkan IRT 3PL - MST-R karşılaştırmalarında aradaki farkların azaldığı sunucuna ulaşılmıştır.

Tablo 4.13. *Normal dağılıma sahip veri setlerinin kuramlara göre puan sıraları farkı (MST-S)*

MST_S	KTK_IRT 3PL		KTK_MST-S		IRT 3PL_MST-S	
	min.	max.	min.	max.	min.	max.
100 Kişi	-26	25	-26	27	-19	19
1000 Kişi	-468	**535**	-384	492	-123	173
10000 Kişi	-4026	4445	-3531	4510	-1270	1261

Normal dağılıma sahip "MST Shaping" veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırılarak minimum ve maksimum değerleri hesaplanmıştır. Tablo 4.13.'ten de görüleceği üzere KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarında 100, 1000 ve 10000 kişilik veri setlerinde puan sıralamalarındaki farklar (100 kişide: 27 sıra, 1000 kişide: 535 sıra ve 10000 kişide: 4510 sıraya kadar sınav katılımcılarının puan sıraları farklı sıralanabilmektedir.) önemli derecede artarkan IRT 3PL - MST-S karşılaştırmalarında aradaki farkların azaldığı sunucuna ulaşılmıştır.

Tablo 4.14. *Normal olmayan (sola çarpık) dağılıma sahip veri setlerinin kuramlara göre puan sıraları farkı (MST-R)*

MST_R	KTK_IRT 3PL		KTK_MST-R		IRT 3PL_MST-R	
	min.	max.	min.	max.	min.	max.
100 Kişi	-37	32	**-97**	19	**-97**	19
1000 Kişi	**-504**	253	-498	222	-83	170
10000 Kişi	-3962	1861	-4224	1863	-369	1003

Normal dağılıma sahip "MST Routing" veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırılarak minimum ve maksimum değerleri hesaplanmıştır. Tablo 4.14.'ten de görüleceği üzere KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarında 100, 1000 ve 10000 kişilik veri setlerinde puan sıralamalarındaki farklar (100 kişide: 97 sıra, 1000 kişide: 504 sıra ve 10000 kişide: 4224 sıraya kadar sınav katılımcılarının puan sıraları farklı sıralanabilmektedir.) önemli derecede artmaktadır. IRT 3PL - MST-R karşılaştırmalarında ise aradaki farklar azalmakla birlikte 100 kişilik veri seti için artan bir seyir izlemektedir.

Tablo 4.15. *Normal olmayan (sola çarpık) dağılıma sahip veri setlerinin kuramlara göre puan sıraları farkı (MST-S)*

MST_S	KTK_IRT 3PL		KTK_MST-S		IRT 3PL_MST-S	
	min.	max.	min.	max.	min.	max.
100 Kişi	-18	15	-21	14	-15	14
1000 Kişi	-333	235	-352	256	-93	108
10000 Kişi	-4222	2814	-4045	2592	-887	1588

Normal dağılıma sahip "MST Shaping" veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırılarak minimum ve maksimum değerleri hesaplanmıştır. Tablo 4.15.'ten de görüleceği üzere KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarında 100, 1000 ve 10000 kişilik veri setlerinde puan sıralamalarındaki farklar (100 kişide: 21 sıra, 1000 kişide: 352 sıra ve 10000 kişide: 4222 sıraya kadar sınav katılımcılarının puan sıraları farklı sıralanabilmektedir.) önemli derecede artarkan IRT 3PL - MST-S karşılaştırmalarında aradaki farkların azaldığı sunucuna ulaşılmıştır.

Tablo 4.16. *Normal olmayan (sağa çarpık) dağılıma sahip veri setlerinin kuramlara göre puan sıraları farkı (MST-R)*

MST_R	KTK_IRT 3PL		KTK_MST-R		IRT 3PL_MST-R	
	min.	max.	min.	max.	min.	max.
100 Kişi	**-67**	40	-32	34	-45	**65**
1000 Kişi	-321	222	-157	231	-180	267
10000 Kişi	-3333	4023	-3651	3955	-858	2547

Normal dağılıma sahip "MST Routing" veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırılarak minimum ve maksimum değerleri hesaplanmıştır. Tablo 4.16.'dan da görüleceği üzere KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarında 100, 1000 ve 10000 kişilik veri setlerinde puan sıralamalarındaki farklar (100 kişide: 67 sıra, 1000 kişide: 321 sıra ve 10000 kişide: 4023

sıraya kadar sınav katılımcılarının puan sıraları farklı sıralanabilmektedir.) önemli derecede artarmaktadır. IRT 3PL - MST-R karşılaştırmalarında ise aradaki farklar azalmakla birlikte 100 kişilik ve 1000 kişilik veri setleri için artan bir seyir izlemektedir.

Tablo 4.17. *Normal olmayan (sağa çarpık) dağılıma sahip veri setlerinin kuramlara göre puan sıraları farkı (MST-S)*

MST_S	KTK_IRT 3PL		KTK_MST-S		IRT 3PL_MST-S	
	min.	max.	min.	max.	min.	max.
100 Kişi	-35	32	-20	17	-18	29
1000 Kişi	-271	260	-254	288	-140	330
10000 Kişi	-4209	3168	-3326	2869	-1047	2308

Normal dağılıma sahip "MST Shaping" veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırılarak minimum ve maksimum değerleri hesaplanmıştır. Tablo 4.17.'den de görüleceği üzere KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarında 100, 1000 ve 10000 kişilik veri setlerinde puan sıralamalarındaki farklar (100 kişide: 35 sıra, 1000 kişide: 288 sıra ve 10000 kişide: 4209 sıraya kadar sınav katılımcılarının puan sıraları farklı sıralanabilmektedir.) önemli derecede artarkan IRT 3PL - MST-S karşılaştırmalarında aradaki farkların azaldığı sunucuna ulaşılmıştır.

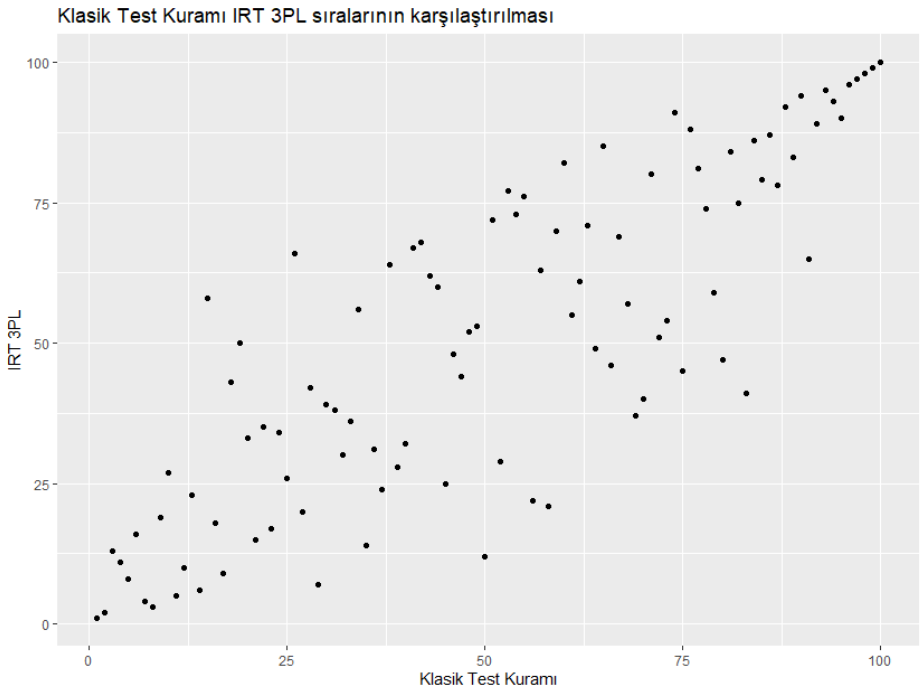
Yukarıda minimum ve maksimum değerler çerçevesinde ana hatları ile verilmiş olan puan sıralama farklarına yönelik analiz sonuçları aşağıda yer alan grafiklerle detaylı olarak her bir veri seti için ayrı ayrı sunulmuştur:

4.4.1. Normal Dağılıma Sahip Veri Setlerinin Kuramlara (KTK, IRT 3PLve MST) Göre Puan Sıraları Farkı

"KTK, IRT ve MST yöntemleri ile elde edilen puanlarla yapılan sıralamalar farklılık göstermekte midir?" araştırma sorusu kapsamında gerçekleştirilen analizlerden normal dağılıma yönelik elde edilen sonuçlar şu şekildedir:

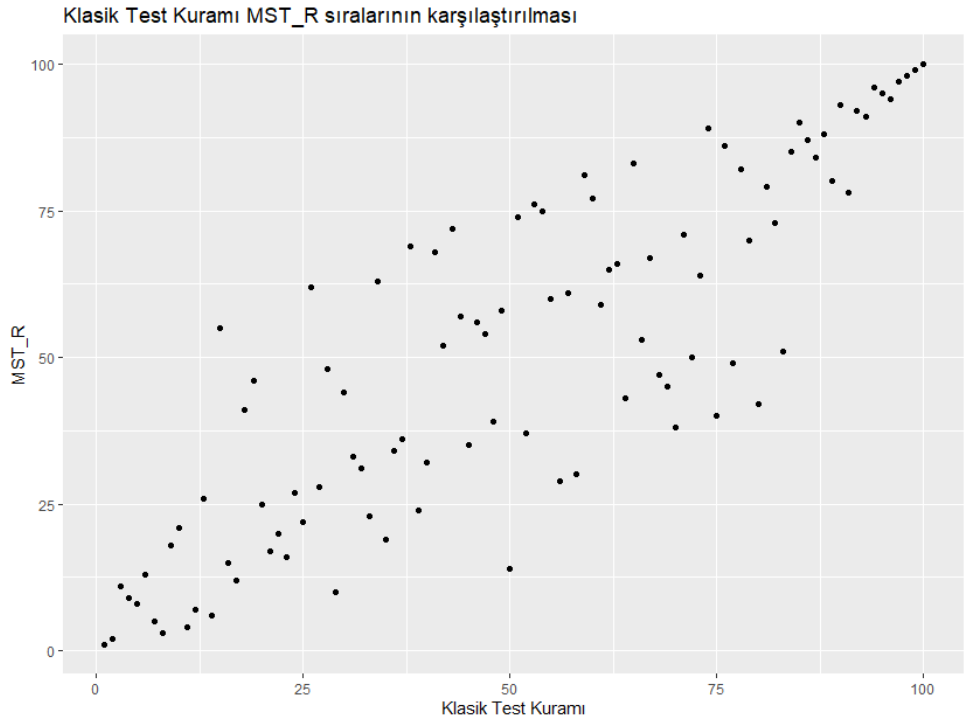
4.4.1.1. "Routing" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal dağılıma sahip "Routing" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



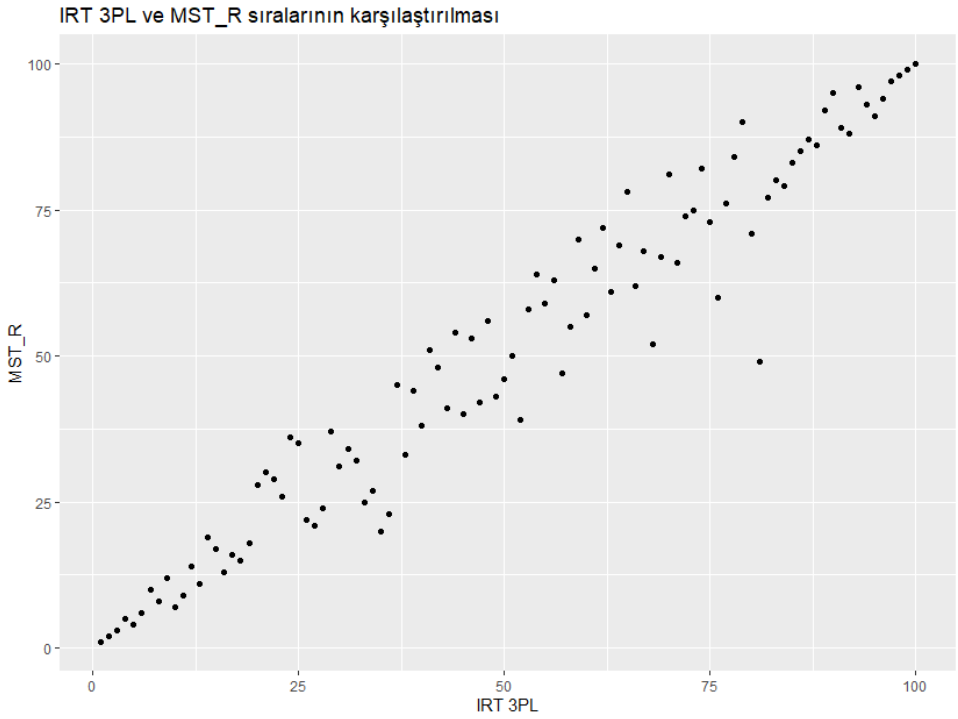
Grafik 4.7. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 100 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 100 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.8. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 100 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 100 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

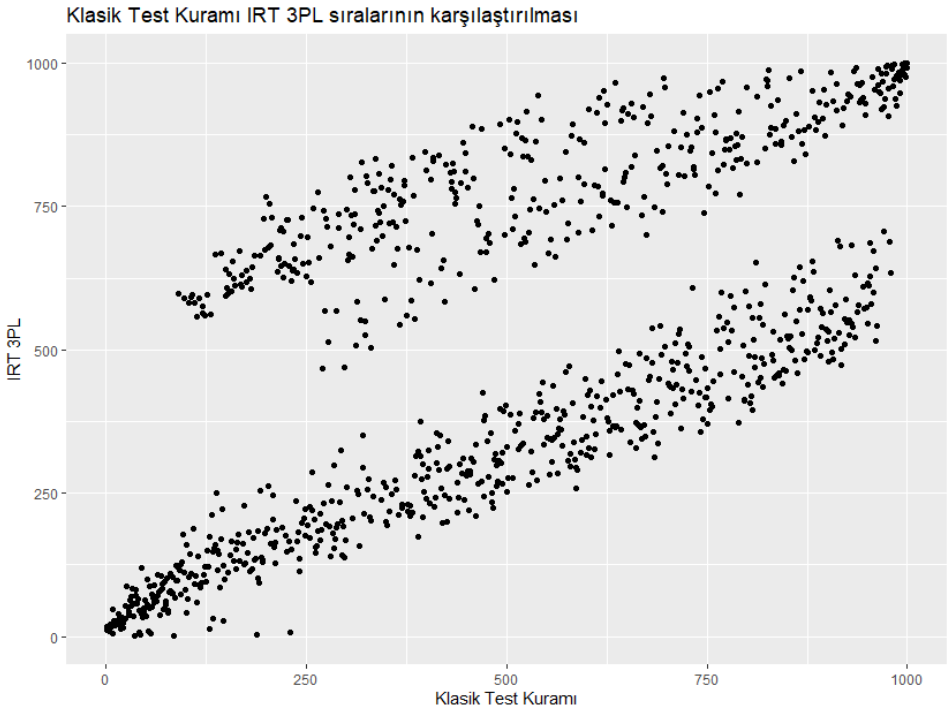


Grafik 4.9. *"Routing"* d *göre üretilmiş normal dağılıma sahip 100 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

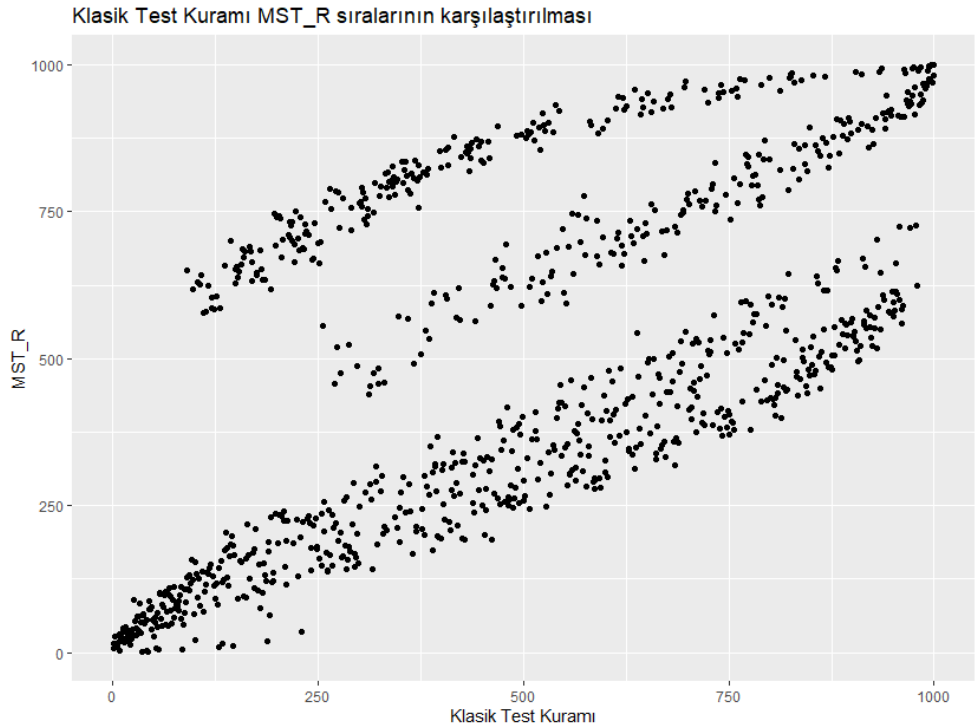
4.4.1.2. "Routing" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal dağılıma sahip "Routing" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



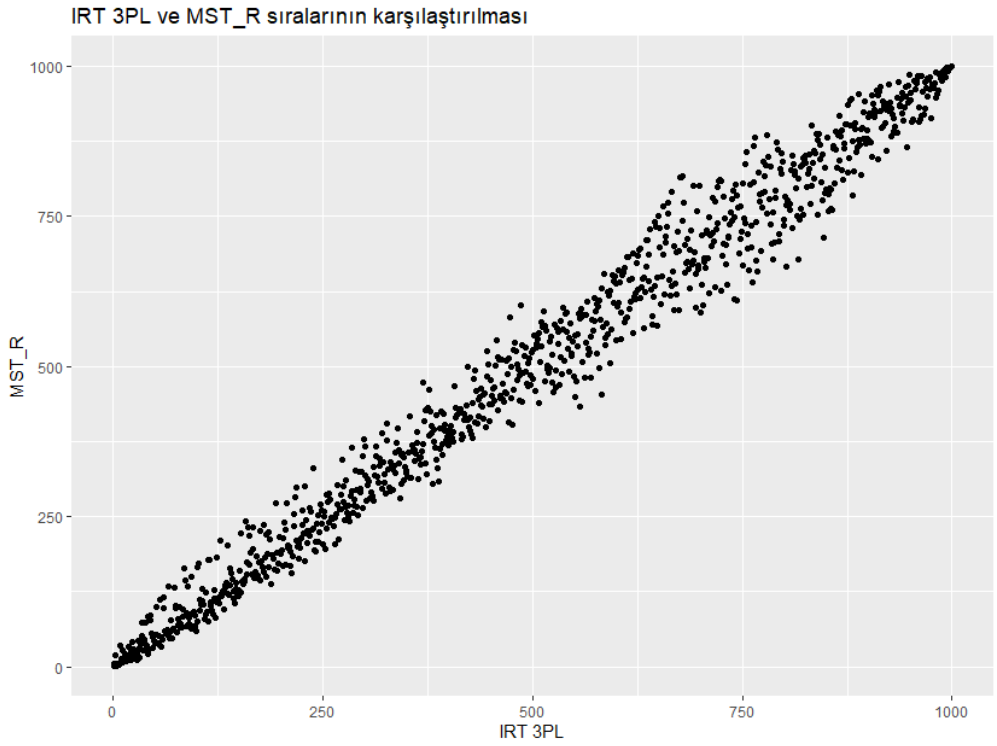
Grafik 4.10. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 1000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.11. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 1000 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

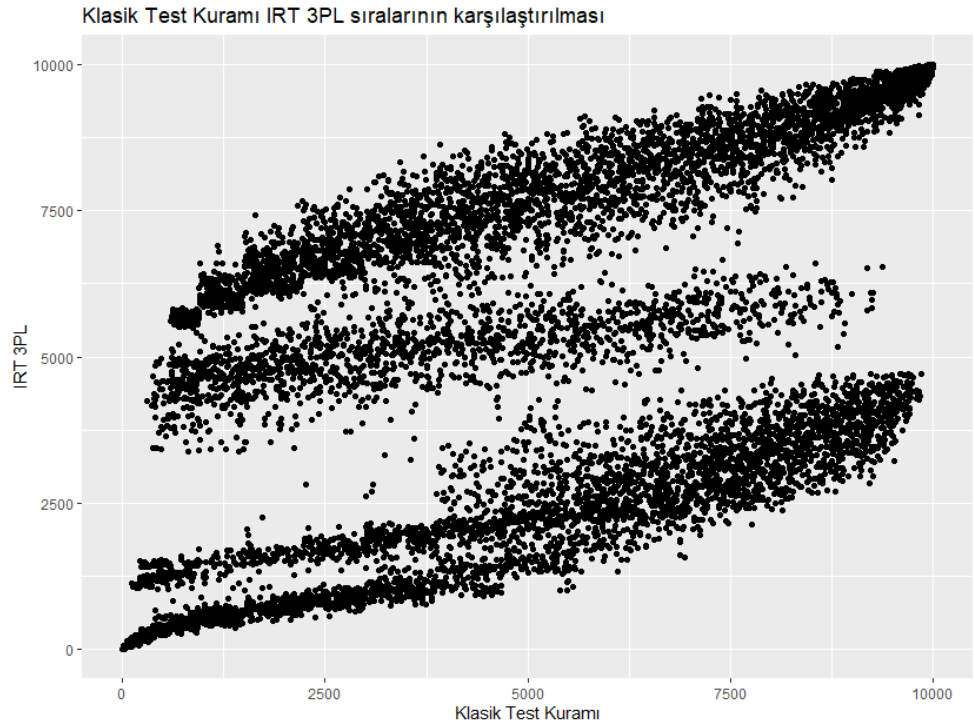


Grafik 4.12. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 1000 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 1000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

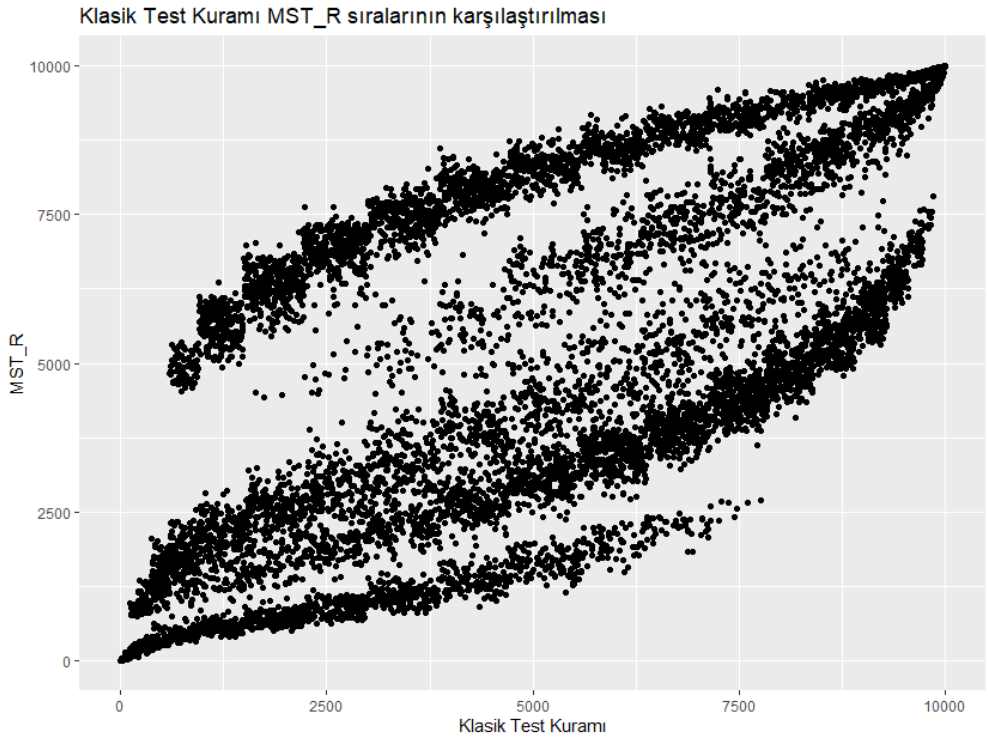
4.4.1.3. "Routing" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal dağılıma sahip "Routing" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



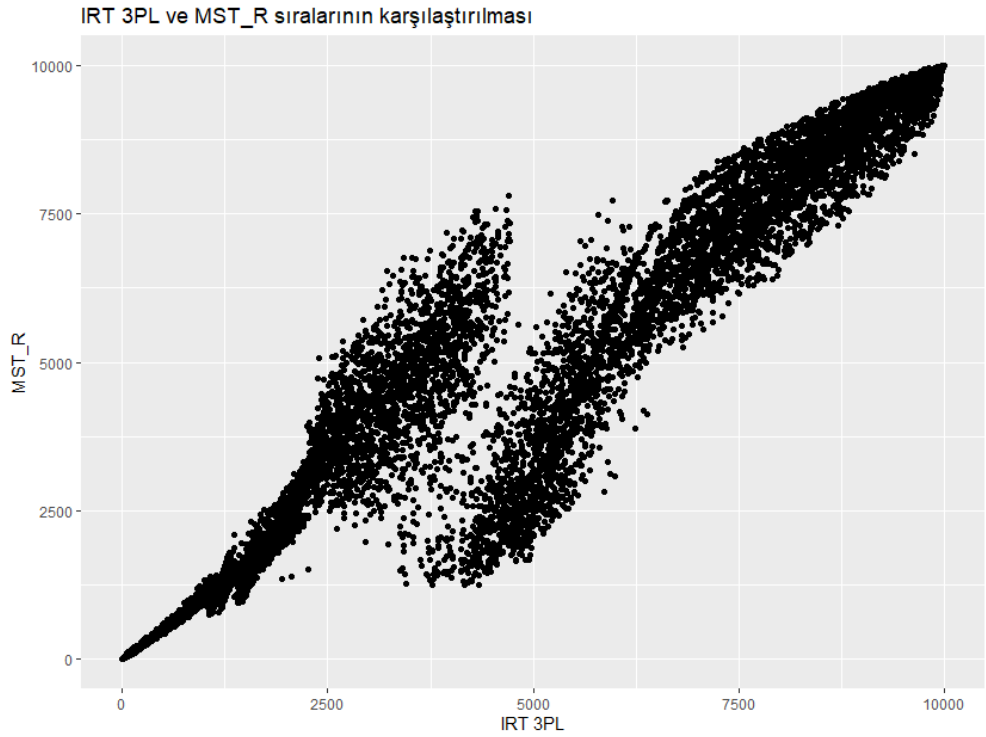
Grafik 4.13. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 10000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.14. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 10000 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

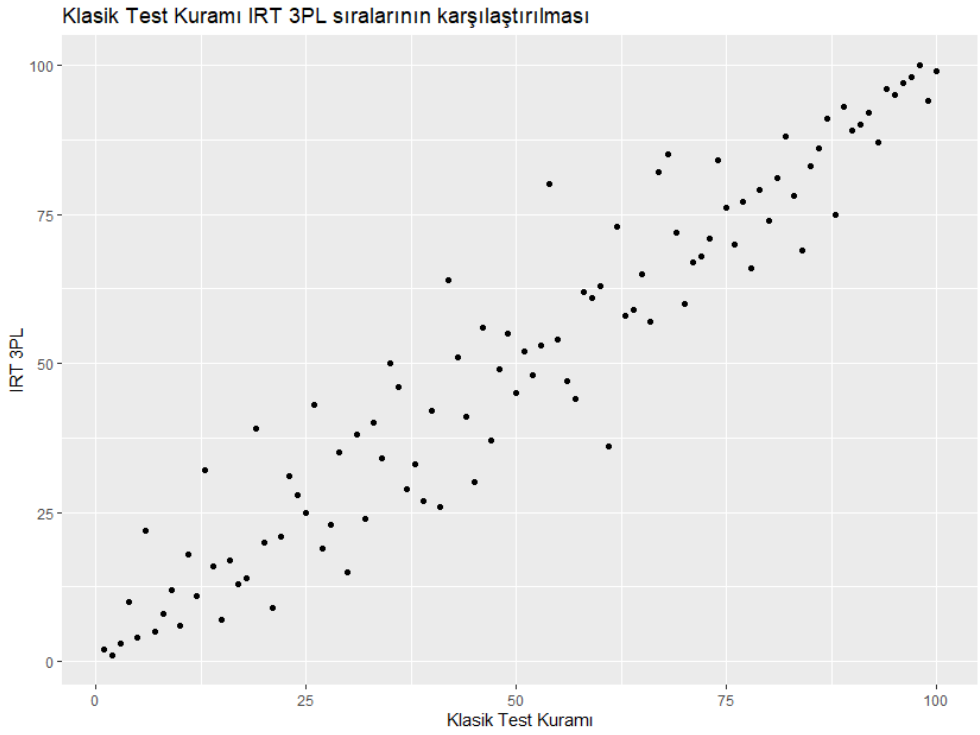


Grafik 4.15. *"Routing" yöntemine göre üretilmiş normal dağılıma sahip 10000 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 10000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

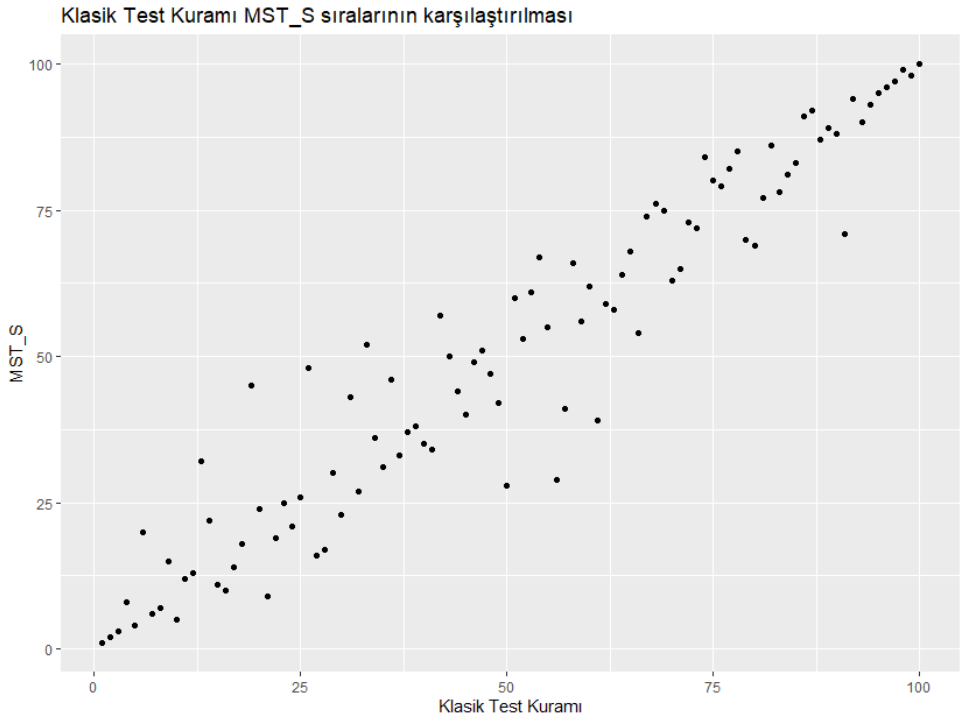
4.4.1.4. "Shaping" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PLve MST puan sıra farkları

Normal dağılıma sahip "Shaping" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



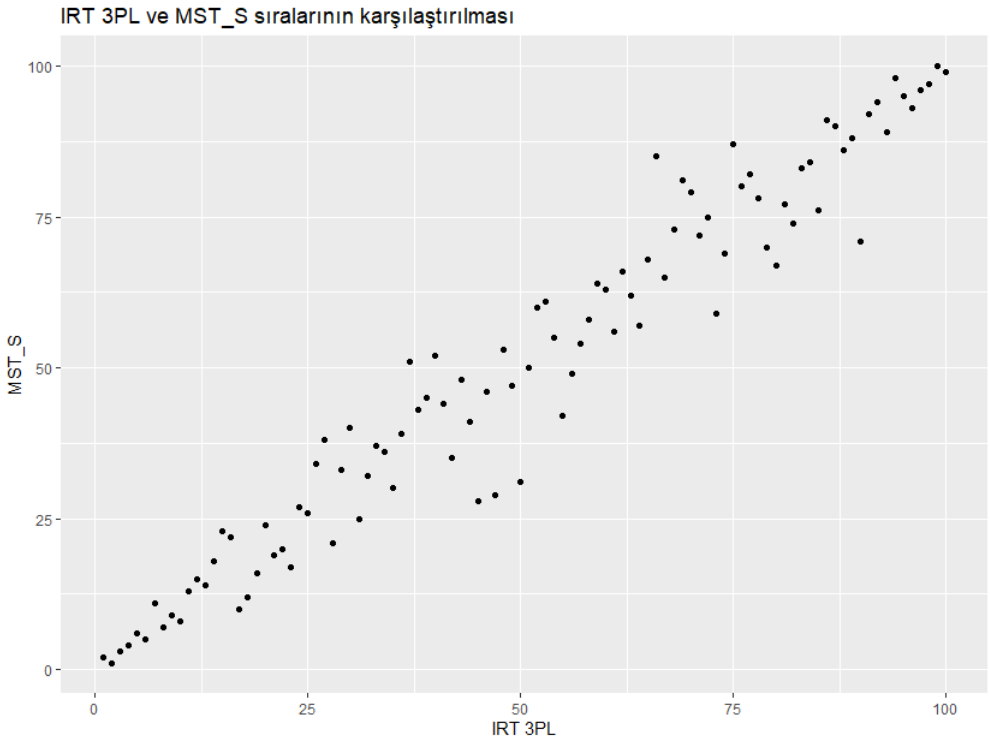
Grafik 4.16. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 100 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 100 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.17. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 100 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 100 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

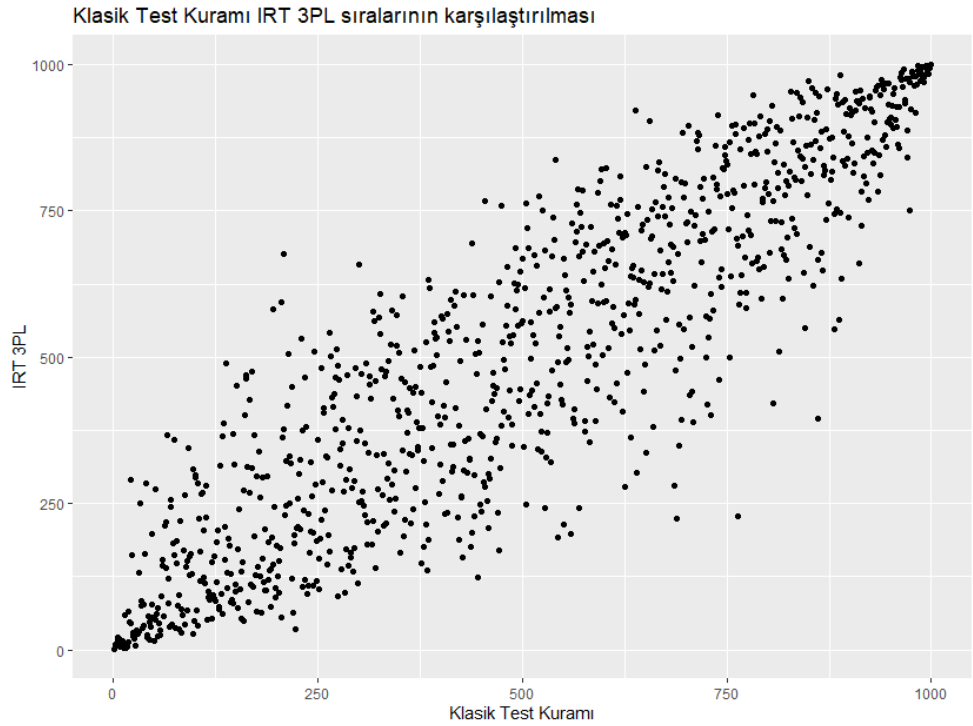


Grafik 4.18. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 100 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

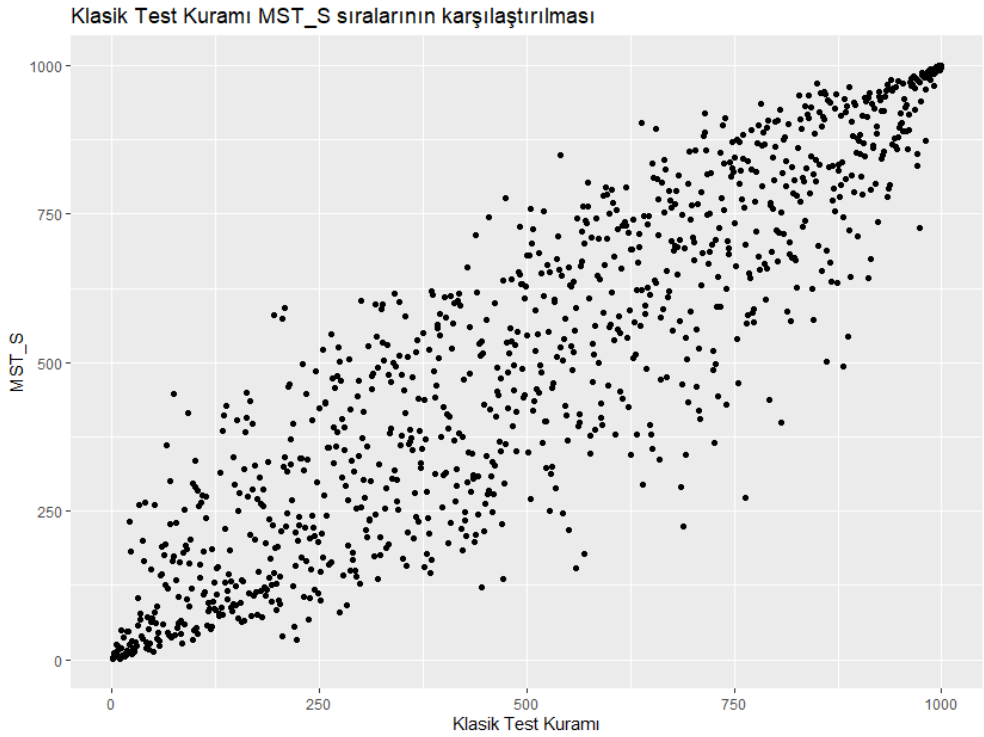
4.4.1.5. "Shaping" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal dağılıma sahip "Shaping" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



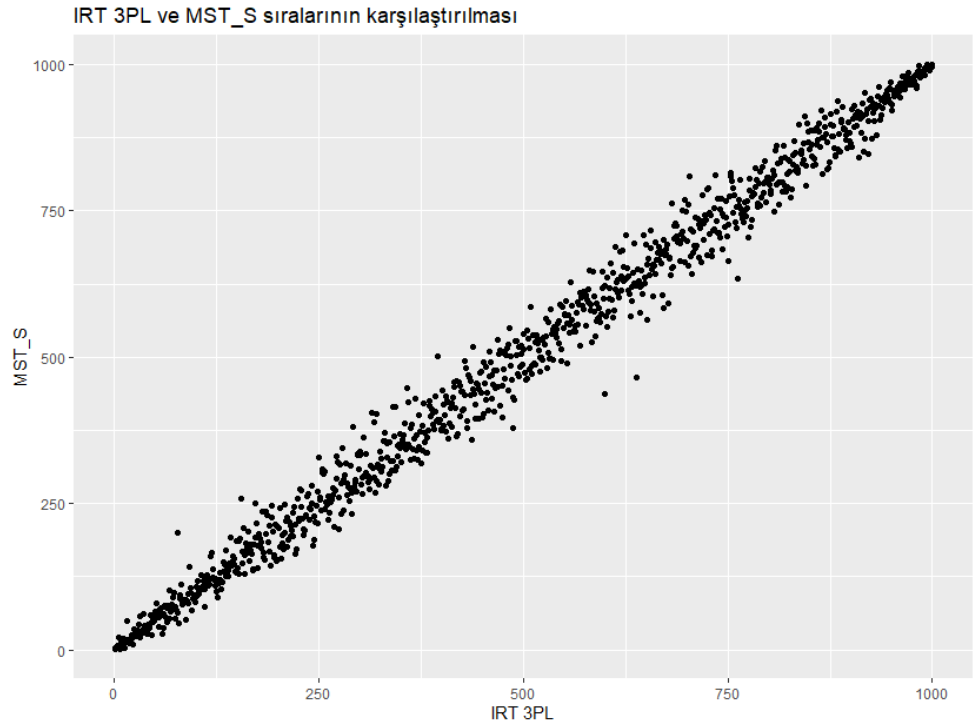
Grafik 4.19. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 1000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.20. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 1000 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

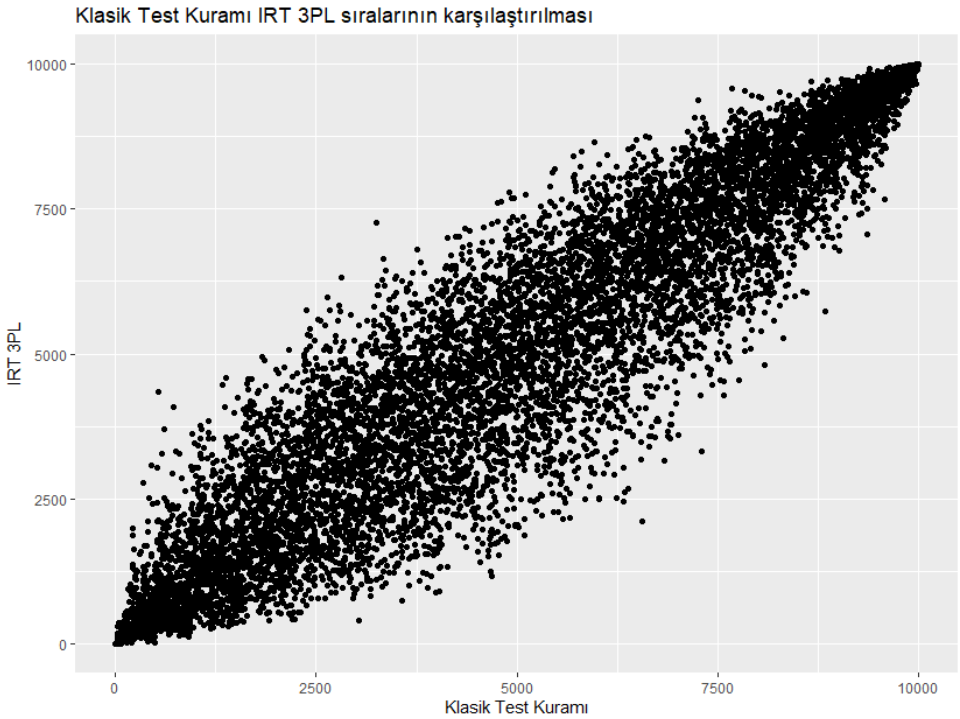


Grafik 4.21. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 1000 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

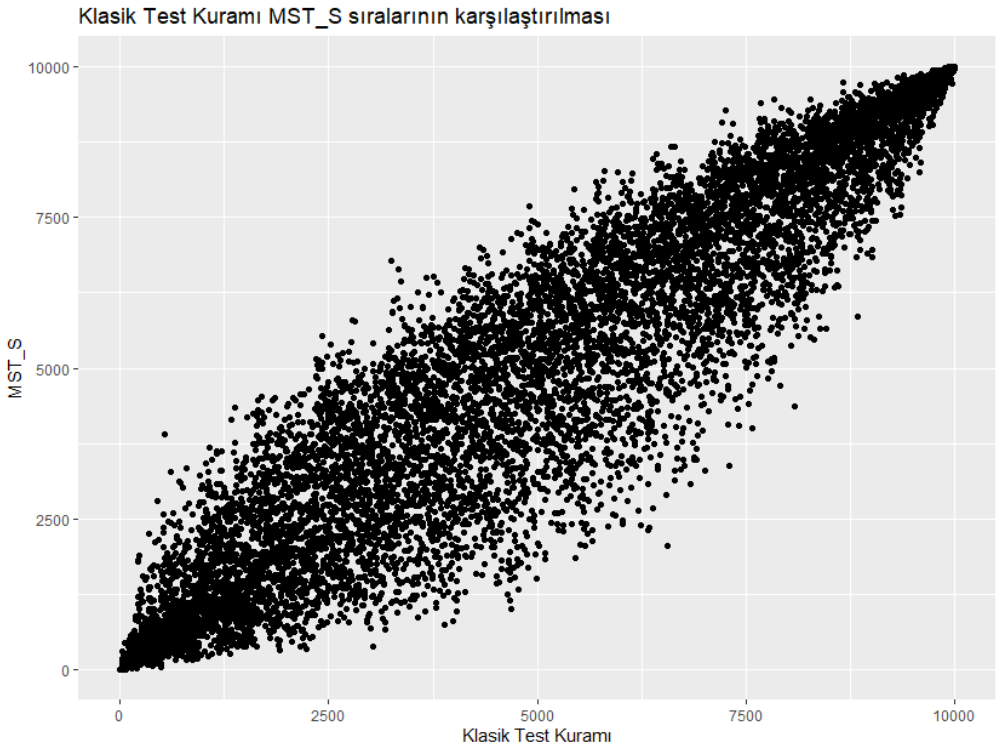
4.4.1.6. "Shaping" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal dağılıma sahip "Shaping" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



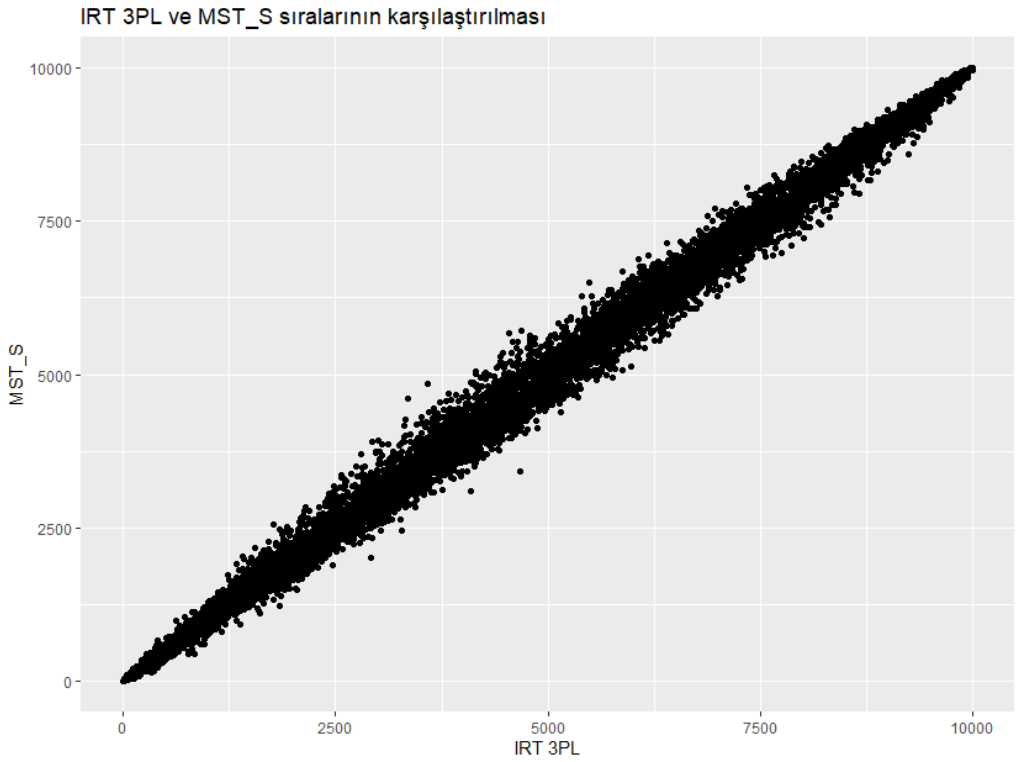
Grafik 4.22. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 10000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.23. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 10000 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.24. *"Shaping" yöntemine göre üretilmiş normal dağılıma sahip 10000 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

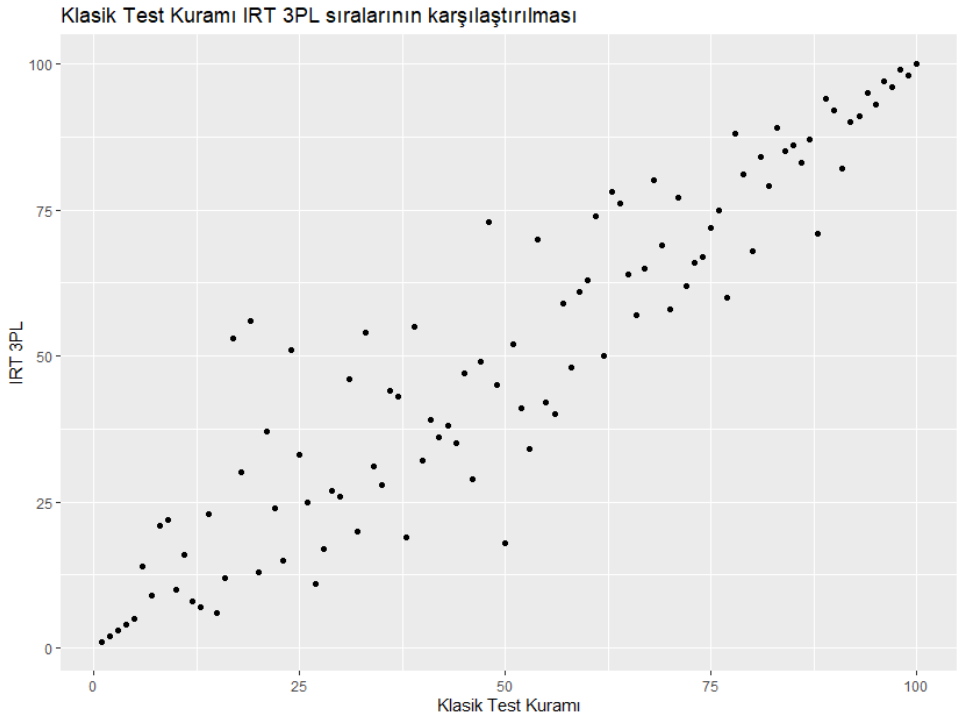
IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

4.4.2. Normal Olmayan (Sola Çarpık) Dağılıma Sahip Veri Setlerinin Kuramlara (KTK, IRT ve MST) Göre Puan Sıraları Farkı

"KTK, IRT ve MST yöntemleri ile elde edilen puanlarla yapılan sıralamalar farklılık göstermekte midir?" araştırma sorusu kapsamında gerçekleştirilen analizden normal olmayan (sola çarpık) dağılıma yönelik elde edilen sonuçlar şu şekildedir:

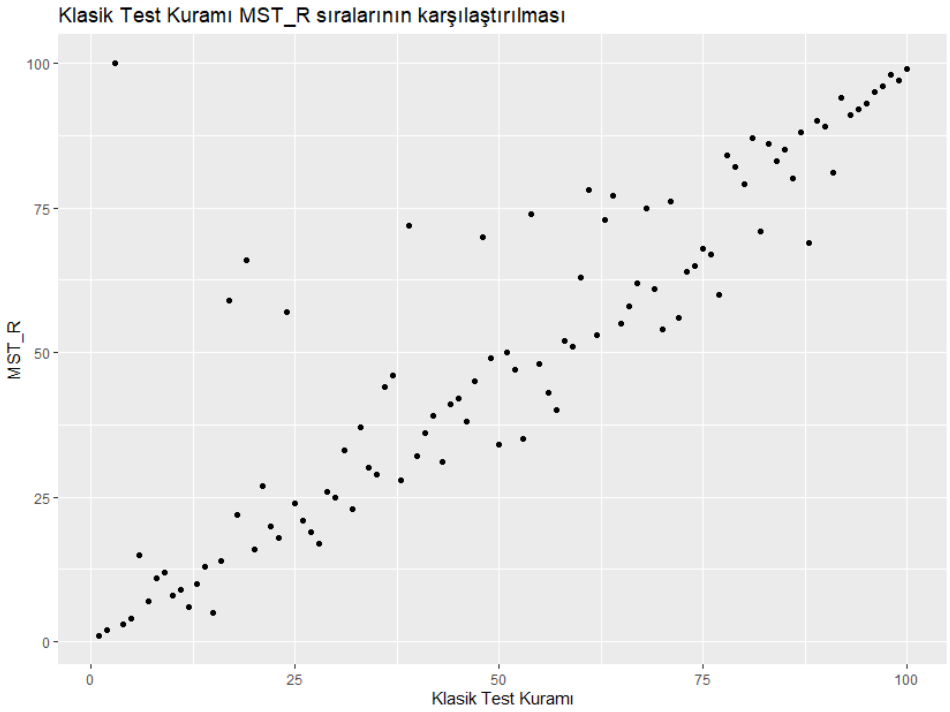
4.4.2.1. "Routing" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sola çarpık) dağılıma sahip "Routing" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



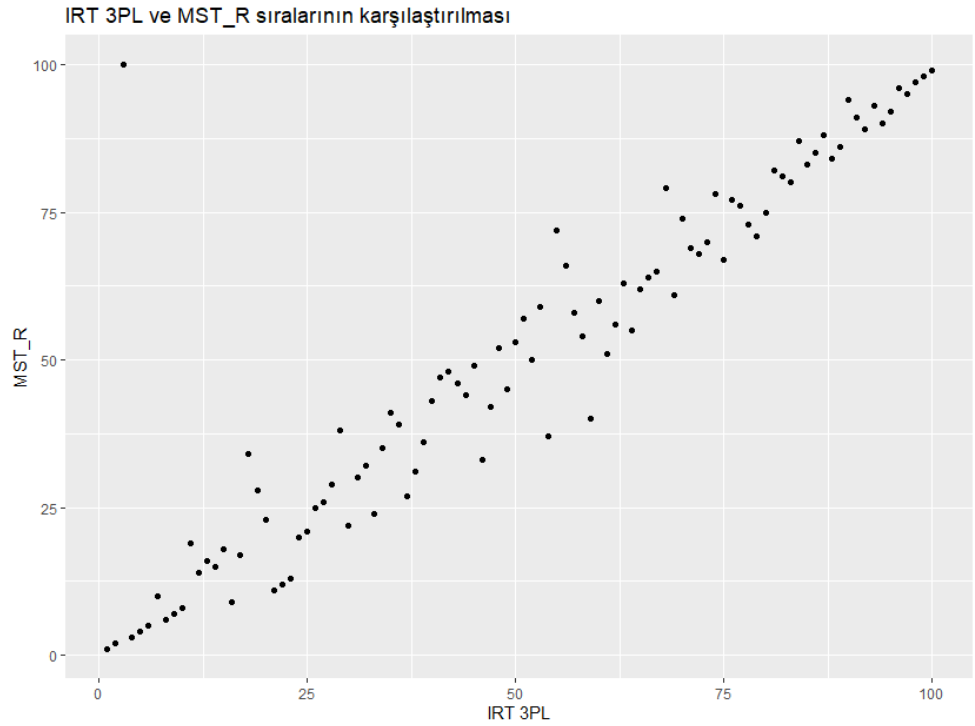
Grafik 4.25. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 100 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.26. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 100 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

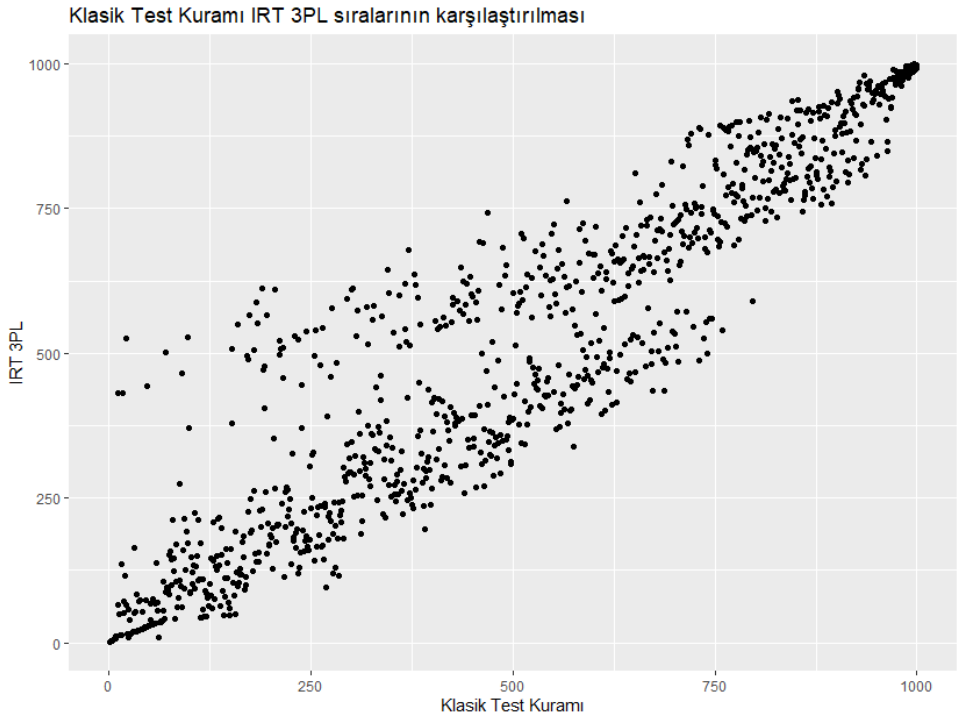


Grafik 4.27. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 100 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

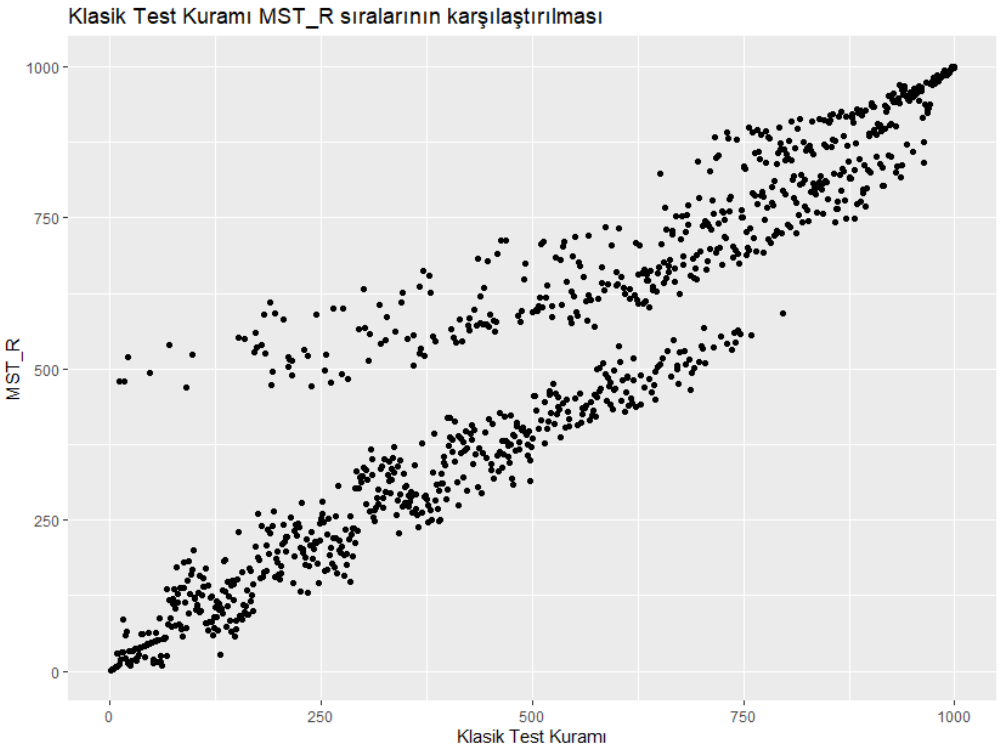
4.4.2.2. "Routing" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sola çarpık) dağılıma sahip "Routing" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



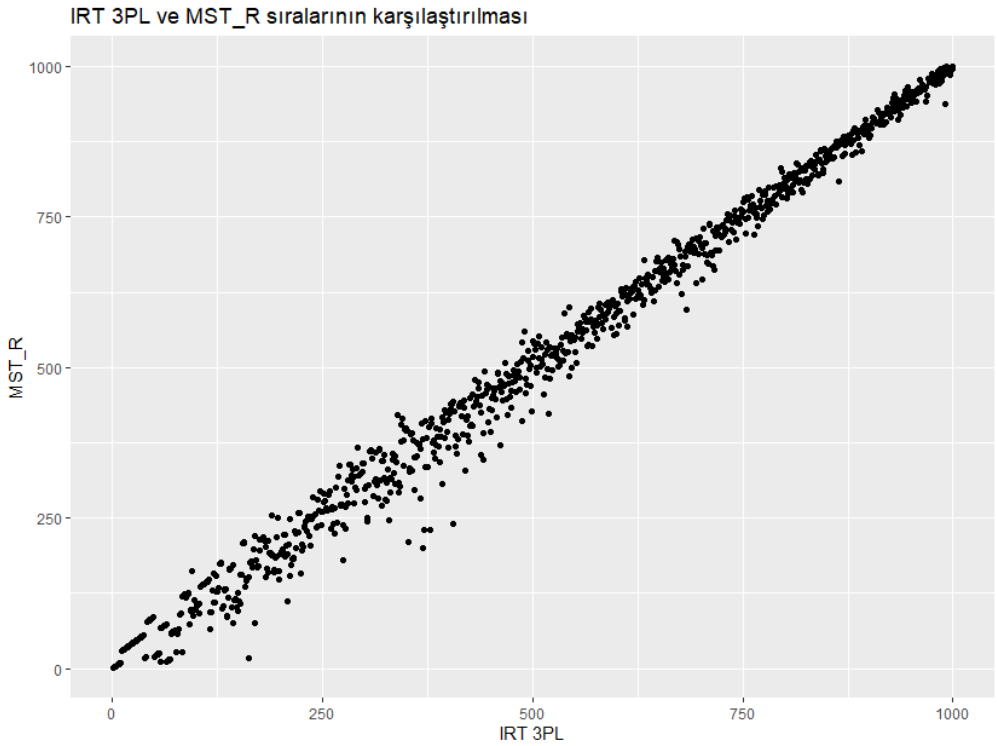
Grafik 4.28. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.29. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

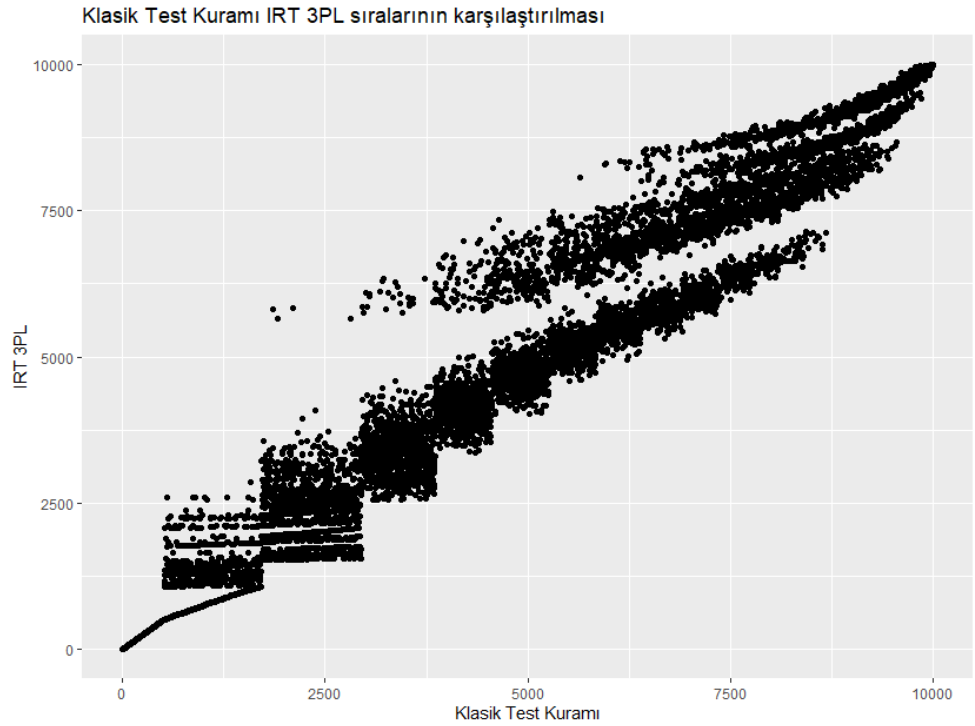


Grafik 4.30. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 1000 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 1000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

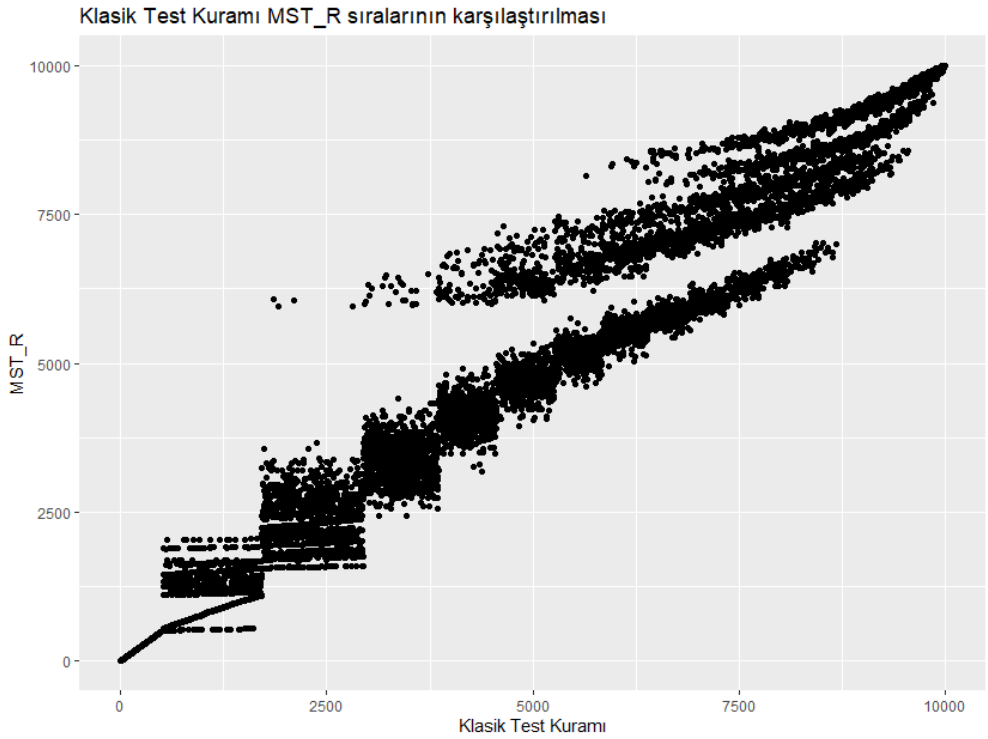
4.4.2.3. "Routing" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sola çarpık) dağılıma sahip "Routing" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



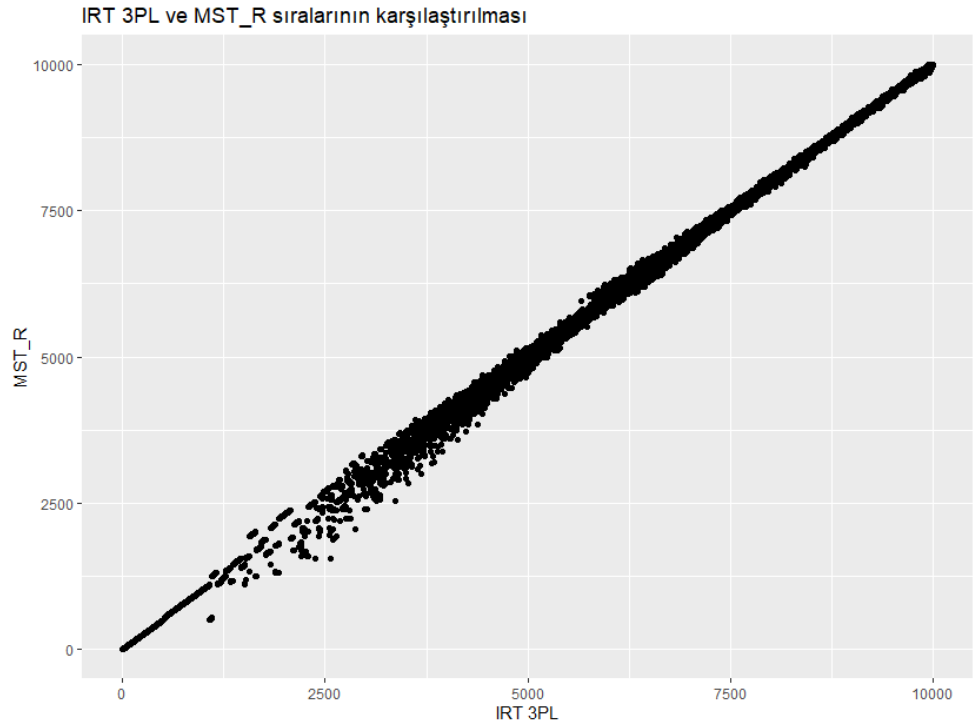
Grafik 4.31. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.32. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

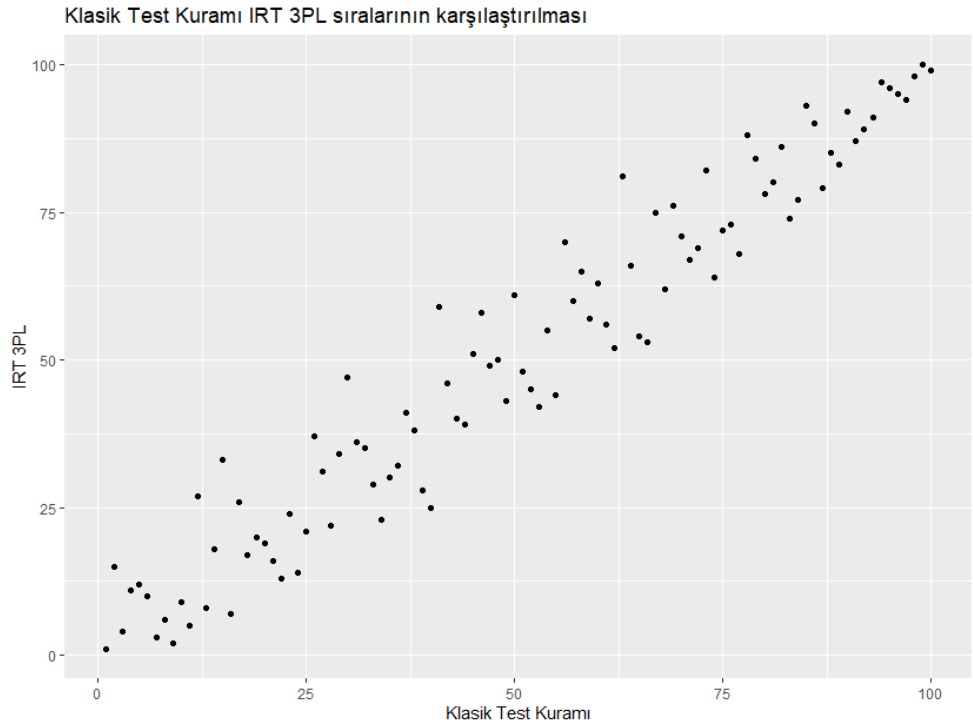


Grafik 4.33. *"Routing" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 10000 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 10000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

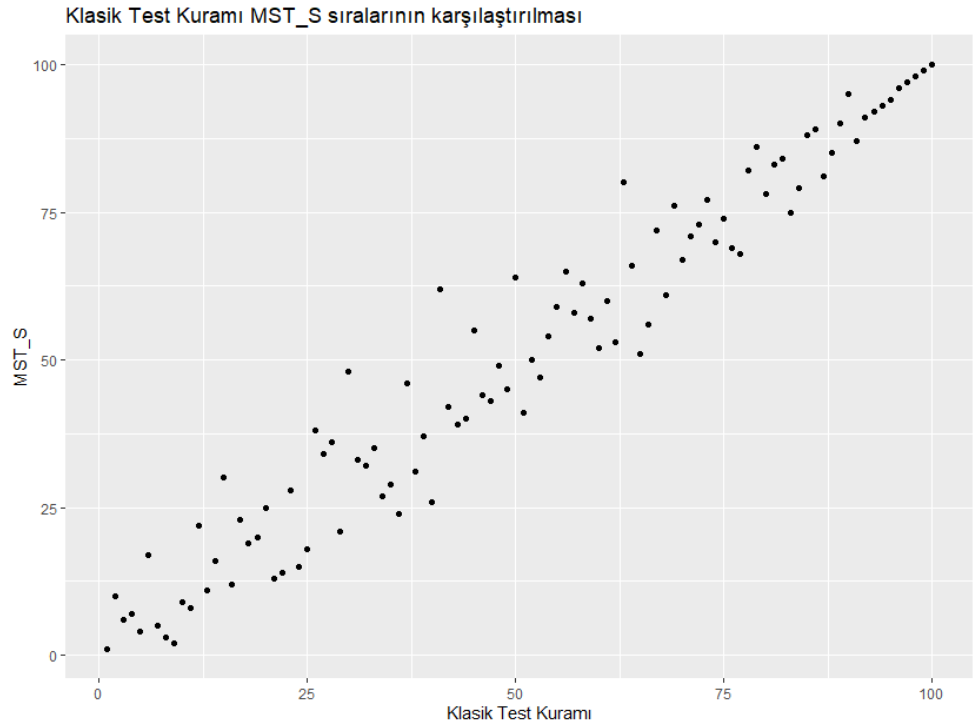
4.4.2.4. "Shaping" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sola çarpık) dağılıma sahip "Shaping" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



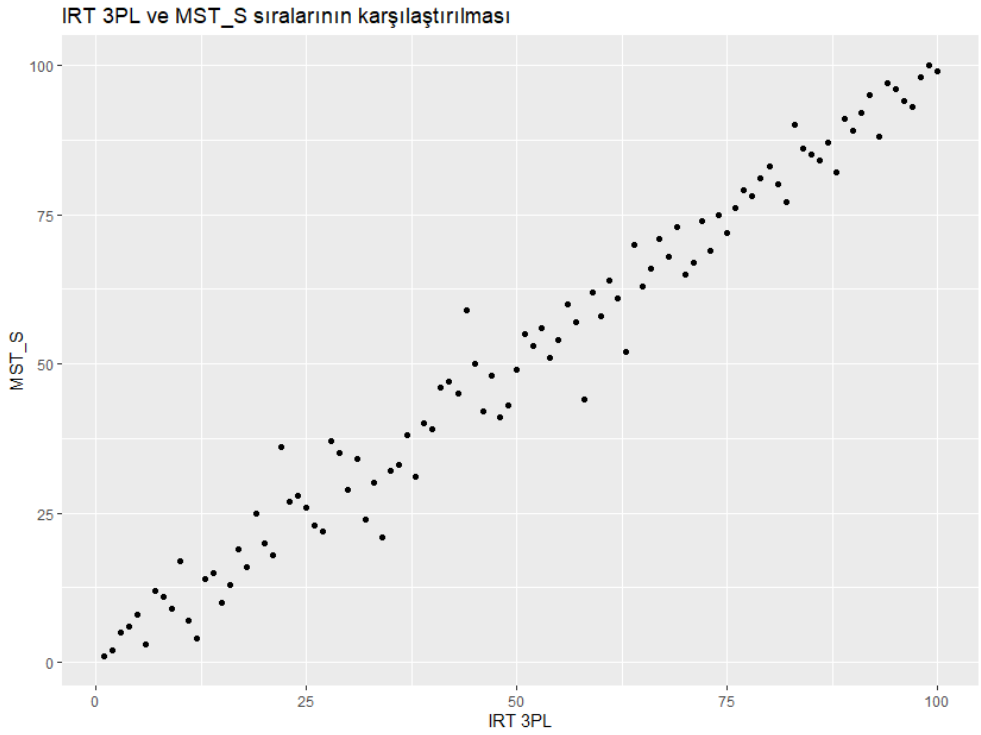
Grafik 4.34. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 100 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.35. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 100 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

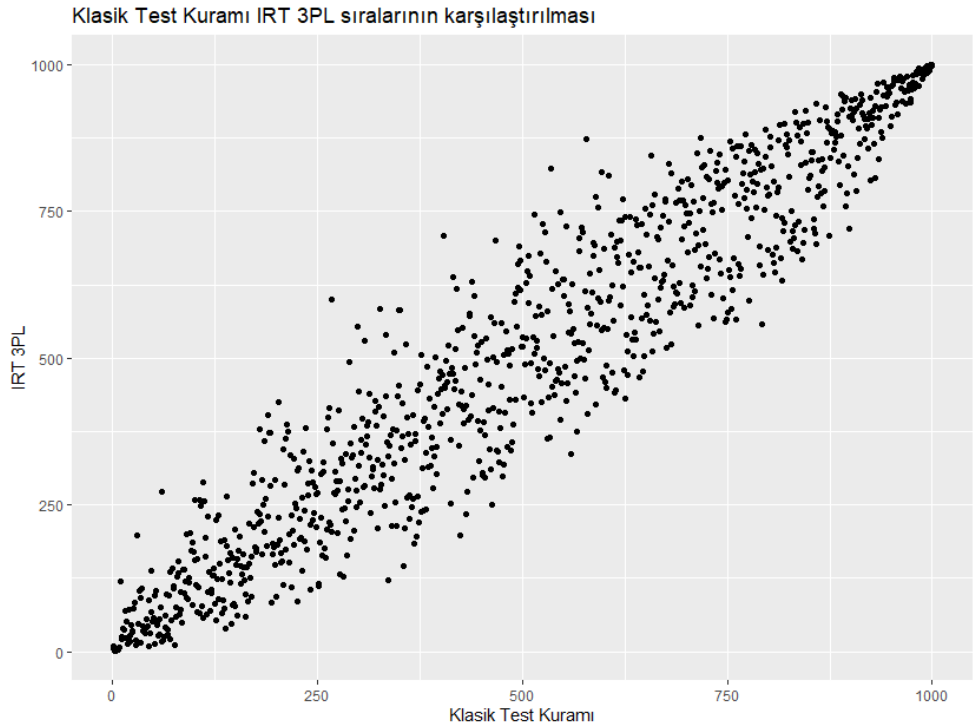


Grafik 4.36. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 100 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

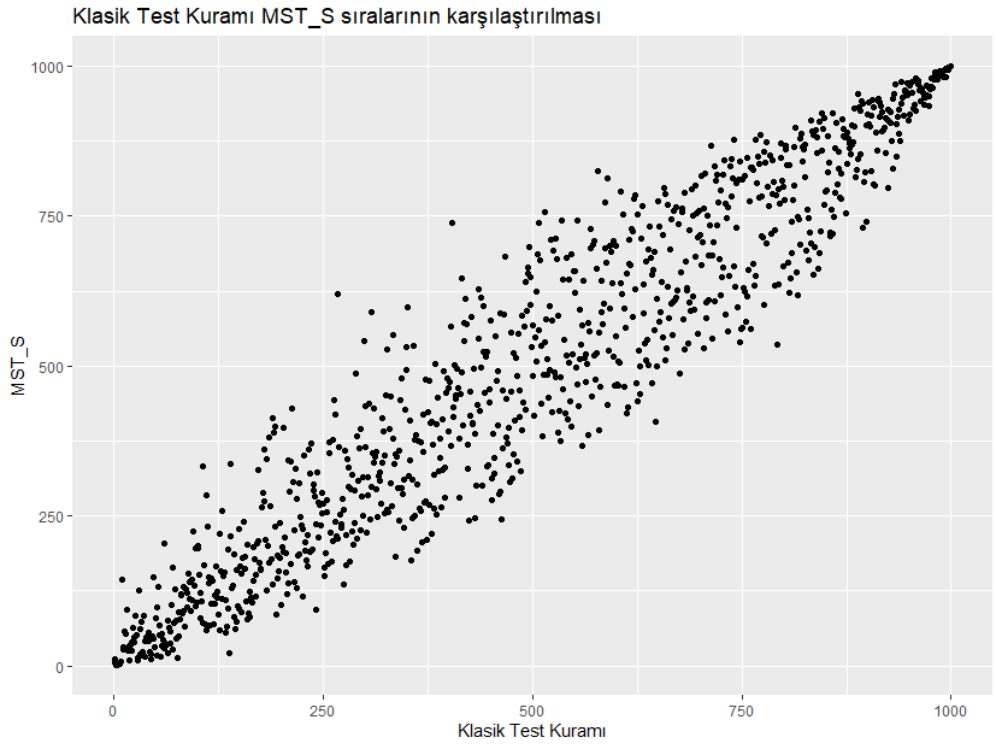
4.4.2.5. "Shaping" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sola çarpık) dağılıma sahip "Shaping" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



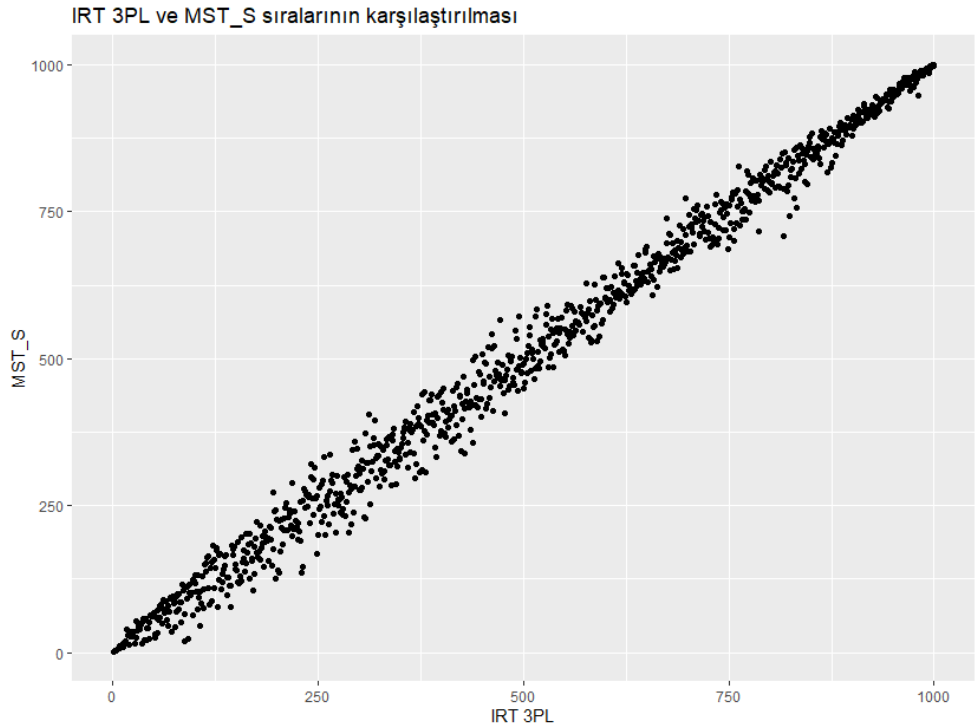
Grafik 4.37. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.38. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

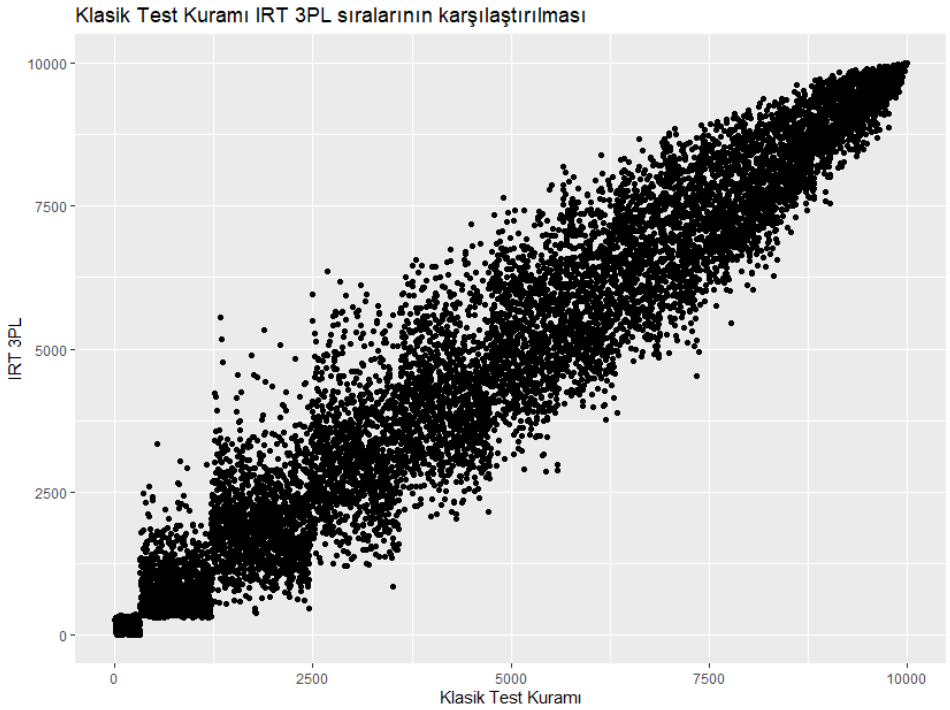


Grafik 4.39. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 1000 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 1000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

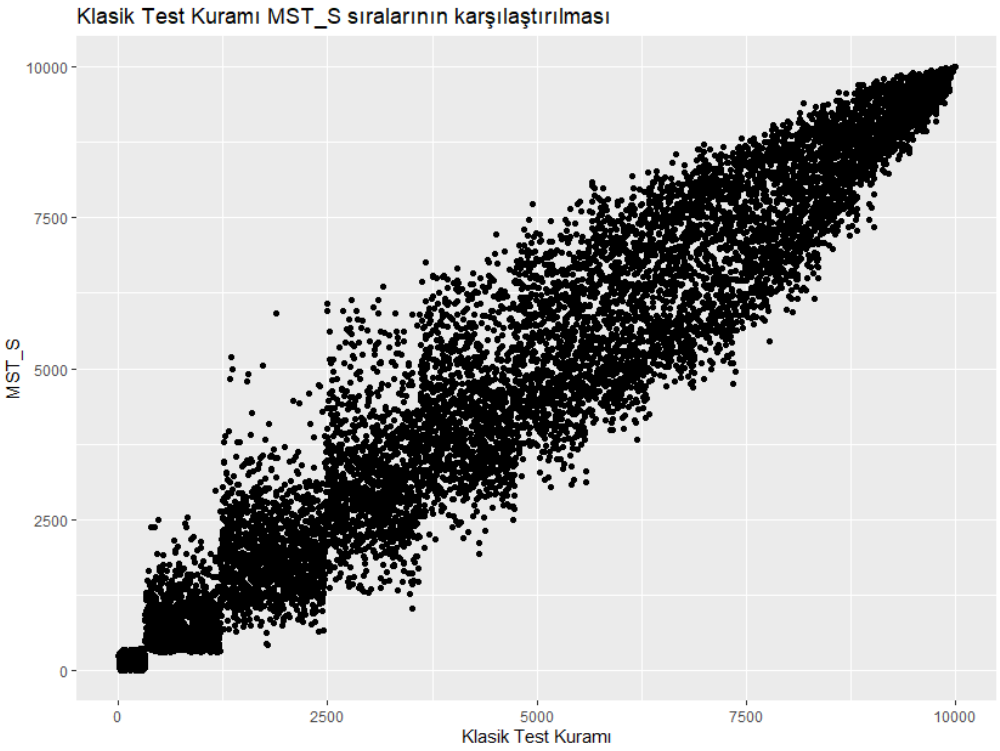
4.4.2.6. "Shaping" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sola çarpık) dağılıma sahip "Shaping" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



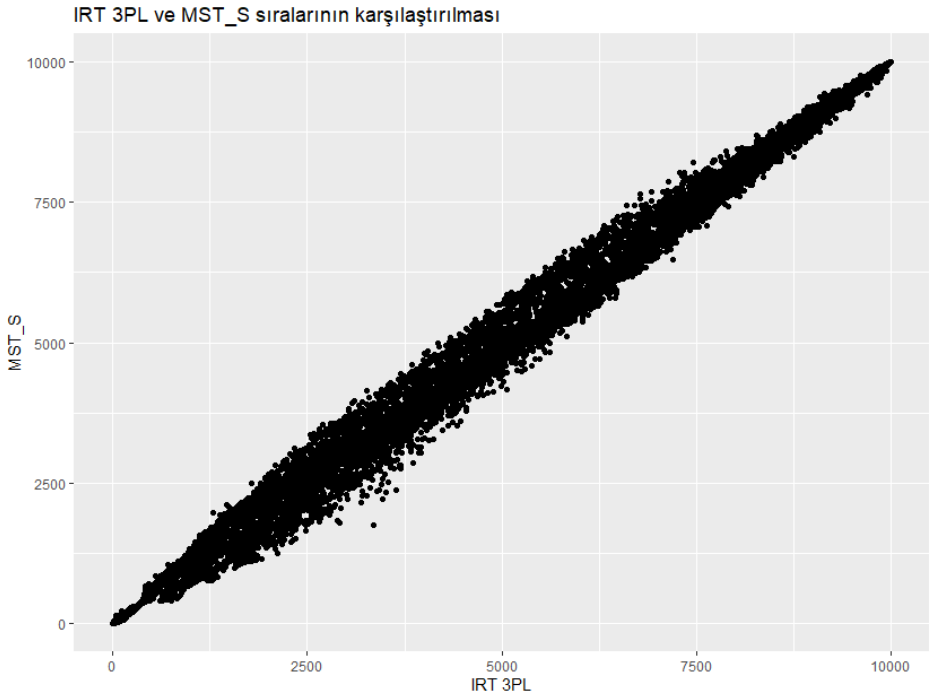
Grafik 4.40. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.41. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.42. *"Shaping" yöntemine göre üretilmiş normal olmayan (sola çarpık) dağılıma sahip 10000 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

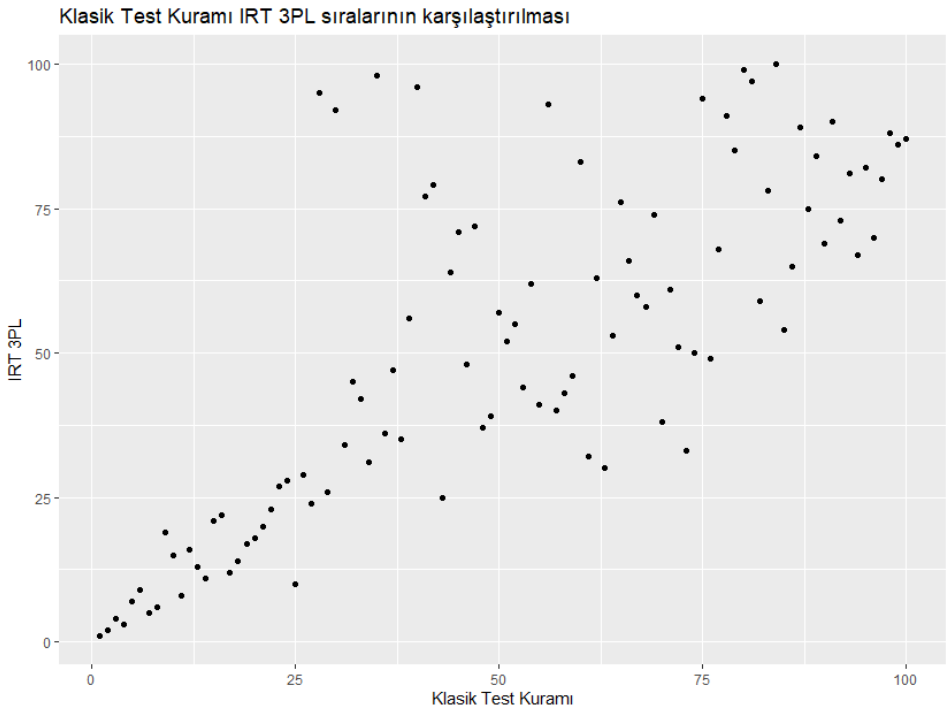
IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 10000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

4.4.3. Normal Olmayan (Sağa Çarpık) Dağılıma Sahip Veri Setlerinin Kuramlara (KTK, IRT ve MST) Göre Puan Sıraları Farkı

"KTK, IRT ve MST yöntemleri ile elde edilen puanlarla yapılan sıralamalar farklılık göstermekte midir?" araştırma sorusu kapsamında gerçekleştirilen analizden normal olmayan (sağa çarpık) dağılıma yönelik elde edilen sonuçlar şu şekildedir:

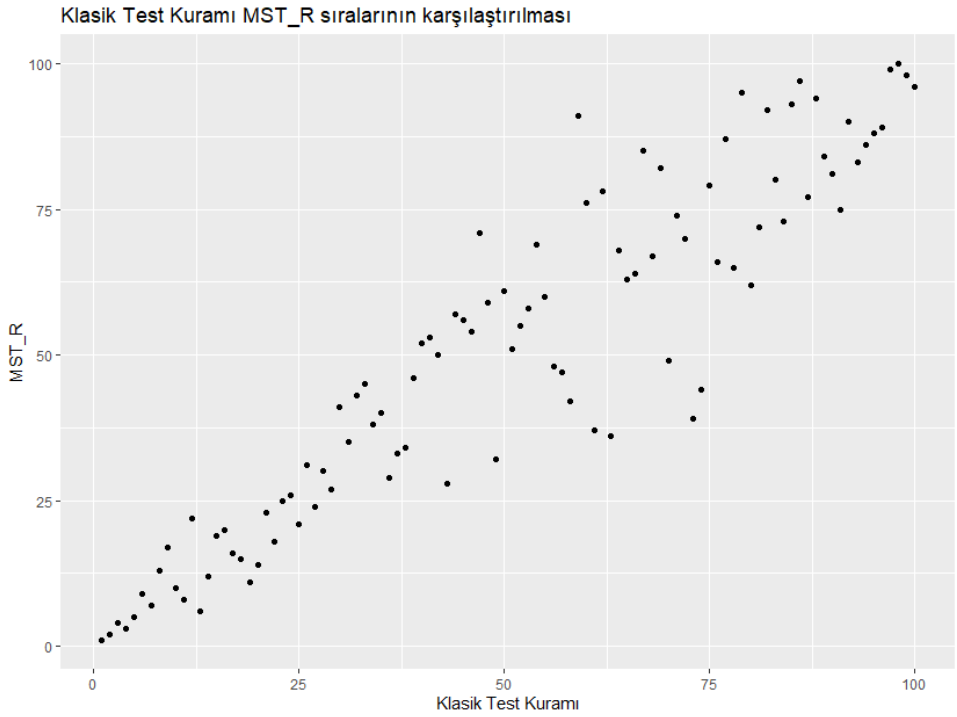
*4.4.3.1. "Routing" yöntemine göre üretil*miş *100 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları*

Normal olmayan (sağa çarpık) dağılıma sahip "Routing" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



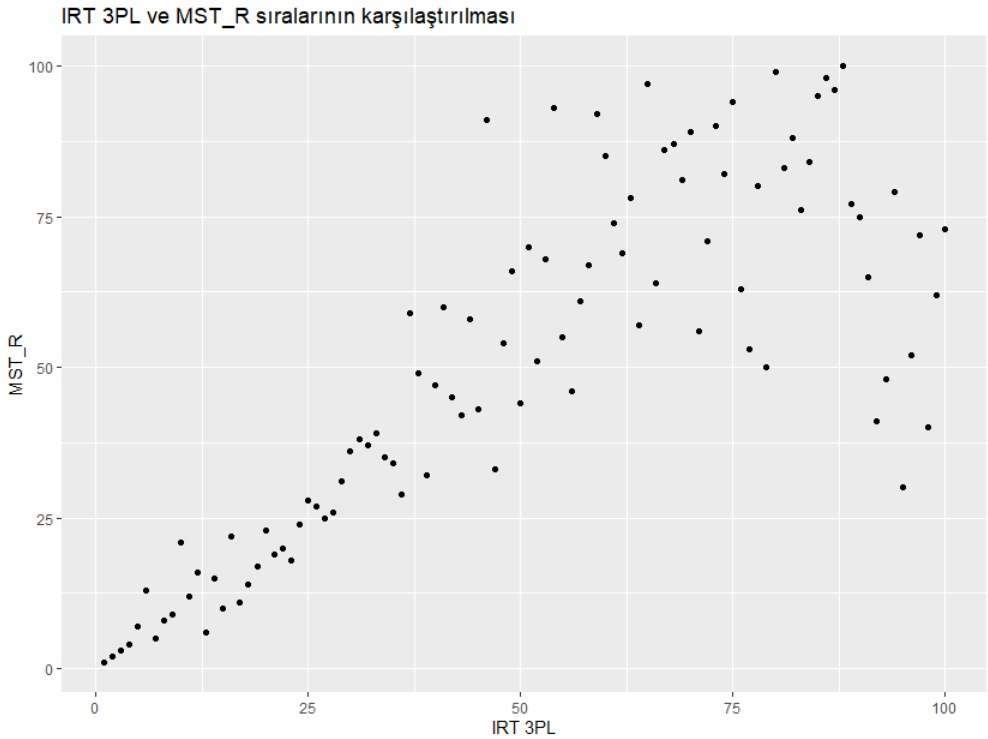
Grafik 4.43. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 100 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının büyük ölçüde değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.44. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 100 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

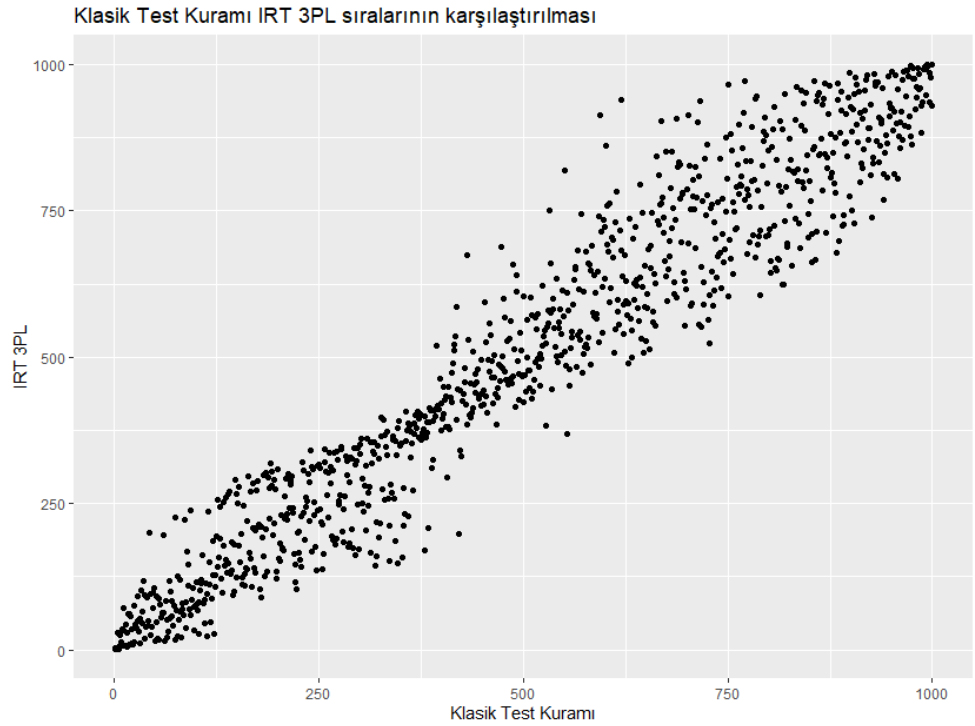


Grafik 4.45. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 100 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmıştır.

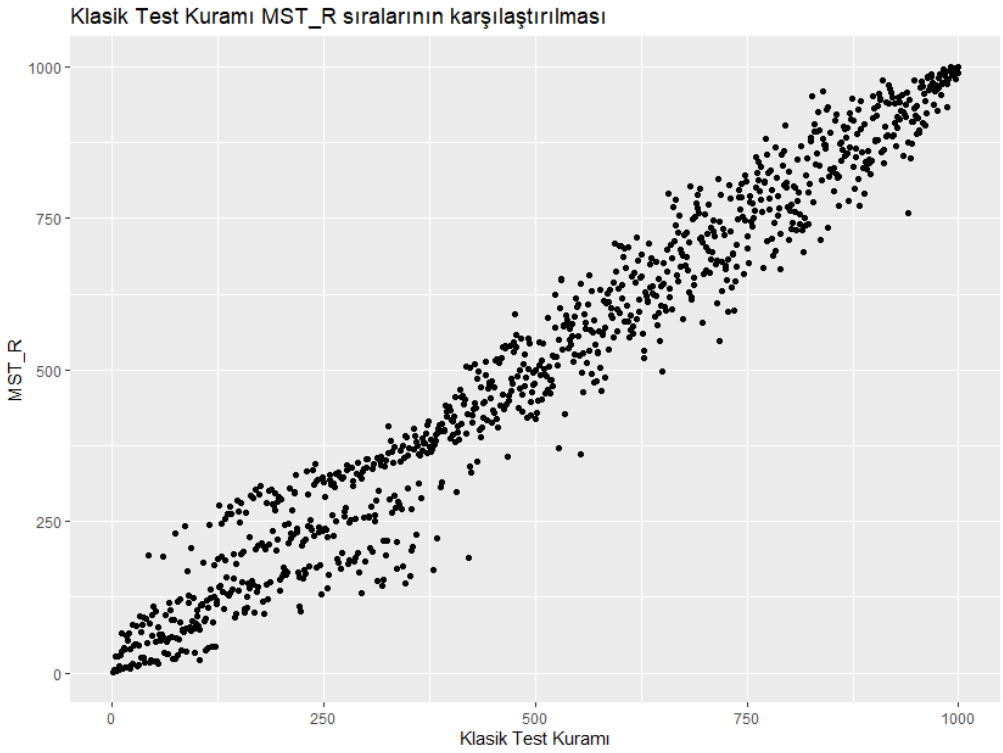
4.4.3.2. "Routing" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sağa çarpık) dağılıma sahip "Routing" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



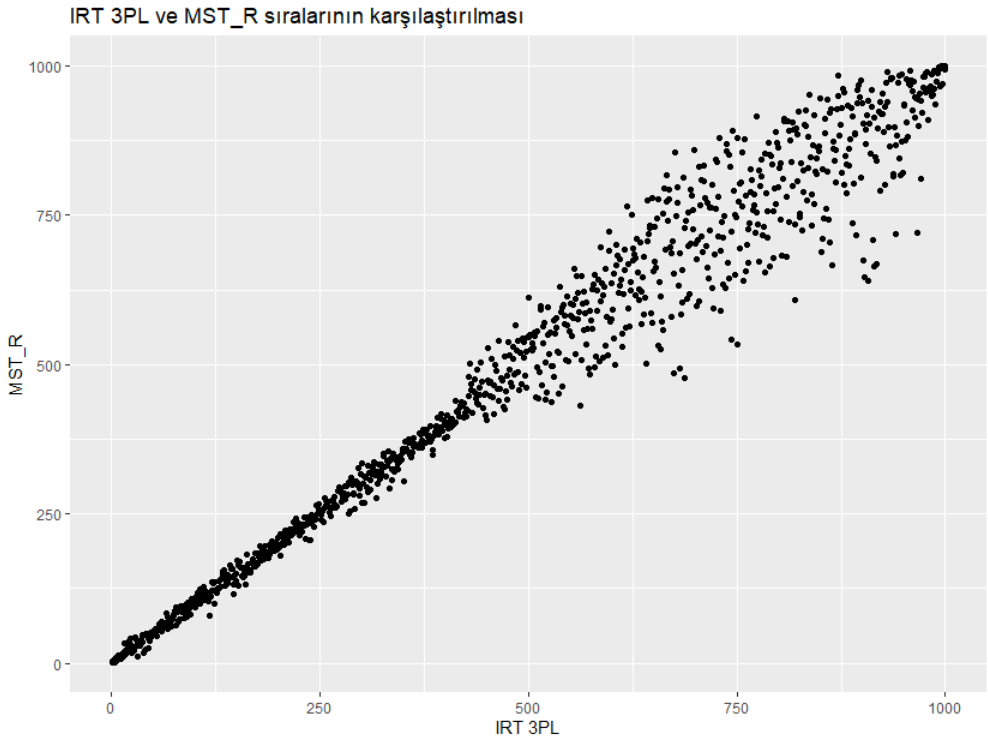
Grafik 4.46. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.47. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

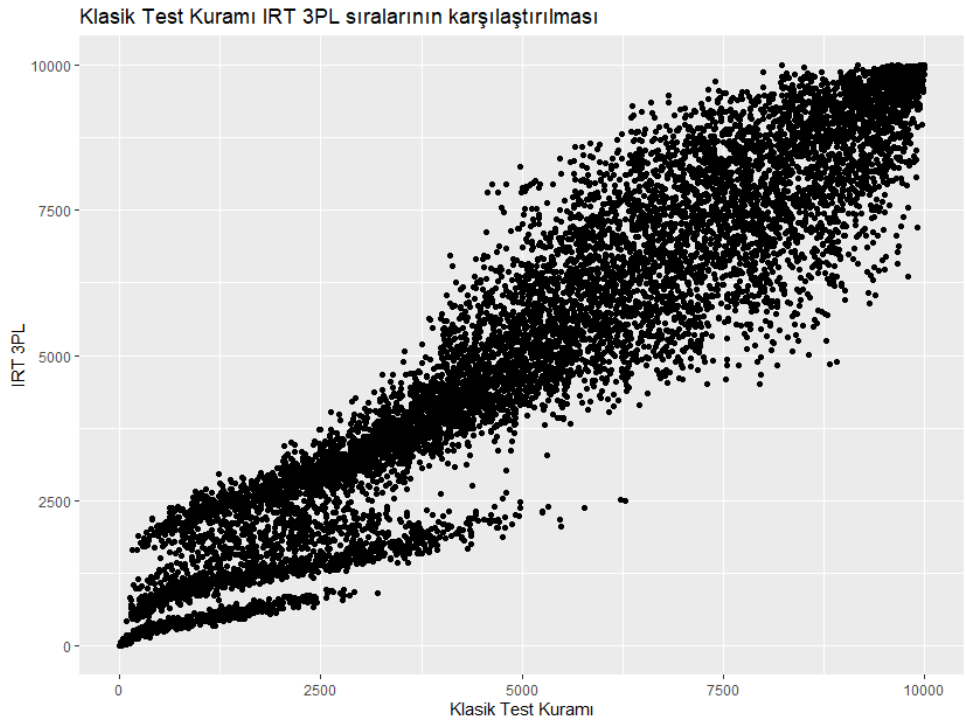


Grafik 4.48. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 1000 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 1000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmakla birlikte son sıralarda saçılımın arttığı tespit edilmiştir.

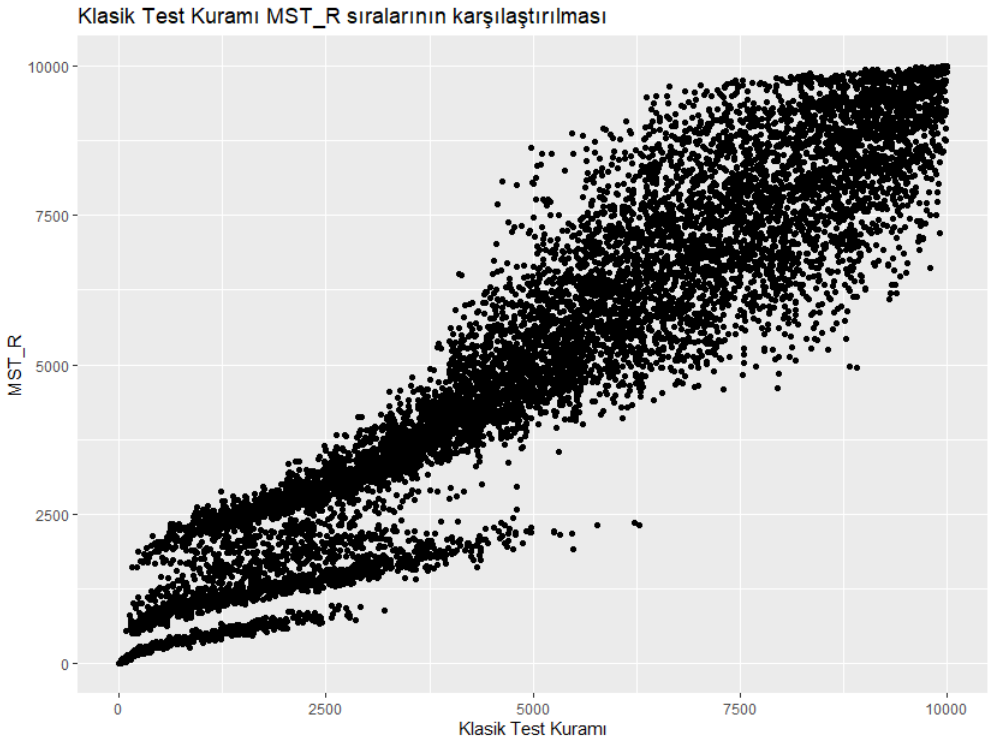
4.4.3.3. "Routing" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sağa çarpık) dağılıma sahip "Routing" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



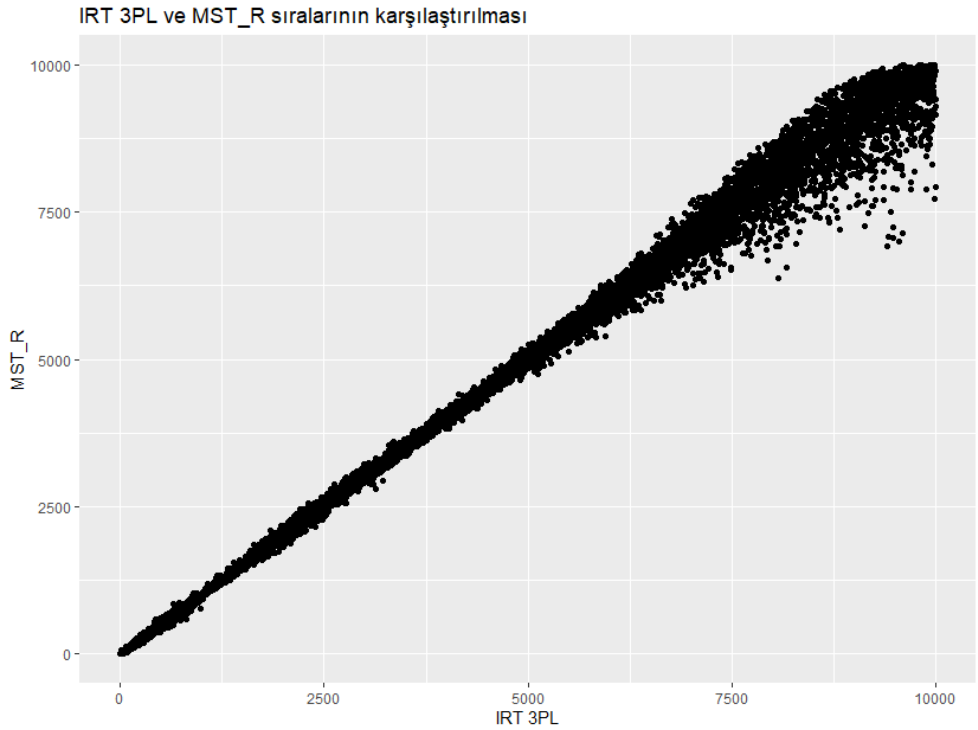
Grafik 4.49. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.50. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve MST-R puan sıralarının karşılaştırması*

KTK ve MST-R puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

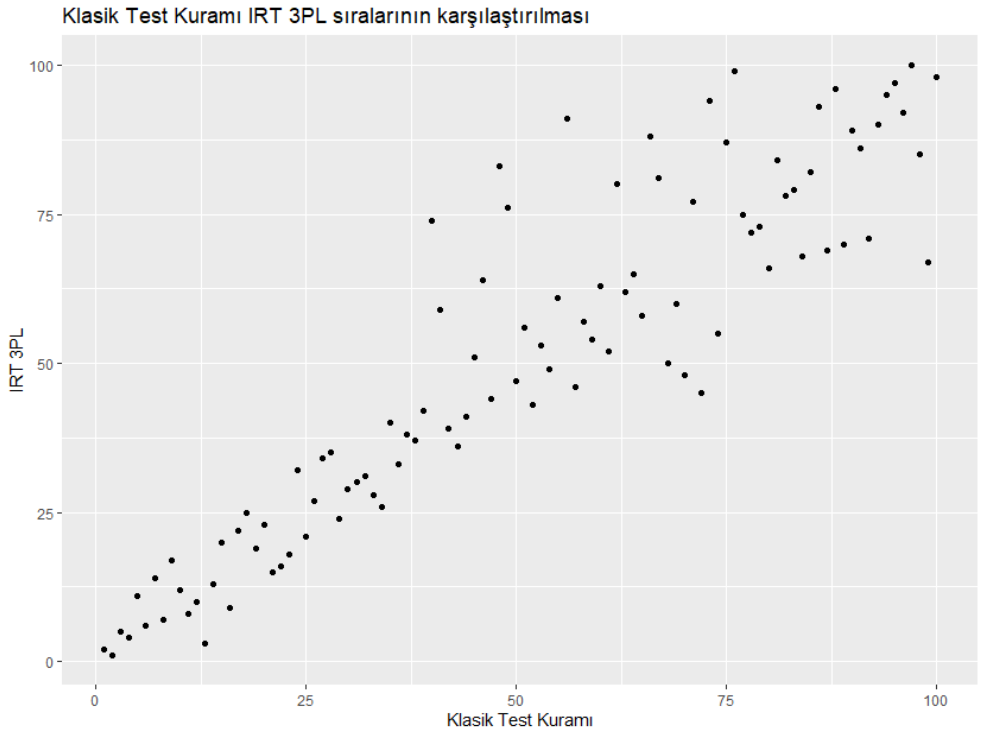


Grafik 4.51. *"Routing" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 10000 kişilik veri seti için IRT 3PL ve MST-R puan sıralarının karşılaştırması*

IRT 3PL ve MST-R yöntemlerinin puan sıralarının karşılaştırıldığı 10000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-R karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-R yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-R) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmakla birlikte son sıralarda saçılımın arttığı tespit edilmiştir.

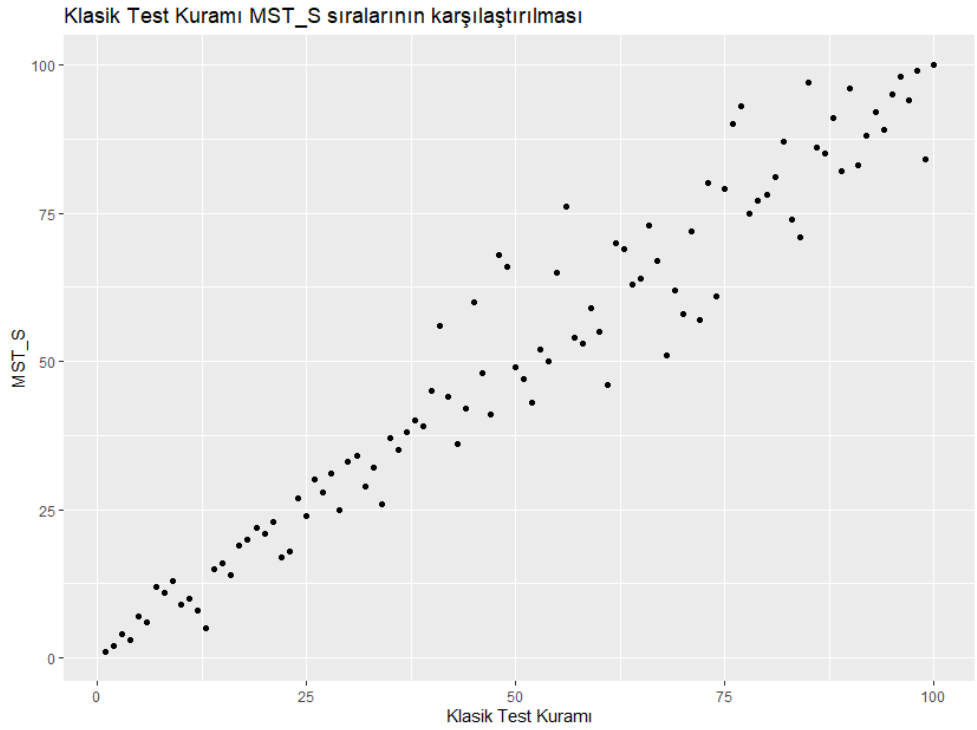
4.4.3.4. "Shaping" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sağa çarpık) dağılıma sahip "Shaping" yöntemine göre üretilmiş 100 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



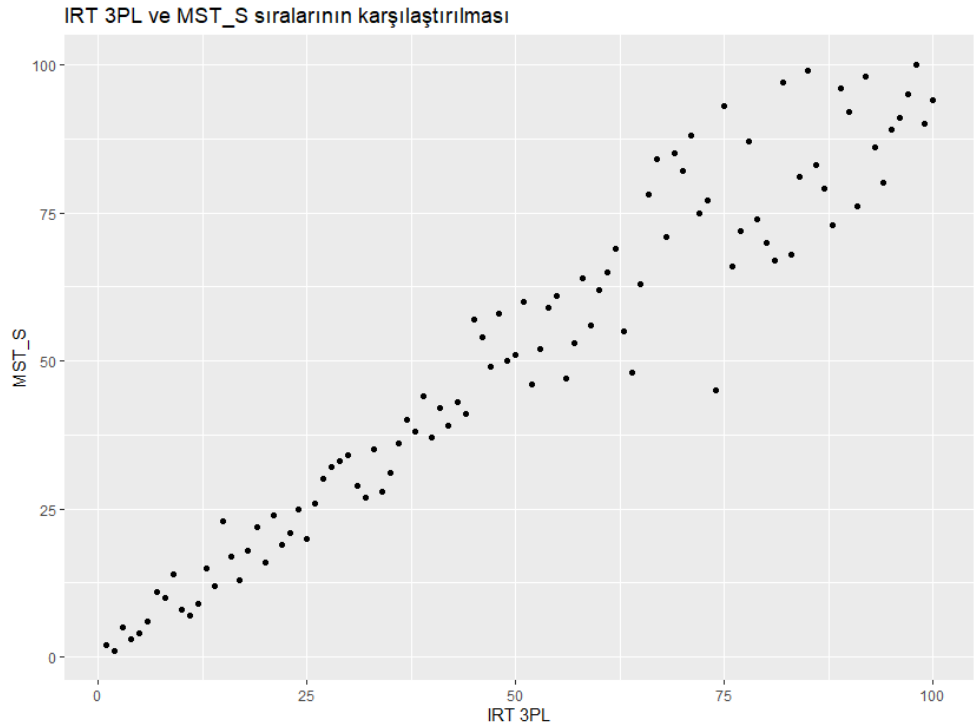
Grafik 4.52. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 100 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.53. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 100 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 100 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

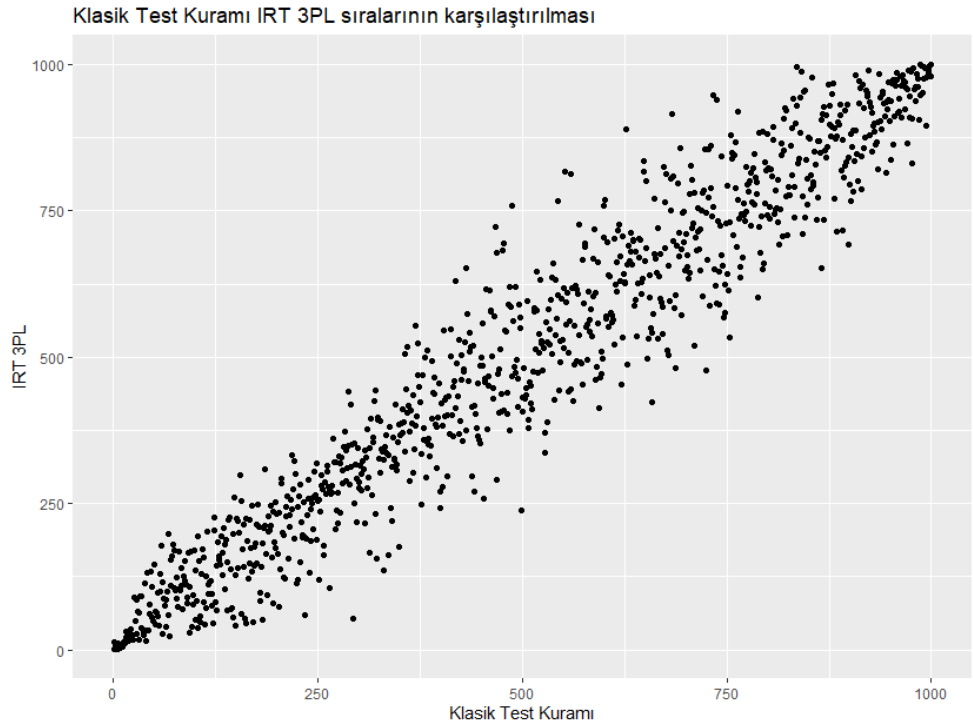


Grafik 4.54. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 100 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 100 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmakla birlikte son sıralarda saçılımın arttığı tespit edilmiştir.

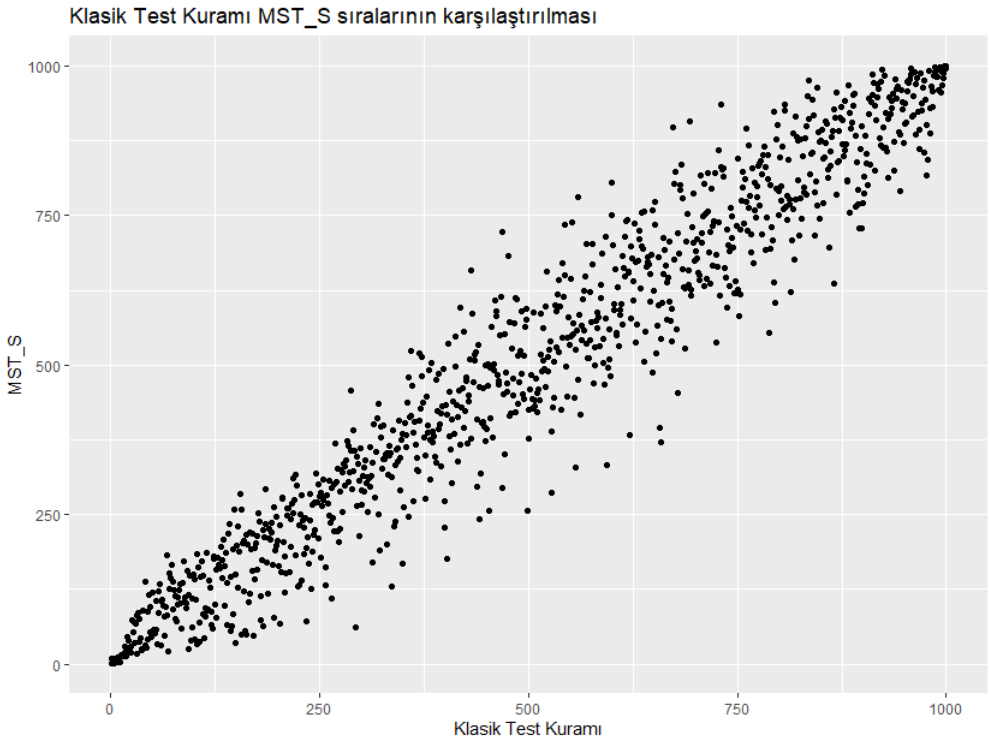
4.4.3.5. "Shaping" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sağa çarpık) dağılıma sahip "Shaping" yöntemine göre üretilmiş 1000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



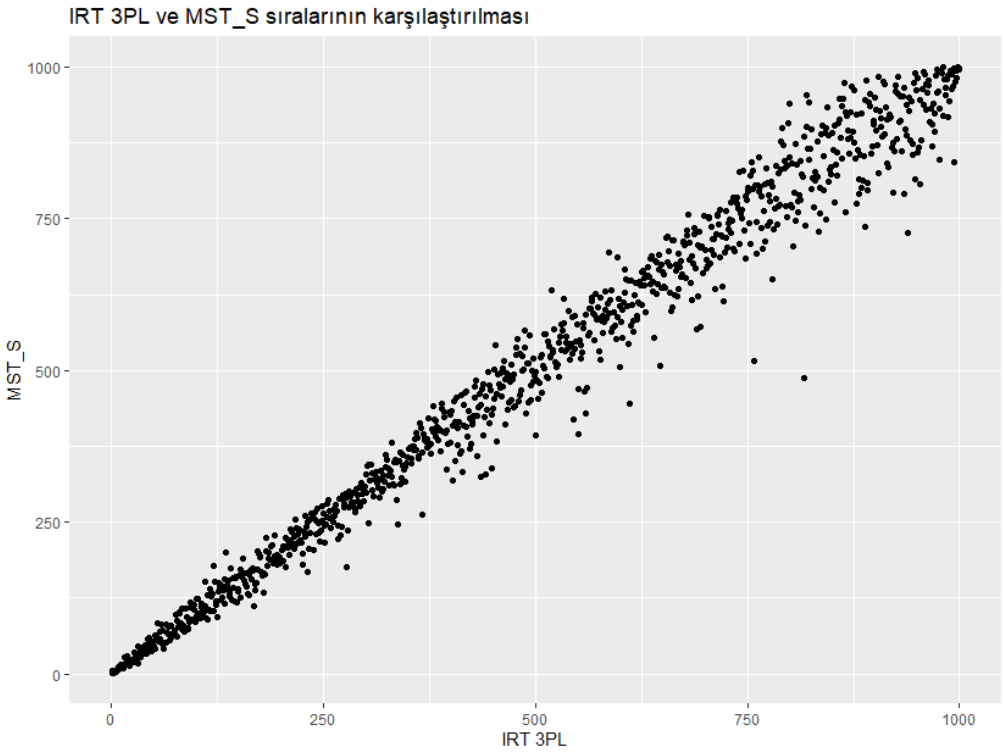
Grafik 4.55. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.56. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 1000 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 1000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.

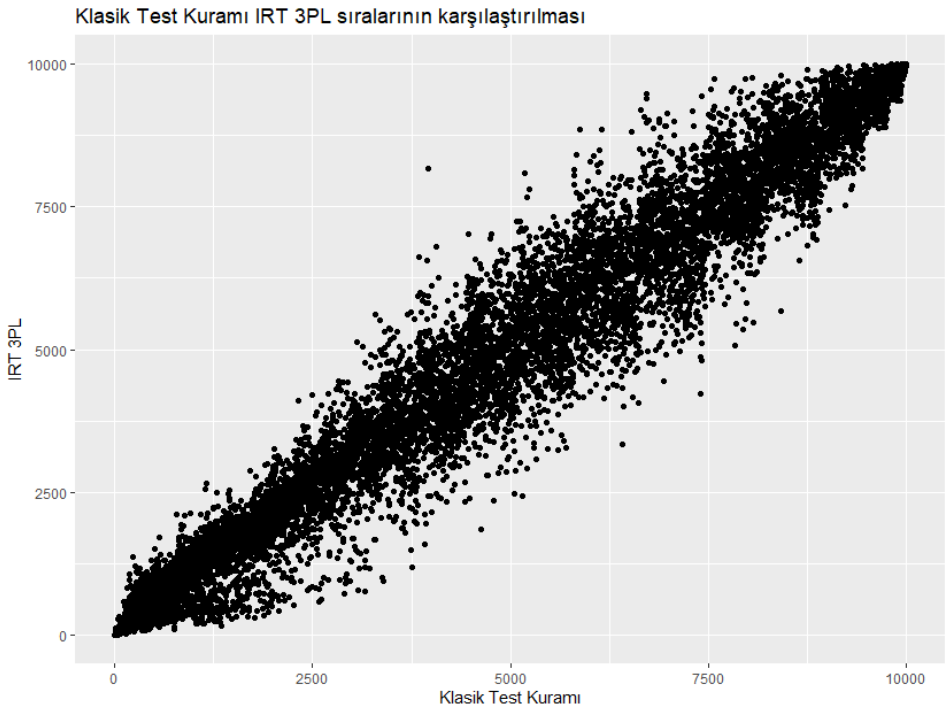


Grafik 4.57. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 1000 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 1000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmakla birlikte son sıralarda saçılımın arttığı tespit edilmiştir.

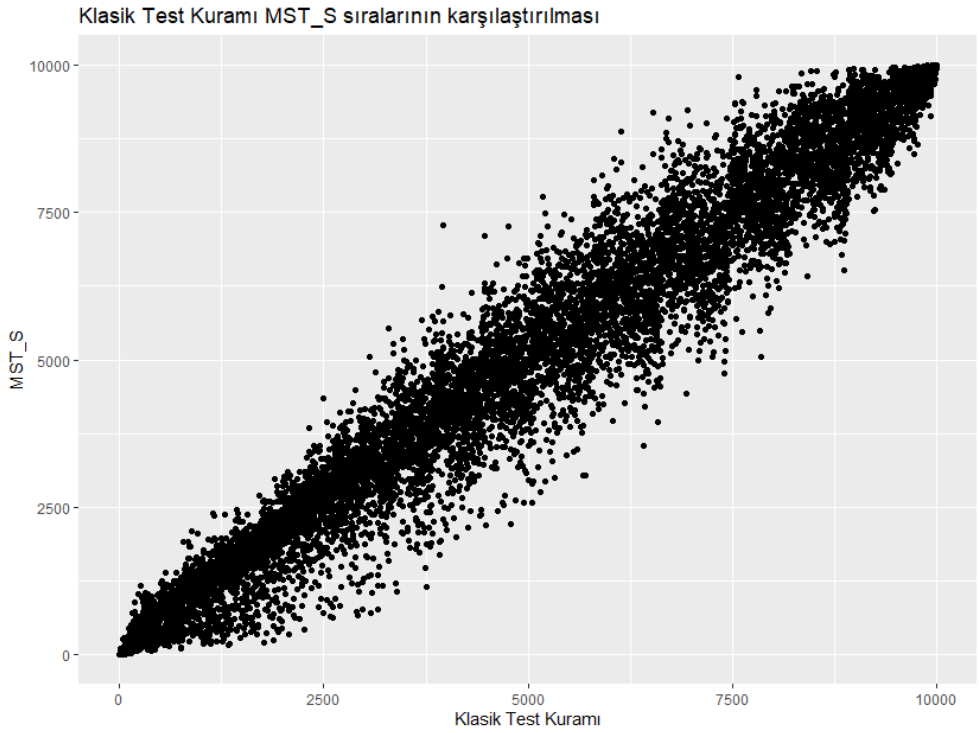
4.4.3.6. "Shaping" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST puan sıra farkları

Normal olmayan (sağa çarpık) dağılıma sahip "Shaping" yöntemine göre üretilmiş 10000 kişilik veri seti için KTK, IRT 3PL ve MST yöntemlerinin puan sıralamaları birbiri ile karılaştırıldığı grafikler aşağıda detaylı olarak sunulmuştur:



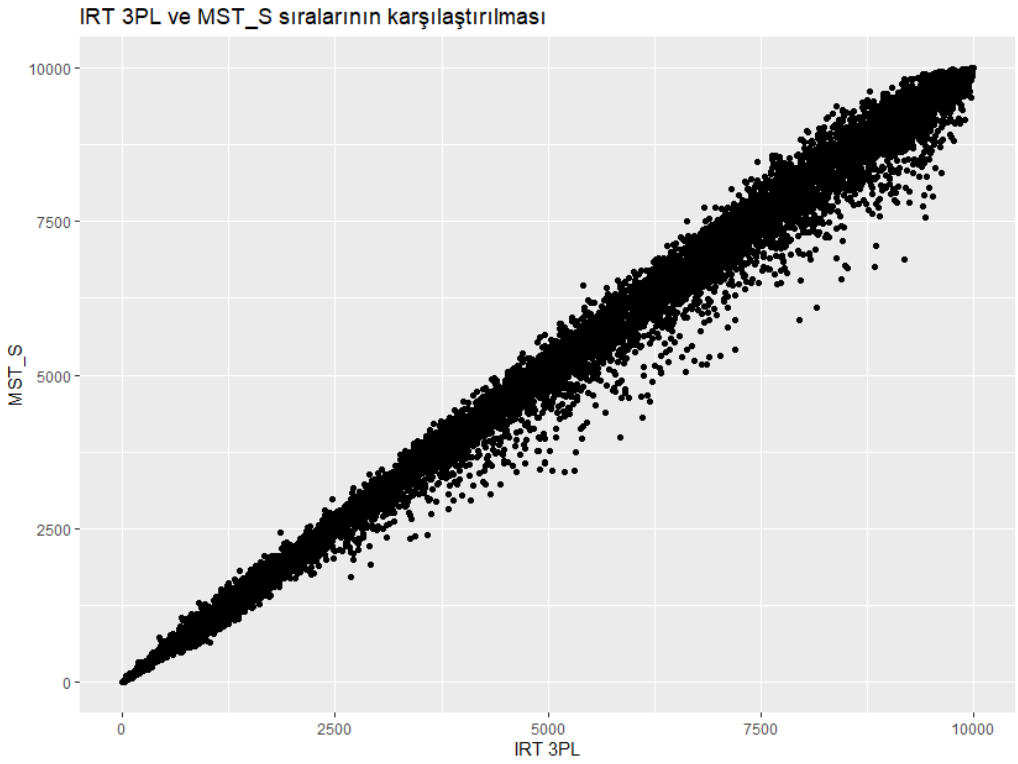
Grafik 4.58. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve IRT 3PL puan sıralarının karşılaştırması*

KTK ve IRT 3PL puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıralarının değişmediği tespit edilmiştir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.59. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 10000 kişilik veri seti için KTK ve MST-S puan sıralarının karşılaştırması*

KTK ve MST-S puan sıralarının karşılaştırıldığı 10000 kişilik veri setinde ise aynı şekilde uç kısımlardaki yetenek düzeylerinde yer alan sınav katılımcılarının sıraları değişmemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının ise saçılımları artarak puan sıralamaları önemli derecede farklılık göstermektedir.



Grafik 4.60. *"Shaping" yöntemine göre üretilmiş normal olmayan (sağa çarpık) dağılıma sahip 10000 kişilik veri seti için IRT 3PL ve MST-S puan sıralarının karşılaştırması*

IRT 3PL ve MST-S yöntemlerinin puan sıralarının karşılaştırıldığı 10000 kişilik verisetinde ise uç kısımlardaki yetenek düzeyinde bulunan sınav katılımcılarının puan sıraları bu veri setinde de aynı kalarak herhangi bir farklılık göstermemektedir. Orta kısımlarda bulunan yetenek düzeyindeki sınav katılımcılarının puan sıralamalarının da KTK - IRT 3PL ve KTK - MST-S karşılaştırmalarına oranla daha düşük düzeyde bir saçılım göstermektedir. Diğer bir ifadeyle IRT 3PL ve MST-S yöntemlerinin puan sıralamaları diğer karşılaştırmalara (KTK - IRT 3PL ve KTK - MST-S) oranla dahaz az farklılık gösterdiği sonucuna ulaşılmakla birlikte son sıralarda saçılımın arttığı tespit edilmiştir.

5. TARTIŞMA

Bu başlık altında "Bulgular" bölümünde sunulmuş olan analiz sonuçları farklı kuramlara (KTK, IRT ve MST) göre "Standart Hata Değerleri", "Farklı Yöntemlere Göre Elde Edilen Yetenek Ölçülerinin Korelasyon Katsayıları", "Farklı Yöntemlerin Verilere Ne Kadar Uyuştuğunu Gösteren AIC Değerleri" ve "Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları" bağlamında ele alınarak araştırma soruları temelinde tartışılmıştır.

Ölçüm doğruluğu tahminlerinin standart hata değerleri aracılığıyla elde edilebilir olması ve bu durumun MST ölçümleri için de geçerli olmasından dolayı (Park vd., 2017) bulgular bölümünde de ifade edildiği üzere tez kapsamında gerçekleştirilen analizlerde ilk olarak KTK, IRT 3PL ve MST (MST-R ve MST-S) yöntemlerine göre elde edilen puanların "Standart Hata Değerleri" karşılaştırılmıştır. Bilindiği üzere standart hata değerlerine yönelik olarak yapılan hesaplamalarda büyük hata değerine sahip araçlar küçük hata değerine sahip araçlara göre daha az tercih edilmesinin sebebi (Overton, 2012, s. 116-123) küçük hata değerine sahip araçların daha hassas ölçümler gerçekleştirebildiğinin kabul edilmesidir (Mollenkopf, 1949; Overton, 2012; Reise ve Haviland, 2005). Gardner (1989) yıllar önce başarı testlerinin yanlış kullanılmasından (doğru kullanılmayan test sunumundan) dolayı ölçme hataları, değerlendirme problemleri ve puanlama hatalarının yaşandığını belirtmiştir. Doğru kullanılmayan test sunumunda meydana gelebilecek puanlama hataları KTK, IRT 3PL ve MST yöntemleri özelinde tartışılmıştır.

Klasik Test Teorisi (KTK), varsayımlarının test edilmesinin, sonuçlarının yorumlanmasının ve genelleştirilmesinin kolay olması nedeniyle testlerin geliştirilmesinde ve yanıtlayanların puanlarının yorumlanmasında yaygın olarak kullanılmaktadır (Crocker ve Algina, 2008). Uygulayıcıların çoğunun KTK yöntemlerine aşina olması sebebiyle yaygın olarak kullanılmakla birlikte son zamanlarda IRT yöntemleri de popülerlik kazanmaktadır. LeBeau ve arkadaşları (2020) KTK ve IRT'yi karşılaşırdıkları çalışmalarında IRT parametrelerinin yetenek tahmininin standart hatasını düşürmeye yardımcı olduğunu ifade etmektedirler. Diğer bir ifadeyle IRT yöntemine göre hesaplanan yetenek tahminlerinde KTK'ya oranla daha düşük SEM değerleri tespit edilmiştir. Raju ve arkadaşlarının (2007) ise KTK ve IRT kapsamında CSEM (Conditional Standard Error of Measurement/Koşullu Standart Ölçüm Hatası) değeri

hesaplamaları/analizleri IRT lehine sonuçlanmıştır. CSEM, belirli bir gerçek puan veya yetenek seviyesindeki ölçüm hassasiyeti seviyesini göstermektedir (Li, 2020). Tüm puanlarda sabit bir CSEM'e sahip olmak yalnızca puan yorumlamayı ve puan raporlamayı basitleştirmekle kalmamakta aynı zamanda testin adil olmasına da katkıda bulunmaktadır. Bu sebeple standart hata değeri karşılaştırmalarında CSEM değeri de dikkate alınmıştır. AL-khadher ve Albursan (2017) çalışmalarında KTK ve IRT 2PL modelini standart hata değeri üzerinden gerçekleştirmiş oldukları karşılaştırmada IRT lehine sonuçlara ulaşmışlardır. Kline (2005) ise IRT'yi kendi içerisinde 1PL, 2PL ve 3PL modellerini standart hata değerleri bakımından karşılaştırmış ve IRT 3PL modeli en çok bilgi sağlayan ve en düşük hata değerine sahip model olarak tespit etmiştir. Tez araştırması kapsamında elde edilen IRT 3PL lehine düşük standart hata değerleri Kline'nin (2005) bulguları ile benzerlik gösterirken diğer araştırmaların bulgularında KTK ve IRT'ye göre hesaplanan SE/SEM/CSEM değerleri bağlamında elde edilen sonuçlarla da örtüşmektedir. Analizlere MST yöntemlerinin (MST-R ve MST-S) dahil edilerek hem KTK hemde IRT 3PL standart hata değerleri bağlamında karşılaştırılması ise tez çalışmasının diğer çalışmalarla farklılaştığı özgün yönü olarak ifade edilebilir.

Eğitim bilimleri alanında olduğu kadar tıp bilimlerinde de özellikle klinik bağlamlarda bireysel değişimin değerlendirilmesi, klasik test teorisi (KTK) veya madde tepki kuramı (IRT) metodolojileri kullanılarak yapılabilmektedir. Reise ve Henson (2003) IRT'nin KTK'ya göre üstün yönlerini tartıştıkları çalışmalarında IRT yönteminin bilişsel ölçümlerin yanı sıra kişilik ölçümlerinde de kullanılmasının önemine dikkat çekmektedirler. Bu yöntemler bireysel değişim değerlendirmesi gibi klinisyenlerin bir bütün olarak hasta gruplarının ortalama iyileşme oranından (KTK) ziyade bireysel hastalar için tedavinin etkinliği (IRT) ile ilgilendikleri klinik uygulamalarda önemli bir rol oynamaktadır. Jabrayilov, Emons ve Sijtsma (2016) çalışmalarında klinik psikologların hastalarındaki bireysel değişimlerin değerlendirilmesinde KTK ve IRT yöntemlerini standart hata değerleri bağlamında karşılaştırarak bireysel değişiklik tespitinde IRT'nin KTK'dan üstün olduğunu ortaya koymuşlardır. Yine klinik araştırmalarda bireysel değişimin istatistiksel olarak değerlendirildiği bir diğer çalışma Reise ve Haviland (2005) tarafından yapılmıştır. Yazarlar ölçek bilgisi (ölçüm hassasiyeti) ne kadar yüksekse ölçümün standart hatasının o kadar düşük olacağını ifade ettikleri çalışmalarında IRT'nin KTK'dan üstün olduğunu tartışmışlardır. Bu çalışmalara

paralel olarak tez araştırmasının genel sonuçları IRT 3PL modelini en düşük standart hata değerine (ölçüm doğruluğuna) sahip yöntem olarak ortaya koyarken KTK'yı en yüksek standart hata değerine (ölçüm doğruluğuna) sahip yöntem olarak tespit etmiştir. Araştırmanın standart hata değerlerine yönelik KTK ve IRT bulguları Jabrayilov, Emons ve Sijtsma (2016) ile Reise ve Haviland (2005)'ın bulguları ile uyumludur. Tez araştırmasının özgün olduğu ve farklılaştığı yönleri ise şu şekilde ifade edilebilir: MST test sunum yönteminin analizlere dahil edilerek KTK'dan önemli derecede düşük oranlarda ve IRT'ye oldukça yakın oranlarda standart hata değerlerine (ölçüm doğruluğuna) sahip olduğu tespit edilmiştir. MST'nin ayrıca kendi içerisinde "Routing (MST-R)" ve "Shaping (MST-S)" yöntemleri ile elde edilen puanların standart hata değerleri de karşılaştırılmış ve birbirlerine oldukça yakın standart hata değerleri bulunmuştur. Ulaşılan bu bulgular MST yönteminin IRT'ye yakın ölçüm hassasiyetine sahip olduğu yönünde önemli bilgiler sunmaktadır.

Pohl (2014) büyük ölçekli araştırmalarda uyarlamalı test yönteminde çok aşamalı testin (MST) özel bir biçimi olan "Longitudinal Multistage Testing (lMST)" yöntemini tanıttığı çalışmasında standart hata değerleri bakımından yeteneği tahmin etmede geleneksel yöntemden (KTK) daha iyi performans gösterdiği sonucuna ulaşmıştır. Ayrıca MST ve CAT yöntemleri ile çok daha verimli sonuçlar elde edilebileceğini vurgulamıştır. Bu araştırmaların bulguları KTK ve MST'ye göre hesaplanan SEM değerleri bağlamında tez araştırması analiz sonuçları ile örtüşürken analizlere MST-R, MST-S ve IRT 3PL yöntemlerinin dahil edilerek karşılaştırılması noktasında farklılaşmaktadır.

Luecht ve Sireci (2012) çalışmalarında CAT'in birçok durumda MST'den daha düşük standart hatalarla daha iyi teta (θ) tahminleriyle sonuçlandığını ifade etmektedirler (Luecht ve Sireci, 2012). Bu sonuçlar araştırma sonuçları ile örtüşmektedir. Macken-Ruiz (2008) MST ve CAT yöntemlerini simülatif olarak karşılaştırdığı tez çalışmasında yeteneği tahmin etmede ölçümün kesinliğini ve dolayısıyla testin gerçek θ (theta)'yı ne kadar iyi tahmin edebildiğini tespit edebilmek için standart hata değerlerini hesaplamıştır. Sonuç olarak yeteneği tahmin etmede en iyi performansı gösteren (en düşük hata değerine sahip) yöntem CAT olarak bulunmuştur. MST yöntemi ise CAT kadar iyi performans göstermese de yeteneği tahmin etmedeki farklılıklar büyük değildir. Bu durum ise MST tasarımının CAT'e uygun bir alternatif olduğunu göstermektedir. Wang (2017) tarafından

gerçekleştirilen tez çalışmasında MST için tasarlanmış farklı madde havuzları altında koşullu standart ölçüm hatası, maruz kalma oranları, IRT puanlama yöntemi ve içerik özelliklerine göre eşleştirildiğinde MST ve CAT'in ölçüm doğruluğunu ve ortalama test uzunluğunu karşılaştırmaya odaklanmıştır. Bu kapsamda elde edilen sonuçlar MST ve CAT arasında benzer ölçüm doğruluğuna işaret etmektedir. Çalışmalar CAT ve MST karşılaştırmalarında ölçüm doğruluğu/yeteneği tahmin etmede standart hata değerlerine yönelik ulaşılan sonuçlar bakımından tez araştırması kapsamında elde edilen bulgularla benzerlik göstermektedir. Bu bulgulara ek olarak KTK - IRT 3PL, KTK - MST-R, KTK - MST-S, MST-R - MST-S yöntemlerinin de birbirleri ile karşılaştırılması ve sonuç olarak MST yöntemlerinin KTK'dan oldukça uzak IRT 3 PL ile yakın değerler vermesi tez araştırmasının söz konusu çalışmalardan farklılaştığı özgün yönüdür.

Han ve Guo (2014) tarafından MST-S'nin test senaryolarının nasıl çalıştığını anlamak adına bir simülasyon çalışması gerçekleştirilmiştir. Araştırma sonucunda doğrudan standart hata değerlerine yönelik bulgulara rastlanmasa da maddelerin CAT ve MST-R koşullarına kıyasla çok daha dengeli bir şekilde kullanıldığı tespit edilmiştir. Ayrıca MST-S yöntemi MST-R ile "CSEE (conditional standard errors of estimation)" değeri açısından karşılaştırıldığında bazı alanlarda MST-S düşük değerlere sahipken bazılarında ise MST-R daha düşük değerlere sahiptir. CAT yöntemi ise MST-R'ye göre daha düşük CSEE değeri göstermektedir. Bu bulgular ile karşılaştırıldığında tez araştırmasında IRT 3PL (CAT), MST-R ve MST-S yöntemlerinin standart hata değerlerine yönelik elde edilen sonuçlar benzeşmektedir.

Rotou ve arkadaşları (2007), IRT 1PL, 2PL ve 3PL modelleri kapsamında KTK, CAT ve MST yöntemlerini ölçüm hassasiyetleri açısından CSEM değerleri ile karşılaştırdıkları çalışmalarında MST'nin 2PL ve 3PL modelleri için KTK'dan, 1PL ve 2PL modelleri için CAT testinden daha iyi performans gösterdiğini ortaya koymuşlardır. Ayrıca MST, 3PL modeli için ise CAT ile aynı performansı göstermiştir. Bu çalışmada her üç yönteminde hata değerleri bağlamında ele alınmış olması ve sonuçların MST lehine veya CAT ile eşdeğer ya da CAT'e yakın bulunmuş olması tez araştırması ile benzerlik gösterirken KTK ve CAT, KTK ve MST ile CAT ve MST yöntemlerinin ayrıca birbirleri ile karşılaştırılmamış olması noktasında ise farklılaşmaktadır. Tez araştırmasında KTK ve CAT (IRT 3PL), KTK ve MST (MST-R ve MST-S yöntemleri

ayrı ayrı ele alınmıştır) ile CAT (IRT 3PL) ve MST (MST-R ve MST-S) yöntemleri birbirleri ile karşılaştırılmış olup MST (MST-R ve MST-S) standart hata değerleri bağlamında CAT'e yakın bir seyir izlerken KTK'dan önemli derecede uzaklaşmaktadır.

Yukarıda yer alan açıklamalardan yola çıkarak tez araştırması kapsamında her üç yöntemin standart hata değerleri bağlamında bütüncül karşılaştırması Tablo 5.1.'de sunulmuştur.

Tablo 5.1. *Veri setlerinin tamamına yönelik standart hata değerleri*

	Normal Dağılım			Normal Olmayan Dağılım (Sola Çarpık)			Normal Olmayan Dağılım (Sağa Çarpık)		
MST_R	**KTK_se**	**IRT_se**	**MST_R_se**	**KTK_se**	**IRT_se**	**MST_R_se**	**KTK_se**	**IRT_se**	**MST_R_se**
100 Kişi	**0,87013152**	0,1810704	0,28203	**0,71965076**	0,2930234	0,33582	**0,60913817**	0,2283728	0,39721
1000 Kişi	**0,84187527**	0,2739308	**0,26777**	**0,7086569**	0,2813307	0,33733	**0,59352599**	0,2977713	0,38019
10000 Kişi	**0,90876838**	0,22273809	0,26452	**0,62082535**	0,2675895	0,31973	**0,64604923**	0,2931815	0,3701
MST_S	**KTK_se**	**IRT_se**	**MST_S_se**	**KTK_se**	**IRT_se**	**MST_S_se**	**KTK_se**	**IRT_se**	**MST_S_se**
100 Kişi	**0,7404971**	0,1984967	0,2972	**0,5805668**	0,2255257	0,34198	**0,54247368**	0,2004066	0,40867
1000 Kişi	**0,94392453**	0,27393077	0,29626	**0,65976709**	0,2827763	0,3347	**0,6828017**	0,3390708	0,35256
10000 Kişi	**0,87392764**	0,3388127	**0,29014**	**0,64587738**	0,3102667	0,35355	**0,67000614**	0,3620553	0,38371

Tablo 5.1.'den de görüleceği üzere farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) MST-R ve MST-S yöntemleri ile simülatif olarak elde edilen veri setleri KTK, IRT 3PL ve MST (MST-R ve MST-S yöntemleri ayrı ayrı ele alınmış olmak kaydıyla) yöntemlerine göre analiz edilmiş ve standart hata değerleri açısından karşılaştırılmıştır. Analiz sonuçlarından elde edilen bulgular göstermektedir ki; IRT 3PL yöntemine göre elde edilen puanların standart hatası diğer iki yöntemle karşılaştırıldığında en düşük hata değerine sahip yöntem olarak tespit edilmiştir. MST (MST-R ve MST-S) test sunum yönteminin ise KTK'ya göre önemli derecede daha az hata ile kestirim yapabilmesinin yanı sıra bazı örneklem gruplarında IRT 3PL'den daha az hata ile kestirim yapabildiği sonuçlara da ulaşılmıştır. KTK yöntemi

ise tüm örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) diğer iki yönteme (MST ve IRT 3PL) göre dikkate değer oranda yüksek hata değerleri ile yetenek kestirimleri (puan hesaplamaları) yapmaktadır.

MST'nin IRT yöntemine yakın değerler vermesi bu yöntemin geniş kitlelere çevrimiçi sınav uygulayan sistemler tarafından CAT yöntemine alternatif olarak kullanılabileceğinin önemli bir göstergesidir. KTK yöntemine yönelik olarak elde edilen yüksek oranlardaki hata değerlerinin ise ölçme ve değerlendirme alanı için ifade ettiği anlam şudur: Gerçekleştirilen ölçümlerde olası bütün varyans kadar hatanın söz konusu olabileceği diğer bir ifadeyle yetenek kestirimlerinin (puanların) tamamının ya da tamamına yakınının yanlış hesaplanmış olabileceği ihtimalini kuvvetle desteklemesidir. Ulaşılan bu değerler ölçme ve değerlendirme aşamasında çok daha hassas ölçümler yapabilen alternatif yöntemlere (IRT/MST) olan ihtiyacın önemini vurgular nitelikte bir bulgudur.

Açık ve uzaktan öğrenme sistemleride dahil olmak kaydıyla halihazırda eğitim sistemlerinde en yaygın kullanılan ölçüm teorisinin geleneksel KTK olduğu gerçeği dikkate alındığında elde edilen bu hata değerlerinin önemi çok daha net bir biçimde ortaya çıkmaktadır. Ayrıca açık ve uzaktan öğrenme sistemlerinde birey materyal etkileşiminin barındırdığı transaksiyonel uzaklığa KTK'dan elde edilen puanların taşıdığı bu kısıtlılıklar da eklendiğinde sorunun büyüklüğü daha net görünmektedir.

Standart hata değerlerine yönelik analizlerde elde edilen bulguları kuvvetlendirmek adına gerçekleştirilen "Kendall's Tau-b" korelasyon analizi sonuçları da bu bulguları destekler niteliktedir.

Hwang (2002), KTK ile IRT modellerini korelasyon değerleri bakımından karşılaştırdığı çalışmasında KTK ve IRT 3PL modeli arasında diğerlerine oranla daha düşük seviyede korelasyon tespit etmiştir. Bu sonuç tez araştırması kapsamında gerçekleştirilen KTK ve IRT 3PL "Kendall's Tau-b" korelasyon analizi sonuçları ile uyum göstermektedir. Tez araştırmasında ayrıca KTK - MST-R, KTK – MST-S, IRT 3PL – MST-R ve IRT 3PL – MST-S yöntemleri bağlamında "Kendall's Tau-b" korelasyon analizleri gerçekleştirilmiştir. KTK – MST-R ve KTK – MST-S karşılaştımalarında da benzer şekilde düşük korelasyon değerleri tespit edilirken IRT 3PL – MST-R ve IRT 3PL

– MST-S karşılaştırmalarında yüksek korelasyon değerlerine ulaşılmıştır. Korelasyon analizlerine MST yöntemlerinin eklenmiş olması tez araştırmasının farklılaştığı özgün yönüdür.

Pohl (2014), büyük ölçekli araştırmalarda uyarlamalı test yönteminde çok aşamalı testin (MST) özel bir biçimi olan "Longitudinal Multistage Testing (lMST)" yöntemini tanıttığı çalışmasında tüm korelasyonlar açısından yeteneği tahmin etmede geleneksel yöntemden (KTK) daha iyi performans gösterdiği sonucuna ulaşmıştır. Bu araştırmaların bulguları KTK ve MST'ye göre hesaplanan korelasyon değerleri bağlamında tez araştırması analiz sonuçları ile örtüşürken analizlere MST-R, MST-S ve IRT 3PL yöntemlerinin dahil edilerek karşılaştırılması noktasında farklılaşmaktadır.

Macken-Ruiz (2008) tarafından gerçekleştirilen çalışmada hem CAT hemde MST temelinde bilinen ve tahmin edilen *θ(*theta) değerleri arasındaki Pearson korelasyonları testlerin her birinin bilinen yetenek tahminini ne kadar iyi geri kazandığının bir ölçüsü olarak hesaplanmıştır. Bilinen ve tahmin edilen *θ*(theta) arasındaki en yüksek korelasyonu CAT vermiştir. MST yöntemi CAT'ten düşük korelasyona sahip olmakla birlikte CAT'e yakın değerlerine sahiptir. Verilerin rastgele dağılım esasına uyduğunun kabul edildiği parametrik bir dağılımda sürekli değişkenlerin birbirleriyle doğrusal ilişkisinin derecesine "*Pearson*" korelasyon testi uygulanarak bakılmaktadır (Chen ve Popovich, 2002; Dawson ve Trapp, 2004; Pagano ve Gauvreau, 2018). Macken-Ruiz (2008) çalışmasında bilinen ve tahmin edilen *θ*(theta) değerlerini ele aldığı için Pearson korelasyonunu kullanmıştır. Bu çalışmanın tez araştırması ile benzeşen noktası ise yetenek tahmininde CAT ve MST'nin birbirlerine yakın korelasyona sahip olmalarıdır. Tez araştırması kapsamında gerçekleştirilen "Kendall's Tau-b" korelasyon analizi sonuçları da CAT ve MST arasında yüksek korelasyon vermiştir. Değişkenlerden birinin ya da her ikisinin de sıralı değişken olması durumunda kullanılan test "*Kendall'ın tau-b (τb)*" korelasyon analizidir (Arndt, Turvey ve Andreasen, 1999; Chen ve Popovich, 2002). Tez çalışması kapsamında incelenen KTK, IRT ve MST temelli puanların sıralama amacıyla kullanılma ihtimallerinin yüksek olması sebebiyle *Kendall'ın tau-b (τb)* katsayısının kullanımı tercih edilmiştir. Bu açıdan değerlendirildiğinde tezde KTK, IRT(CAT) ve MST (MST-R ve MST-S) yöntemlerinin *Kendall'ın tau-b (τb)* korelasyon analizi ile sıralama anlamında ele alınmış olması ise bu araştırmanın farklılaştığı yönü olarak nitelendirilebilir.

"Kendall's Tau-b" korelasyon katsayısına göre farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) gerçekleştirilen analiz sonuçlarına göre MST ve IRT 3PL yöntemlerinin birbirine karşı büyük oranda uyum göstermesi ve KTK yöntemi ile uyumsuzlaşması standart hata değerleri aracılığıyla önemli derecede hatalı puanlama yapıldığı bulgusunu doğrular/destekler niteliktedir. Bulgular bölümünde yer alan analiz sonuçlarında görüldüğü üzere KTK'nın korelasyon katsayısı diğer kuramlara (IRT 3PL/MST) oranla önemli derecede farklılaşmaktadır. Bu durumda tıpkı standart hata değerlerinde olduğu gibi böylesi hata barındıran bir ölçüm yönteminin yerini daha hassas ölçümler gerçekleştirebilen yöntemlerin alması gereğinin sinyallerini vermektedir.

Tez çalışması kapsamında gerçekleştirilen bir diğer analiz ise; KTK, IRT ve MST-R yöntemleri için elde edilen yetenek kestirimlerinden yola çıkarak optimum modeli tespit etmek adına farklı yöntemlerin verilere ne kadar uyuştuğunu gösteren AIC değerlerini hesaplamaya yönelik analizlerdir. Bu aşamada tüm maddelerinde "Eksik" veri barındırmasından kaynaklı olarak MST-S yöntemi için AIC değeri hesaplanamamıştır. MST-R yöntemi için ise sadece tüm katılımcıların yanıtlamış olduğu "Eksik" veri barındırmayan maddeler için AIC değeri hesaplanabilmiştir. Bulgular bölümünde sunulmuş olan sonuçlar bu sınırlılıklar dahilinde elde edilmiştir.

van Rijn (2014) IRT (1PL, 2PL ve 3PL) modellerini AIC değerleri bağlamında karşılaştırdığı çalışmasında IRT 3PL model en uyumlu model olarak tespit edilmiş olması tez araştırması ile benzerlik göstermektedir. Analizlere ek olarak KTK ve MST-R yöntemlerinin dahil edilerek karşılaştırılması noktasında ise farklılaşmaktadır. Ulaşılan araştırma/analiz bulguları göstermektedir ki; MST-R yöntemi optimum model uyumu açısından IRT 3PL modele yakın değerlere sahip olmakla birlikte doğrudan optimum model olarak tespit edildiği analiz sonuçlarıda mevcuttur. KTK yöntemi ise optimum model uyumu anlamında hem IRT 3PL'den hemde MST-R'den önemli derecede uzaklaşarak veriler ile en az uyumlu olan model olarak tespit edilmiştir. Bulgularda da ifade edildiği üzere KTK yöntemine göre gerçekleştirilen puanlamalar önemli derecede yüksek hata barındırmakta ve optimum model olmaktan çok uzak bir seyir ortaya koymaktadır.

AIC değerleri analizinin en önemli sınırlılığı MST test sunum yönteminin yapısı gereği çok fazla "Eksik" veri barındırmasından dolayı tüm veri setlerinde analizlerin yapılamamış olmasıdır. Analiz programlarında yazılımsal olarak bu hususu dikkate alarak hesaplamalar gerçekleştirebilecek yapıların tasarlanması analiz süreçleri açısından önemli bir katkı sağlayacaktır.

Tez araştırması kapsamında gerçekleştirilen son analizde ise KTK, IRT ve MST yöntemlerine göre puan sıralamalarının farkları incelenmiştir. Verilerin analizi; her üç kurama (KTK, IRT ve MST) göre farklı örneklem koşullarında (örneklem büyüklüğü, örneklem homojenliği ve dağılımın şekli) puan sıralamalarının göstermiş olduğu değişimleri birbirleri ile karşılaştırılarak gerçekleştirilmiştir. Bu yönde bir analize gereksinim duyulmasının sebebi önceki analizlerde tespit edilen "*Kendall'ın tau-b (τb)*" korelasyon katsayılarının oransal değerleri vermesi sıralama anlamında bu yönde bir bilgiyi içermemesidir. Bilindiği üzere Açık ve uzaktan öğrenme sistemleri dahil büyük ölçekli merkezi sınav gerçekleştiren tüm sistemlerde başarılı/başarısız, geçti/kaldı veya işe alımlarda puan sıralamasına göre sonuca karar verilmektedir. Puan sıralamasının bireye yönelik karar vermede bu denli önemli rol oynadığı bir sistemde her bir yönteme göre hesaplanan sınav puanlarının maksimum kaç kişiye kadar yanılarak hesaplanmış olabileceğinin tespitine yönelik analiz sonuçları kaydadeğer nitelikte önemli bulgulardır.

Alan yazında bu yönde gerçekleştirilen çalışmaların bulguları şu şekilde özetlenebilir. Zaman ve arkadaşları (2008) öğrenci puanlarını KTK ve IRT yöntemlerine göre ayrı ayrı sıraladıkları çalışmalarında IRT bazında yapılan öğrenci sıralamalarında önemli bir kayma olduğunu gözlemlemişlerdir. Araştırmacılar gerçekleştirdikleri analizlerde 100 öğrenciden sadece 9'unun yerinin sabit kaldığı diğer öğrencilerin tamamının sıralamadaki yerlerinin değiştiği sonucuna ulaşmışlardır. Binh ve Duy (2016) ise çalışmalarında KTK ve IRT yöntemlerine göre gerçekleştirilen puan sıralamaları farklılaştığını tespit etmekle beraber IRT'nin KTK'ya kıyasla daha düzgün ve daha ayrıntılı sonuçlar sunduğunu vurgulamaktadırlar. Ayrıca araştırmacılar bilgisayarla uyarlanabilir testte IRT yönteminin uygulanmasının e-öğrenme sistemlerinde daha iyi bir başarı için öğrenme dokümantasyonu önermesine, uzmanların çevrimiçi test veya diğer akıllı özel ders sistemleri oluşturmalarına yardımcı olabileceğini ifade etmektedirler. Ayanwale ve Adeleke (2020) sınav katılımcıları için iki zıt çerçeve çalışması (KTK ve IRT) kapsamında elde edilen puanların aynı metrik ölçeğe dönüştürülerek puan

sıralamalarının oluşturulduğu araştırmalarında önemli sıralama farklılıkları bulduklarını ifade etmektedirler. Örneğin, KTK'da en yüksek sınav katılımcıları 70 puan alırken, IRT sıralamasında bu göreceli durum değişmiş ve 69.210 puan alan katılımcı sıralamanın en üstünde yer almıştır. Bu durum IRT sıralamasının daha üstün olduğunu göstermektedir. Çünkü 69.210 puan alan bir sınav katılımcısı kolay sorularda yanlış seçenekler seçerek daha az puan kaybına uğrarken, KTK'da 70 puan alan bir sınav katılımcısı zor sorulara yanıt veremediği için göreceli konumu daha fazla puan kaybıyla değişmiştir. Bu nedenle araştırmacılar IRT yönteminin test geliştirme ve puanlamada KTK'dan daha etkili olduğu sonucuna varmışlardır. Mutiawani, Saputra ve Subianto (2022) benzer şekilde KTK ve IRT yöntemlerine göre katılımcıların puan sıralamalarını analiz ettikleri çalışmalarında katılımcı sıralamalarında önemli derecede artış ve azalışlar gözlemlemektedirler. Bu açıdan değerlendirildiğinde tez araştırması farklı kuramlara göre puan sıralaması farklarının analiz sonuçları bakımından yukarıda belirtilen araştırma bulguları ile örtüşmekle beraber KTK, IRT ve MST yöntemlerini hem "Routing" hem de "Shaping" temelinde ayrı ayrı ele alarak bir bütün olarak analiz edilmiş olması araştırmanın özgün yönünü vurgulamaktadır.

Tüm bu araştırmaların ortak bir bulgusu şudur ki; öğrencilerin puan sıralamaları uç kısımlarda (en düşük ve en yüksek yetenek düzeyinde bulunan öğrencilerin yer aldığı sıralamalar) değişiklik göstermemekte olup orta yetenek düzeyinde bulunan öğrencilerin sıralamaları önemli ölçüde değişiklik göstermektedir. Bu bulgu tez kapsamında KTK, IRT ve MST yöntemlerinin birbirleri ile puan sıralama farkları bağlamında karşılaştırıldığı analizlerde elde edilen sonuçlarla örtüşmektedir.

Yukarıda yer alan tartışma ve değerlendirmelerden yola çıkarak alan yazında önceden gerçekleştirilen çalışmaların bulguları ile tez çalışması özelinde elde edilen bulgular doğrultusunda konuyu bütüncül olarak şu şekilde ifade edilebiliriz/tartışabiliriz:

Geleneksel KTK puanlama stratejisi ölçülen özellik ile bireylerin performansları arasında doğrusal ilişki varsayması, elde edilen parametrelerinin örnekleme bağımlı olması, standart hatayı ve güvenilirliği testi alan tüm bireyler için eşit kabul etmesi gibi bir takım ölçüm kısıtlılıkları nedeniyle uzun süredir eleştiri altında olan bir puanlama yöntemidir (Arnold, 1996; Crocker ve Algina, 2008; Kline, 2005; Magno, 2009; Rusch vd., 2017; Weis ve Yoes, 1990). KTK'ya göre bireysel sınava girenlerin yeterliliği doğru

yanıtlanan madde sayısı cinsinden rapor edilmektedir. Örneğin herhangi bir genel sınavdaki test maddelerinin kalitesi her zaman sınava girenlerin yanıtlarının madde analizi yoluyla incelenmesiyle ölçülmektedir. Madde analizi; bu maddelerin ve bir bütün olarak testin kalitesini değerlendirmek için öğrencilerin bireysel test maddelerine verdiği yanıtları inceleyen bir süreçtir. Doğru yanıt sayısına göre tespit edilmiş sınav puanına sahip öğrencilerin farklı yanıt kalıplarına (yani farklı maddelerde doğru yanıtlara) sahip olabilmeleri ve bu nedenle test tarafından ölçülen aynı yeterlik düzeyine sahip olmama durumları ölçümün kalitesini etkileyen bir husustur. Test maddelerinin kalitesi ile ilgili raporlar ise genellikle madde güçlük indeksleri (madde üzerindeki doğru yanıtların oranı) ve madde ayırt ediciliği indeksleri ile sınırlıdır. Ancak bu tür indekslerle ilgili önemli bir sorun, bunların test edilen sınava girenlerin grubuna bağlı olması ve bu nedenle test maddelerinin ölçüm kalitesini yeterince yansıtmamasıdır (Adedoyin ve Mokobi, 2013, s. 992-993). Bu tür bir puanlama yöntemi ise Klasik Test Kuramı'nın önemli bir sınırlılığı ve zayıf kaldığı yönü olarak nitelendirilmektedir. Ülkemiz dahil eğitim sistemlerinin tüm kademelerinde yaygın olarak uygulanan bu sistemle devam edildiği takdirde sınavı geçmemesi gereken bir birey geçmiş ya da aksine geçmesi gereken bir birey başarısız olarak değerlendirilerek hatalı bir ölçme ve değerlendirme gerçekleştirilmiş olmaktadır. Diğer bir ifadeyle ölçme ve değerlendirmenin temel amacı olan bireyin zihnindeki gizil (örtük) bilgi miktarı için doğru bir ölçüm yapılamamaktadır. Bu durum eğitim öğretim sürecinin çıktılarının net bir biçimde değerlendirilememesi anlamına gelmektedir. Kearney (1983), geniş ölçekli bir değerlendirmenin temel amaçlarını şu şekilde ifade etmektedir: (a) yerel, bölgesel veya ulusal düzeyde kamuya raporlama yapmak, (b) gereksinimleri belirleyerek kaynakları tahsis etmek ve (c) terfi ve mezuniyet hakkında karar vermektir (s. 9-11). Değerlendirme, süreç boyunca öğrencinin kültürel, dilsel ve etnik geçmişini göz önünde bulundurarak öğrenciyi bütüncül olarak görmelidir. Bu anlamda mevcut geleneksel değerlendirme modeli özellikle eğitimciler tarafından yetersiz/eksik/sorunlu bulunduğundan müdahale ve problem çözmeyi vurgulayan çağdaş bir değerlendirme modeli desteklenmektedir (Overton, 2012). Örneğin, çok yüksek standartlarda öğretim malzemeleri tasarlanarak bireylerin hizmetine yüksek nitelikte içerikler sunulmuş olabilir. Ancak bu denli yüksek oranlarda hata barındıran bir ölçme ve değerlendirme yöntemi benimseyen eğitim öğretim sistemi bu aşamanın öncesinde yapılan hiçbir yatırımın sonucundan net olarak haberdar olamayacaktır. Sağlıklı bir ölçüm yapılabildiği oranda öğrenme çıktılarının başarısından ya da başarısızlığından

bahsedebilmek mümkündür. Bunun için hassas ölçüm kabiliyetine sahip olan düşük standart hata değerleri ile ölçüm yapabilen yöntemlerin (IRT ve MST gibi) tercih edilmesinin gereği öne çıkmaktadır.

Modern puanlama stratejisi olarak bilinen IRT ölçme ve değerlendirme alanında kullanılan bir diğer puanlama yöntemidir. Bilgisayarda bireye uyarlanmış testlerde en bilinen ve operasyonel anlamda en yaygın kullanılan CAT puanlama yöntemi olarak IRT modellerini kullanmaktadır (Weiss ve Kingsbury, 1984). CAT kullandığı IRT puanlama yöntemi ile madde ayırt ediciliği=a, madde güçlüğü=b ve şans başarısı=c (sorunun şansla doğru yanıtlanma yüzdesi) gibi parametrelerin ölçümüne olanak tanımaktadır. CAT sisteminde bu parametreler sayesinde geniş soru havuzları bireylere uyarlanarak her bireye özgü test oluşturulması sağlanabilmektedir. CAT test sunum yönteminde maddeler her bir sınav katılımcısının önceki maddelere verdiği yanıtlara göre sınav katılımcısının altta yatan gizil yeteneğinin tahmininindeki kesinliği hedefleyecek ve en üst düzeye çıkaracak şekilde seçilmektedir (Rotou, 2007). Bireye uyarlanmış testlerin yapısı çoktan seçmeli test sunum tekniği ile gerçekleştirilen sınavlarda aynı yetenek düzeyindeki katılımcıya aynı zorlukta farklı sorular sorulabilmesine imkân vermektedir. Sistemin bu özelliği eğitim öğretim sürecinin temel problemi niteliğindeki güvenlik sorununa çözüm getirmeye aday bir fonksiyondur. Wang, Zheng ve Chang (2014) çalışmalarında bilgisayar tarafından yönetilen bireye uyarlanmış testlerin daha yüksek derecede kontrol sağlamaya olanak tanıdığını ve test güvenliği arttırma fırsatı sunduğunu ifade etmektedirler. CAT test sunum yönteminin belirgin bir diğer avantajı ise daha kısa bir test potansiyelinin bulunmasıdır (Mead, 2006; Rotou, 2007; Weiss ve Kingsbury, 1984; Yan, von Davier ve Lewis, 2014, s. 19). Uyarlanabilir testlerin temel amacı, testi her bir sınav katılımcısının yetenek düzeyine göre uyarlamak ve sınavı birey için ne çok kolay ne de çok zor hale getirmeden yetenek düzeyine göre testi uygulamaktır (Lord, 1968). Çok kolay veya çok zor olan maddeler, sınava giren kişinin yetenek düzeyini tahmin etmek için çok az bilgi sağladığından, uyarlamalı testler, çok fazla ölçüm doğruluğundan ödün vermeden test uzunluğunu etkili bir şekilde azaltabilmektedir. Örneğin bir sınav katılımcısı için çok kolay veya çok zor olan maddeler, bir içerik spesifikasyonunu karşılamak veya başka bir maddenin aşırı maruz kalmasını önlemek için bir maddeye gereksinim duyulmadıkça uygulanmaz. Maddelerin sınava giren kişinin yetenek

düzeyine göre uyarlanması geleneksel KTK yöntemine oranla daha verimli olan uyarlanabilir testlere imkân tanımakta (Lord, 1980; Weiss, 1982) ve doğal olarak sınava giren katılımcıların eşdeğer bir hassasiyet düzeyine ulaşmak için daha az maddeyi yanıtlamasını gerektirmektedir (Schnipke ve Reese, 1997). CAT'ler çok verimli testlerdir çünkü geleneksel bir kağıt kalem testinden daha az öğeyle istikrarlı bir yeterlilik tahminine ulaşılabilmektedir (Weiss ve Kingsbury, 1984).

IRT puanlama yönteminin kullanıldığı CAT sistemi KTK'nın sınırılıklarını büyük oranda çözmekle beraber kendi içinde bazı sorunlar barındırmaktadır. CAT'in yüksek düzeyde akademik avantajları olmakla birlikte hukuki, mali, kültürel ve sosyal bazı dezavantajları bulunmaktadır (Glas ve van der Linden, 2003; Ockey, 2012, s. 347). Örneğin; Hukuk sistemleri eşitlik ilkesi gereği aynı soruyu doğru yanıtlamış iki bireyin birbirinden farklı puan almasını onaylayacak şekilde evrilmemiştir. Bu anlamda akademik olarak adil bir yöntem olmasına rağmen CAT'e dayalı puanlama hukuken tartışmalara açık bir konudur. Mali açıdan değerlendirildiğinde ise bilgisayarlı bireye uyarlanabilir test uygulamaları büyük soru bankalarına, iyi yetişmiş ekiplere ve yüksek teknolojik altyapıya gereksinim duymaktadır. Tüm bu uygulamaların ciddi anlamda emek, uzmanlık, maliyet ve zaman gerektirmesi CAT sisteminin sürdürülebilirliğini güçleştirmektedir. Kültürel ve sosyal anlamda ise toplumsal olarak bireyler soruların açıklanması konusunda ısrarcı bir tutum sergilerken CAT tüm soruların saklanması konusunda yüksek bir standarda sahiptir. CAT test sunum yönteminin sıklıkla eleştirilen yönlerinden biri de sınav katılımcılarının maddeleri atlamalarına ve tamamlanmış maddelere tekrar dönmelerine izin vermemesidir (Hendrickson, 2007; Macken-Ruiz, 2008; Luecht, 2003; Wainer, 1993; Yan, Lewis ve von Davier, 2014). Sınav güvenliğinin manipüle edilmesini önlemek için uygulanmakta olan bu yöntem sınav katılımcılarında gereksiz stres oluşmasına sebep olmaktadır. Bu koşullar altında CAT sisteminin bazı kültürlerde kabul görmesi zor bir durumken sınav güvenliği konusunda bireye uyarlanmış testlere olan gereksinim ise üstesinden gelinmesi gereken önemli bir konudur. Yetenek, başarı, yeterlilik, giriş ve profesyonel lisans testlerini içeren geniş ölçekli bilişsel değerlendirmede, madde yanıt teorisi (IRT), ölçek oluşturma, analiz ve puanlama için baskın psikometrik paradigmadır (Resie ve Henson, 2003). Kişilik verilerine yönelik IRT yöntemi uygulamaları olmasına rağmen (Chernyshenko vd., 2001; Fraley, Waller ve

Brennan, 2000; Harvey ve Murry, 1994; Steinberg, 1994), geleneksel klasik test teorisi (KTK) hâkim psikometrik yöntem olmaya devam etmektedir. IRT yönteminin daha önce sayılan dezavantajları bu durumun sebepleri arasında olabilir.

Bilgisayarda bireye uyarlanmış testlerde IRT puanlama stratejisi ile uygulanan bir diğer test sunum yöntemi ise MST'dir. MST test sunum yöntemi "Routing" ve "Shaping" modelleri olmak üzere iki farklı şekilde uygulanabilmektedir. MST, CAT'e alternatif bir test sunum yöntemi olup ana hatları ile uygulama biçimi şu şekildedir: CAT'de, her bir madde sınava giren kişinin önceki maddelere verdiği yanıtlara göre madde havuzundan anlık seçilirken MST'de paneller her aşamada çeşitli zorluk seviyelerinde (kolay, orta ve zor olmak üzere) sabitlenmiş belirli sayıda modülerden (maddelerin bulunduğu bloklar) oluşmaktadır (Zheng ve Chang, 2015). Uygulama sırasında her bir sınav katılımcısına önceden monte edilmiş paralel panellerden biri rastgele atanmaktadır ve kişi ilk modüldeki yetenek seviyesine göre bir sonraki aşamada ilgili modüle yönlendirilmektedir. Kişinin sınavın tamamında aldığı modüller grubuna ise yol ya da rota adı verilmektedir (Yan, von Davier ve Lewis, 2014). Yani CAT madde bazlı bir test montaj yöntemini benimserken MST modül bazlı bir test montaj yöntemini benimsemektedir. MST test sunum yönteminin modül bazlı yapısı sayesinde ise sınava girenler yanıtlarını gözden geçirmek ve değiştirmek için mevcut aşama içinde serbestçe ileri geri hareket edebilme imkanına sahiptir (Han ve Guo, 2014, s. 122-123; Macken-Ruiz, 2008; Luecht, 2003; Zheng ve Chang, 2015, s. 109-111). MST'nin bu özelliği bireylerin stres sebebinin ortadan kaldırılmasını sağlamakla birlikte soruların bir bütün olarak aynı anda ve tek seferde sızması önlenerek bireylere herbir modülde aynı yetenek seviyesi için hazırlanmış farklı sorular sunulabilmektedir. Bu durumun MST'nin CAT test sunum yöntemine alternatif olarak değerlendirilmesinde önemli bir katkısı sağlayacağı düşünülmektedir. Ayrıca MST'nin modül bazlı yapısı sürecin daha düşük maliyetle ve mevcut kültüre daha uygun bir biçimde işlemesine olanak tanımaktadır. MST, kağıt ve kalem modunda teste izin verdiğinden diğer uyarlanabilir test tasarımlarının uygulanamadığı değerlendirmelerde geleneksel teste (KTK) bir alternatif teşkil edebilmektedir (Pohl, 2014). Multistage yöntem CAT'e göre özellikle mali anlamda önemli avantajlar sağlamaktadır. Örneğin; daha az emekle daha sürdürülebilir bir sistem kurulabilmesi, soru bazında daha düşük harcamaya olanak tanıması, Multistage sistemlerin açık kaynak kodlu olarak üretilebilmesi (aynı durum CAT için mümkün

olmakla birlikte çok daha fazla yetişmiş elemana gereksinim bulunmaktadır) gibi konularda önemli avantajları barındırmaktadır. CAT soruları bireye madde bazlı olarak yöneltirken MST modül bazlı olarak yöneltmektedir. Bu özellikleri sayesinde MST test sunum yöntemi güvenlik ve maliyet avantajı yönünde beklentileri karşılayabileceğinin sinyallerini vermektedir. Bu yönü MST test sunum yönteminin güvenlik, maliyet ve kültürel anlamda bir adım önde olmasına olanak tanımaktadır.

Çevrimiçi testlerin en büyük dezavantajı daha önce sınava giren bireylerin sınava daha sonra giren katılımcılarla bilgi paylaşması ve bu bilginin sosyal platformlarda hızla yayılması ciddi anlamda güvenlik sorunlarına neden olmaktadır. Son zamanlarda bilgisayar tarafından yönetilen testlerin sınav uygulayıcıları tarafından giderek daha fazla benimsenmesi test güvenliği konusunu ön plana çıkarmaktadır. Bu anlamda güvenlik konularını test etmek ve test güvenliğini değerlendirmek için yapılan çalışmaların madde havuzlama ve görüntüleme konusunda gerçekleştirilmiştir (Chang ve Zhang, 2002; Chang, 2004; Stocking, 1994; Wang, Zheng ve Chang, 2014; Way, 1998). MST algoritması ile hazırlanmış bir test sunum yönteminde sınava giren her bir katılımcı tarafından aşama başına yalnızca bir test parçası görülebilmekte ve madde havuzundaki tüm sorular tek seferde sunulmadan güvenli bir biçimde test sunumu gerçekleştirilmektedir. Sınav güvenliği konusunun eğitim öğretim sisteminin hemen her kademesinin ölçme ve değerlendirme aşamasının en temel problemleri arasında yer alması sebebiyle analiz sonuçlarından elde edilen bulgular kayda değer nitelikte göstergelerdir.

Özetle yukarıda bahsi geçen konular temelinde optimum test sunum yöntemini tespit edebilmek adına tez araştırması kapsamında gerçekleştirilen "Standart Hata Değerleri", "Farklı Yöntemlere Göre Elde Edilen Yetenek Ölçülerinin Korelasyon Katsayıları", "Farklı Yöntemlerin Verilere Ne Kadar Uyuştuğunu Gösteren AIC Değerleri" ve "Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları"na yönelik analiz sonuçları göstermektedir ki;

MST yöntemi standart hata değerlerine yönelik gerçekleştirilen analiz sonuçlarına göre KTK'dan önemli ölçüde farklılaşarak IRT 3 PL modele daha yakın duran bir ölçüm hassasiyeti sergilemektedir. Bu sonuçlar ülkemizde dahil olmak üzere eğitim sistemlerinin ölçme ve değerlendirme aşamasında en yaygın kullanılan geleneksel KTK

yönteminin ne kadar büyük oranlarda hata içerebileceğini ortaya koyması bakımından dikkate değer bir bulgudur. Bu bulgular sadece madde analizinde yapılan farklı uygulamaların hatalarını göstermekte ve ölçüm sürecine kadar var olan hataları (ölçme aracının hatası, ölçme aracını besleyen unsurların hatası, kapsam hatası, örneklem hatası gibi) içermemektedir. Mevcut sistemin yalnıca madde ananlizinde bu denli önemli oranlarda hatalarla yetenek kestiriminde bulunuyor olması tüm paydaşların (karar alıcılar, araştırmacılar, öğretmenler, öğrenciler vb.) yüksek ölçüm hassasiyetine sahip test sunum yöntemlerine (IRT ve MST vb.) ivedilikle yönelinmesi gereğini vurgulamaktadır.

KTK, IRT ve MST yöntemlerinin yetenek kestirimi/ölçüm hassasiyeti konusundaki farklılıklarını detaylandırmak ve "Standart Hata Değerleri"ne yönelik ulaşılan bulguları desteklemek adına bir sonraki analiz aşaması *Kendall'ın tau-b (τb)* korelasyon analizi olarak gerçekleştirilmiştir. Daha öncede ifade edildiği üzere KTK, IRT ve MST temelli puanların sıralama amacıyla kullanılma ihtimallerinin yüksek olması sebebiyle *Kendall'ın tau-b (τb)* korelasyon katsayısının tespitine yönelik korelasyon analizi tercih edilmiştir. Analiz sonuçları incelendiğinde IRT ve MST yöntemleri yüksek korelasyon gösterirken KTK – IRT 3PL, KTK – MST-R ile KTK – MST-S düşük korelasyon göstermiştir. Bu bulgulardan elde edilen sonuçlar geleneksel KTK yönteminin kullanılması özellikle sıralamaya bağlı olarak öğrenci başarısına karar verilen sınavlarda çok ciddi oranlarda hatalı hesaplamalar yapılıyor olabileceğini ortaya koymaktadır. Örneğin; geçmesi gereken bir öğrenci için kalması yönünde karar verilmesi ya da kalması gereken bir öğrencinin geçmiş olması gibi. Benzer şekilde puan sıralaması ile yapılan işe alımlarda da aynı hatanın yapılıyor olması söz konusudur.

KTK, IRT ve MST yöntemlerinin karşılaştırılmasına yönelik gerçekleştirilen bir sonraki analiz adımı ise optimum modelin tespiti adına AIC (Akaike Information Criterion) değerlerinin hesaplanması için yapılan analizlerdir. Bu aşamada tüm maddelerinde "Eksik" veri barındırmasından kaynaklı olarak MST-S yöntemi için AIC değeri hesaplanamamıştır. MST-R yöntemi için ise sadece tüm katılımcıların yanıtlamış olduğu "Eksik" veri barındırmayan maddeler için AIC değeri hesaplanabilmiştir. Bu sınırlılıklar dahilinde/çerçevede elde edilen bulgulara göre sonuçlar yine önceki analiz bulgularını destekler/doğrular niteliktedir. KTK'nın optimum modelin tespiti yönündeki analizlerde de veriler ile daha az uyumlu olan model olarak görünmektedir. IRT ve MST birbirine yakın değerler izlemekle birlikte MST yönteminin optimum model olarak tespit

edildiği veri setleri de bulunmaktadır. Bu nedenle MST modellemelerinin sınav süreçlerine uygulanması durumunda CAT'e oldukça yakın ölçüm hassasiyetiyle yetenek kestirimi yapılabileceği yönünde umut vadetmektedir.

KTK, IRT ve MST yöntemlerinin karşılaştırılmasına yönelik son analiz ise "Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları"nın tespiti yönünde detaylandırılarak gerçekleştirilmiştir. Analiz sonuçları bu aşamada da önceki analiz aşamaları olan "Standart Hata Değerleri", "Korelasyon Katsayısı", "AIC (Akaike Information Criterion) Değerleri"ne yönelik analiz bulguları ile örtüşmektedir. Ulaşılan sonuçlarda KTK - IRT 3PL, KTK – MST-R ve KTK – MST-S karşılaştımalarında saçılım artmakta ve sıralamalar kaydadeğer oranlarda farklılaşmaktadır. IRT 3PL – MST-R ve IRT 3PL – MST-S karşılaştırmalarında ise saçılım azalmakta ve doğal olarak kuramlar arası sıralama farklarıda azalmaktadır. Analiz sonuçlarında tespit edilen bir diğer önemli bulgu da uç yetenek düzeylerindeki bireylerin sıralamalarında herhangi bir değişme olmamasıdır. Tüm veri setlerine ait detaylı tablo ve grafik gösterimleri bulgular bölümünde sunulmuştur. Açık ve uzaktan öğrenme sistemlerininde dahil olduğu tüm merkezi sınavlarda kararların sıralama üzerinden verildiği gerçeğinden yola çıkarsak ne denli önemli standart hata değerleri ile sonuca gidildiğini çok daha net bir biçimde ifade etmiş oluruz.

Araştırma sonuçları göre CAT yönteminin hukuki, mali, kültürel ve sosyal anlamdaki dezavantajlarına MST yöntemi aracığılıyla çözüm bulabilmek mükmün görünmektedir. Her üç kurama göre "Standart Hata Değerleri", "Farklı Yöntemlere Göre Elde Edilen Yetenek Ölçülerinin Korelasyon Katsayıları", "Farklı Yöntemlerin Verilere Ne Kadar Uyuştuğunu Gösteren AIC Değerleri" ve "Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları"nın karşılaştırıldığı analiz sonuçları da bu durumu destekler niteliktedir. Analiz bulguları incelendiğinde MST test sunum yöntemi hata değeri olarak IRT yöntemine çok yakın hata değerlerine ulaşıldığı görülmektedir. Bu bulguların ifade ettiği anlam şudur: MST yöntemine göre hazırlanmış bir test montajı CAT'in dezavantajlarının önüne geçmekle birlikte CAT'e yakın sonuçlar vererek KTK'nın sınırlılıklarını da aşmaktadır. Eğitim öğretim sistemleri yayın olarak KTK kültürünü benimsemiş durumda olduğundan CAT sisteminin toplumsal olarak kabul görmesi önemli bir zaman dilimini gerektirmektedir. Toplumsal anlamda bir değişiklik gerçekleştirilirken toplumun bu değişime hazır olması sürecin sağlıklı işlemesinin bir ön

koşulu olarak kabul edildiğinden MST modellemesinin sınav süreçlerine uyarlanması çok daha umut vericidir. MST yöntemi toplumun halihazırda alışık olduğu kültüre uygun bir test tasarımı olup özellikle CAT yönteminin kültürel uyumsuzluklarına çözüm niteliğinde ölçme ve değerlendirme imkânı sunmaktadır.

Tez çalışması kapsamında odaklanılan MST test sunum yöntemi kullanılarak elde edilen yetenek kestirimleri (puanlar) ile aynı verilerden elde edilen KTK ve IRT'ye dayalı yetenek kestirimleri arasında simülatif bir ortamda MST lehine anlamlı bir farklılık olup olmadığı yönündeki ana araştırma sorusundan yola çıkarak ulaşılan "Standart Hata Değerleri", "Korelasyon Katsayısı", "AIC değerleri" ve "Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları"na yönelik bulgular uzaktan öğrenme alanında yapılacak ölçme ve değerlendirme adına önemli veriler sunmaktadır. Araştırmada simülatif olarak gerçek hayat koşullarında karşılaşılabilecek tüm varyanslar dikkate alınarak analizlerin gerçekleştirilmiş olması pratikte karşılaşılabilecek olası sonuçlara yönelik önemli bulgular barındırmaktadır. Elde edilen sonuçlar sayesinde mevcut duruma ve diğer test sunum yöntemlerinin kullanıldığı durumlara dair fikir sahibi olma şansı bulunmaktadır. Literatürde bugüne kadar özellikle açık ve uzaktan öğrenme bağlamında ölçüm hassasiyetini konu alarak KTK, IRT ve MST yöntemlerini bir arada inceleyerek karşılaştıran çalışma bulunmaması tez araştırmasının öncü ve özgün yönü olarak nitelendirilebilir.

Tez araştırması kapsamında elde edilen analiz bulgularında/sonuçlarında MST yöntemlerinin düşük hata değerlerine sahip olması uzaktan eğitim sınavlarında da MST yöntemlerinin uygulanmasını hem kültürel anlamda hemde ölçüm kalitesi ve sınav güvenliği anlamında önemini vurgulamaktadır.

Bu tez araştırmasında analizler ölçme değerlendirme aşamasına kadar tüm süreçlerin olması gerektiği gibi olduğu varsayılararak (soru kalitesinin tam olduğu ve düzgün bir biçimde hazırlandığı vb.) gerçekleştirilmiştir. Soru kalitesinin tam olmadığı bir sistemde sağlıklı bir ölçme ve değerlendirmeden bahsedebilmek mümkün görünmemektedir. Bu nedenle sistemin başarısı için arka plandaki soru kalitelerinin etraflıca ele alınması dikkatle ve hassasiyetle üzerinde durulması gereken önemli bir konudur. Test teorilerine paralel olarak arka planda işe koşulması gereken değişkenler devreye girebildiği ve hedef kitlenin bu soru tipine uyumu sağlanabildiği noktada çok

daha sağlıklı ve güvenilir bir ölçme değerlendirme ortamından söz etmek mümkündür. Bu bağlamda değerlendirildiğinde/ele alındığında görülmektedir ki analiz bu sürecin (testing) sadece tek bir aşamasıdır. Doğru başarıyı doğru yetkinliği test edebilecek nitelikte test sorularının geliştirilmiş olması gerekmektedir. Bu niteliğin sağlanabilmesi ise ancak içerik uzmanları ve alan uzmanlarının inter disipliner çalışmasıyla mümkün olabilmektedir. Analiz öncesinde yapılan ölçme aracının hatası, ölçme aracını besleyen unsurların hatası, kapsam hatası, örneklem hatası gibi daha birçok hata ile analize kadar gelindiği gerçeğini de göz önüne aldığımızda durumun ehemniyetinin/öneminin çok daha dikkate değer boyuta ulaşabileceğinin göstergesidir. Bu nedenle Cansız Aktaş (2008)'ın da ifade ettiği gibi eğitim öğretim bağlamında öğretim yöntemleri ile ölçme ve değerlendirme alanındaki süreçlerin aynı gelişmişlik seviyesinde olması halinde hedeflenen amaca ulaşılması mümkündür. Bates'in (t.y., s. 79-80) kitabında "*Bir istatistik profesörünün sistemle imtihanı*" konulu senaryo örneğinde olduğu gibi istatistik dersinde örneklemin ne olduğunu sorgulama, problem çözme gibi konularda eğitilen öğrencilerin istatistiksel teknikleri ve formülleri ezberlemeleri beklenen bir sınavla test edilmelerinin sonucu olarak kopyaya yönelmeleri bu durumu özetleyen güzel bir örnek olarak sunulabilir. Tıpkı bu örnek olayda ifade edildiği gibi eğitim öğretim sistemi ile ölçme ve değerlendirme yöntemi aynı gelişmişlik seviyesinde olmadığında sonuç gerek öğrenen gerek öğreten ve gerekse yönetici açısından başarısız olarak nitelendirilebilir. Oysaki asıl sorun eğitim şekli ile ölçme ve değerlendirme yöntemi arasındaki uyumsuzluktan kaynaklanmaktadır.

Elbetteki kişinin şans eseri başarılı olup olmadığını ya da kişinin yeteneğini dikkate alabilen ölçme ve değerlendirme yöntemleri daha hassas ve daha hatasız ölçümler sunacaktır. Ancak bu durum birbiriyle zincirleme başka değişkenlerle, faktörlerle ve politikalarla alakalı bir konudur. Örneğin sınav güvenliği konusunda özellikle kopya ile ilgili etik kodlar bireylere önceden verilmiş olmalıdır. Öğretmen eğitimi, aile eğitimi vb. aracılığıyla öğrenene yapabilmenin önemi öğretilmeli ve bu kültür oluşturulmalıdır. Bireyin bu bilince erişebildiği bir kültürde kopya çektirmek mümkün değildir. Tüm bu hususlar dikkate alınarak sistemin bütün açıkları kapatılabildiği noktada test ortamı amaca uygun bir biçimde hazırlanmış olacak ve böyle bir sistemden elde edilen verilerle gerçekleştirilen analizlerde bu yönde çok daha net sonuçlar sağlayabilecektir.

Her üç kurama yönelik analiz sonuçları IRT ve MST'nin KTK'ya göre daha tercih edilebilir uygulamalar olduğunu gösterdi. Araştırma sonuçlarına göre MST test sunum yönteminin KTK'dan daha düşük standart hata değerleri ile daha doğru sıralama yapabiliyor olduğu bulgusu özellikle ölçüm hassasiyeti açısından ölçme ve değerlendirme alanına önemli bir katkı sunmaktadır. MST kapsamında IRT'ye yakın değerlere ulaşılmış olması ise söz konusu katkıyı desteklemektedir. Araştırma sonuçları çevrimiçi sistemlerin yanı sıra yüz yüze eğitim sunan fakat çevrimiçi sınav uygulayan sistemler içinde geçerli sonuçlar barındırmaktadır.

6. SONUÇ

Bireyin sahip olduğu beceriler; gelişim testleri, yetenek ve başarı testleri, zekâ testleri, davranışsal derecelendirme ölçekleri gibi bir takım eğitsel ve psikolojik ölçüm araçları aracılığıyla periyodik olarak araştırılmakta ve ölçülmektedir. Söz konusu testlerin yaşam içindeki rollerinin önemi dolayısıyla bu eğitsel ve psikolojik testlerin oluşturulması, tasarlanması ve değerlendirilmesi süreçleri çok daha önemli bir hal almaktadır. İyi hazırlanmış bir test modeli; test maddeleri ve yetenek arasındaki ilişkileri gerçeğe en yakın doğrulukta belirleyebilmektedir. Bir ölçümün en temel amacı; bireyin özelliklerini geçerli ve yeterli teorik modeller çerçevesinde güvenilir bir formda ölçmek ve bir testin uygulanmasından sonra elde edilen çıktıları bilimsel bir şekilde yorumlayabilmektir.

Geleneksel KTK uzun yıllardır sabit uzunluklu doğrusal kâğıt ve kalem (P&P: Paper and Pencil) testi olarak en yaygın kullanılan standart test sunma yöntemi olmuştur. Son zamanlarda teknolojinin ilerlemesiyle birlikte ölçme ve değerlendirme alanında bilgisayarlı uyarlanabilir test (CAT) ve çok aşamalı test (MST) gibi modern test sunum yöntemlerinin kullanıldığı uygulamaların popülaritesi hem teoride hem de pratikte artmaktadır. CAT modern test sunma yöntemi olarak hâkim uygulama olma rolünü devam ettirmektedir. Ancak CAT alan yazın ve tartışma bölümlerinde de detaylı biçimde belirtilen bir takım hukuki, mali, kültürel ve sosyal dezavantajları sebebiyle uzun süredir eleştiri altındadır. MST test sunum yöntemi ise CAT test sunum yönteminin bazı eleştirilerini ortadan kaldıran alternatif çok aşamalı bir testtir. MST ve CAT arasındaki en belirgin farklardan biri MST'de test sunumunun sınav katılımcılarına CAT sisteminde olduğu gibi madde madde uygulamak yerine modüller halinde aşamalı olarak uygulanmasıdır. CAT'in özel bir durumu olan MST test sunum yöntemi işleyiş biçimi açısından KTK ve IRT yöntemleri arasında bir uzlaşma olarak nitelendirilebilir.

Ölçme ve değerlendirmenin temel amacı olarak yukarıda ifade edilen koşulları hangi yöntemlerin (KTK, IRT, MST) karşılayabilecek nitelikte olduğu sorusu bu tez çalışmasının temel çıkış noktasıdır. KTK, IRT ve MST yöntemlerine göre gerçekleştirilen puanlamalarda (yetenek kestirimlerinde) standart hata bağlamında söz konusu yöntemlerin nasıl konumlandıkları araştırma kapsamında önemli görülmüştür. Bu

sebeple ana yönelimi çok aşamalı testlerin etkililiğini simülatif bir ortamda incelemek olan bu tez çalışmasında araştırma sorularından yola çıkılarak gerçekleştirilen analizlerde optimum yöntemin tespit edilmesine odaklanılmıştır.

Tez araştırması kapsamında öncelikle MST test üretme yöntemi temel alınarak tasarlanan simülasyon çalışması bir ölçme ve değerlendirme sürecinde karşılaşılabilecek olası tüm durumlar (dağılımın normal olduğu veya normalden sapma gösterdiği, homojen veya heterojen olduğu, örneklemin büyük ya da küçük olduğu durumlar gibi) göz önünde bulundurularak gerçekleştirilmiştir. Araştırma kapsamında MST yöntemi ile test üretmenin büyük sistemlerde geleneksel yöntemden daha düşük hata ile puanlama yapmaya imkân tanıyıp tanımadığı ortaya koymaya çalışılmıştır. MST yöntemi ile tasarlamış bir sınav ve bu sınavı alan adayların maddeleri işaretleme biçimleri araştırma soruları doğrultusunda simüle edildiğinde açık ve uzaktan öğrenme sisteminde farklı yöntem ve farklı örneklem büyüklüklerinde ne tür sonuçlar doğurabileceğinin tespiti bu araştırmanın özellikle maliyet ve güvenlik bağlamında en önemli çıktısı olarak değerlendirilebilir. Ayrıca MST'nin kitlesel eğitim hizmeti veren açık ve uzaktan öğrenme kurumlarında uygulanabilirliği ve olası sonuçları hakkında fikir sahibi olabilmek amacıyla gerçek durumu temsil eden simülatif verilerin kullanılması düşük maliyet ile araştırma sorularına yanıt bulunmasına olanak sağlamıştır. Simülasyon çalışmasının MST yöntemi ile tasarlamış bir sınavın açık ve uzaktan öğrenme sisteminde ne tür avantajlar sağlayabileceğinin önceden kestirilerek optimum sınav koşullarının ortaya konulması konusunda önemli ölçüde veri sağlayacağı düşünülmektedir. Dolayısıyla bu araştırma sonuçlarından elde edilen bulgular özellikle geniş ölçekli sınav hizmeti veren kurumlara düşük maliyetle optimum test yöntemini belirlemede rehber niteliğinde veriler sunmayı hedeflemesi bakımından da önem arz etmektedir.

Alan yazında bu alanda çalışmaları bulunan araştırmacıların da önemle üzerinde durduğu gibi test süreçlerindeki güvenlik ve maliyet konularında (Ockey, 2012, s. 347; Rotou, 2007; Weiss ve Kingsbury, 1984; Yan, Davier ve Lewis, 2014, s. 19) MST yönteminin bu simülasyon çalışmasında da CAT'e alternatif olarak değerlendirilebilecek bir test sunum yöntemi olduğunu destekler nitelikte bulgulara ulaşılmıştır. Ayrıca çoğu araştırmacı tarafından belirtilen MST avantajlarından (Han ve Guo, 2014, s. 119-122; Hendrickson, 2007; Luecht, 2003; Macken-Ruiz, 2008; Raborn ve Sarı, 2021; Yan, Lewis ve Davier, 2017, s. 3-20); test maddelerinin modül setlerinden oluşmasından kaynaklı

olarak uygulanmasının ve montajının daha kolay olması, test geliştirmek için CAT yöntemine göre daha düşük düzeyde çaba gerektirmesi, test spesifikasyonlarının ve özelliklerinin ayrıntıları üzerinde CAT'den daha fazla kontrole izin vermesi simülasyon çalışması kapsamında doğrulanırken madde seçim sürecinde istemci bilgisayarlara daha az yük getirmesi, CAT'in aksine, MST sınav katılımcılarının maddeler arasında ileri geri hareket etmelerine ve her bir modül içinde ilk yanıtlarını değiştirmelerine izin verilmesi gibi avantajlarının uyumu araştırma sınırlılıkları kapsamı dışında kalması sebebiyle sınanmamıştır.

Han'ın (2020) vurguladığı gibi bu çalışmada da MST'nin ölçüm verimliliği ve test optimizasyonunu hassasiyetle iyileştirmesi yönündeki avantajları bulunduğu görülmüştür. Wang, Zheng ve Chang'in (2014) çalışmalarında MST'nin bilgisayar tabanlı doğası, yeni madde biçimleri yönünden ifade ettikleri avantajlar bu çalışma ile uyum gösterirken ölçülebilen yeni beceri türleri, daha kolay ve daha hızlı veri analizi, madde yanıt süresi gibi zengin davranış verilerini toplamadaki avantajları yönünden uyumu çalışmanın sınırlılıkları gereği karşılaştırılamamıştır. Wang, Chen ve Jiang (2020) tarafından belirtilen doğrusal testlerin gerektirdiğinden daha az madde kullanarak daha doğru yetenek (θ: theta) tahminleri sağlayabilmesi yönündeki avantajı ise çalışmada bu yönde bir analiz işlemi uygulanmamış olmasından dolayı uyumlulukları test edilememiştir.

Reese, Schnipke ve Luebke (1999)'in testlerin optimal bir şekilde birleştirilmesi için stratejilere odaklanan çalışmaları; dikkatlice oluşturulmuş ve içerik dengeli iki aşamalı bir testin, CAT ve mevcut kağıt ve kalem testinden daha iyi performans gösterdiğini ortaya koymuştur. Bu bağlamda çalışma kapsamında 1-3-3 çok aşamalı test deseni kullanılarak oluşturulan simülasyon sonuçları ise KTK'dan (mevcut kağıt ve kalem testinden) daha iyi performans göstermesi bakımından uyumluluk gösterirken CAT'ten sadece bir kaç veri setinde daha iyi performans gösterdiği sonucuna ulaşılmıştır. Ancak araştırma analizlerinde multistage yöntemin CAT'e önemli derecede yaklaşan sonuçlar vermesi umut vaad eden bir bulgudur.

Bulgular ve tartışma bölümlerinde detaylı olarak ifade edildiği üzere KTK, IRT ve MST (MST-R ve MST-S) yöntemleri araştırma soruları temelinde "Standart Hata Değerleri", "Farklı Yöntemlere Göre Elde Edilen Yetenek Ölçülerinin Korelasyon

Katsayıları", "Farklı Yöntemlerin Verilere Ne Kadar Uyuştuğunu Gösteren AIC Değerleri" ve "Farklı Kuramlara Göre Elde Edilen Sıralamaların Farkları" bağlamında analiz edilerek karşılaştırılmıştır.

Analiz sonuçlarına göre KTK, IRT 3PL ve MST (MST-R ve MST-S) yöntemleri birbirleri ile ölçüm hassasiyeti bakımından karşılaştırıldığında IRT 3PL ile MST (MST-R ve MST-S) birbiri ile yakın bir seyir izlerken KTK önemli derecede farklılaşmakta ve KTK veriler ile daha az uyumlu olan model olarak görünmektedir. Elde edilen bulgularda Z puanı üzerinden hesaplanan KTK standart hata değerlerinin 0.90'ı aşan oranlarla 1'e yaklaşması yapılan ölçümlerde olası bütün varyans kadar hatanın söz konusu olabileceği anlamını taşımaktadır. Bu durum KTK yöntemi ile gerçekleştirilen bir puanlamada gerçekte başarılı kabul edilmesi gereken bir bireyin başarısız olduğu yönünde karar verilmesi ya da aksine başarısız bir bireyin de başarılı olduğu yönünde karar verilmesine yol açabilecek dezavantaja sahip olduğunu göstermektedir. Geleneksel KTK yönteminin ülkemizdeki eğitim kademelerinin büyük çoğunluğunda yaygın olarak kullanıldığı göz önünde bulundurulduğunda bu yöntemin barındırdığı yüksek standart hata değerleri ölçme ve değerlendirme süreçlerinde ciddi anlamda hatalı ölçümler yapılma olasılığının yüksek olduğunun önemli bir göstergesidir. Dolayısıyla bu araştırmanın en çarpıcı sonuçlarından biri mevcut yöntemin (KTK) yeteneği tahmin etmede ciddi anlamda yetersiz kaldığını ortaya koymasıdır. Ayrıca tez araştırması kapsamında ulaşılan sonuçlarda ölçme ve değerlendirme aşamasında hassas ölçümler yapabilen yöntemlere (IRT/MST) olan gereksinim analiz bulgularının ortak paydası olarak tespit edilmiştir. Bu sonuçların tamamı standart hata değerlerinin yanı sıra korelasyon katsayısı, AIC değerleri ve farklı kuramlara göre sıralama farkları hesaplamaları analizlere dahil edilerek doğrulanmış ve önemli bulgular elde edilmiştir.

Bilindiği üzere gerek dünya ölçeğinde gerek ülkemizde çevrimiçi ölçme ve değerlendirmenin geniş kitlelere uygulanan sınavlarda yaygın olarak kullanılmaya başlanmasının temelinde 2019 yılında başlayıp tüm dünyayı saran, bu çalışmanın yapıldığı tarihte de kısmen devam eden Covid-19 pandemisinin yaşam koşullarına getirdiği zorunluluklar yatmaktadır. Ülkemizde büyük yıkımlara sebep olan 6 Şubat 2023 depremi sonucunda da ivedilikle uzaktan eğitime geçilmesi dolayısıyla sınavların çevrimiçi ortamlarda gerçekleştirilmesi kararı alınmıştır. Son dönemlerde yaşanan salgın ve doğal afetler yaşamın pek çok alanında olduğu gibi ölçme ve değerlendirme

yöntemlerinde de zorunlu bir dönüşümü gerekli kılmıştır. Olağanüstü durumlar nedeniyle sınavları hızlı bir biçimde çevrimiçi ortamlara taşımak durumunda kalınması çevrimiçi sınavları ve bu sınavların gerçekleştirilebilmesi için güvenlik, maliyet, ölçüm kalitesi gibi gerekli kriterlerin sağlanabilmesi konularını gündeme getirmiştir.

Salgın ve doğal afetler gibi kendi başına mevcut sonuçları ağır olan böylesi durumlarda eğitim öğretimin yürütülmesinde yaşanacak aksamaların minimuma indirilmesi ölçme ve değerlendirme başta olmak üzere öğretim süreçlerinin çevrimiçi ortamlarda yapılabilmesi konusunda hazır olunmasına bağlıdır. Bir konuda hazır olunması hali hazırda işleyen bir sistemin bulunması anlamına gelmektedir. Bu sebeple mücbir sebepler var olmadan önce tüm yönleriyle sağlıklı bir biçimde işleyen çevrimiçi ölçme ve değerlendirme sisteminin bulunması olağanüstü durumlarda süreçlerin aksamadan yürümesine katkı sağlayacağı düşünülmektedir. Örneğin çevrimiçi ölçme ve değerlendirmeye imkân tanıyan sınavların zorunluluk gerektirmeyen zamanlarda da aktif bir biçimde çevrimiçi platformlarda uygulanması sistemin normal olmayan zamanlarda hazır bulunmasına olanak tanıyacaktır. Bu şekilde işleyen bir sistemin var olması teknik alt yapıdan, tutunda sürecin paydaşlarının hazırbulunuşluklarına kadar etkili bir avantaj sağlayacaktır. Özellikle pandemide yaşanan ani geçiş süreci teknik aksamaların yanı sıra öğrenci velileri dahil tüm paydaşlarda adaptasyon sorununa yol açmıştır. Bu tür aksamaların yaşanmaması için sürecin her dönemde aktif bir biçimde işlemesi önemli rol oynamaktadır.

Çoktan seçmeli test sunum yöntemi ile çok sayıda sınav uygulayan ülkelerde, hızlı akan günlük hayat süreçlerinde çevrimiçi sınavlar bir zorunlulukken sınav güvenliği çözülmesi gereken ana problem olarak karşımıza çıkmaktadır. Bu sebeple eğitim öğretim sistemlerinin bütün kademelerinde hassas ölçüm yapabilen ölçme ve değerlendirme yöntemlerine olan ihtiyacın önemi tüm paydaşlar tarafından kabul görmüş durumdadır. Ölçme ve değerlendirme süreçlerinde genel geçer nitelikteki bu gerçeklikten (hassas ölçüm) yola çıkılarak yapılan analiz sonuçları göstermektedir ki; bu ihtiyaca kültürel, hukuki, mali, güvenlik ve sosyal anlamda en kısa vadedeki çözümü MST yöntemi vadetmektedir. MST bu özelliği nedeniyle kültürel ve hukuki anlamda CAT yönteminin bazı dezavantajlarına da çözüm getirebilme potansiyeline sahip olduğundan tercihinde önceliğine şans verilmesi gereken bir test sunum yöntemi olarak değerlendirilmesi mümkün görünmektedir. Bu sebeple özellikle ölçme ve değerlendirmede MST gibi

hassas ölçümler yapabilen test sunum yöntemlerinin süreçlere entegre edilmesini sağlayan araştırma ve çalışmaların hem dünya ölçeğinde hem de ülkemiz bağlamında önemli katkılar getireceği düşünülmektedir. Özet olarak sunmak gerekirse;

- MST yöntemi KTK'ya göre çok daha düşük standart hata değerlerine sahiptir.
- MST'ye göre elde edilen yetenek ölçüleri IRT ile daha yüksek korelasyona sahiptir.
- MST yöntemi veriler ile IRT'ye yakın değerlerde uyum göstermekle birlikte IRT'ye oranla daha yüksek uyum gösterdiği veri grupları da bulunmaktadır.
- MST yöntemi puan sıralama farkları farkları bağlamında da KTK'dan uzaklaşarak IRT'ye daha yakın sonuçlar vermektedir.
- MST, IRT'ye daha yakın hassasiyetle ölçme yapmaktadır.

7. ÖNERİ

Araştırma sürecinde her ne kadar farklı kuramlar temelinde farklı örneklem koşulları göz önünde bulundurularak belirli bir çeşitlilik ve derinlikte detaylı analizler sonucunda bulgular elde edilmeye çalışılmış olsa da karşılaşılan birtakım sınırlılıklar ve beraberinde getirdiği gereksinimler bağlamında bu tez çalışması gelecekteki araştırmalar için aşağıda belirtilen önerileri barındırmaktadır:

- **Geniş Madde Havuzlarında Bireye Özgü Test Sunumunun Çeşitlendirilerek Çalışılması:** MST yöntemi bu çalışmada 70 soruluk dar bir madde havuzundan elde edilen verilerle gerçekleştirilmiştir. Uygulanan analizler sonucunda tüm yöntemler için madde havuzunun bu sınırlılıklarında önemli bulgular elde edilmiştir. Özellikle MST test sunum yönteminin yetenek kestiriminde IRT yöntemine yakınlaşırken KTK'dan uzaklaşması tez çalışması kapsamında ulaşılan en önemli ve umut vaad eden bulgudur. Bu sonuçlar bağlamında değerlendirildiğinde MST test sunum yönteminin geniş soru havuzlarında uygulanmasının ölçüm kalitesi ve güvenirliğini artıracağı düşünülmektedir. Çünkü geniş madde havuzlarında soru sayısının fazla olmasından kaynaklı olarak bireye özgü test sunumunu çeşitlendirmek çok daha mümkün görünmektedir. Bu sayede aynı yetenek düzeyindeki bireylere aynı zorlukta çok daha farklı soru yapıları ile karşılaşma olanağı tanınarak ölçüm kalitesi ile sınav güvenliği konusunda çok daha nitelikli sonuçlar elde edilebileceği düşünülmektedir. Bu nedenle araştırma sınırlılıklarına dayanılarak tez çalışmasının en önemli önerilerinden biri şudur: gerek uygulayıcılar gerekse araştırmacılar açısından geniş madde havuzlarında çalışmak ölçme ve değerlendirmede sağlıklı sonuçlar elde edebilmenin en önemli anahtarıdır. Bu çalışmada dar bir madde havuzunda gerçekleştirilen analizler sonucu ulaşılan oldukça çarpıcı sonuçlar, geniş madde havuzlarında çok daha dikkate değer boyutlara ulaşılabileceğinin önemli bir göstergesidir.

- **MST Yönteminin Kullanımında İstatistiksel Analizlerin Gerçekleştirildiği Mevcut Programların Yazılımsal Kısıtlılıklarının Çözümüne Yönelik Çalışmaların Teşvik Edilmesi:** MST yönteminin kullanımında özellikle üzerinde durulması gereken bir diğer öneri ise istatistiksel analizlerin

gerçekleştirildiği mevcut programların yazılımsal kısıtlılıkları ile ilgilidir. MST test sunum yönteminin yapı itibariyle çok fazla "Eksik" veri barındırmasından kaynaklı olarak halihazırdaki analiz programlarında istatistiksel analizlerin gerçekleştirilmesi sırasında ciddi anlamda eksiklikler bulunmakta ve analiz aşamasında ciddi kısıtlılıklara sebebiyet vermektedir. MST'nin yapısal anlamdaki bu farklılığından dolayı mevcut analiz programları aracılığıyla analizleri gerçekleştirmek mümkün olmamakta ve gerekli analizleri uygulayabilmek için elle hesaplamak durumunda kalmak önemli zaman kayıplarına yol açmaktadır. Söz konusu istatistiksel analizleri gerçekleştirebilecek mevcut programların tamamı "Eksik" değerleri silmeye veya ağırlıklı ortalamaya yuvarlamaya odaklı bir algoritmayla tasarlanmış durumdadır. Programlar "Eksik" değerleri hata olarak algılamakta ve analizlerini bu yönde gerçekleştirmektedir. Dolayısıyla bu algoritmada hazırlanmış bir analiz uygulaması MST yapısında üretilen verilerin analizini gerçekleştirememektedir. Bilindiği üzere analiz bir sürecin çıktılarının tespiti için olmazsa olmaz bir aşamadır. Bu derece kaydadeğer bulguların emek ve zaman kayıplarına yol açmadan çok daha pratik bir biçimde elde edilebilmesini sağlayan programlara olan gereksinim önemli bir eksiklik ve kısıtlıklık olarak tespit edilmiştir. Bu sebeple MST yapısına uygun formda "Eksik" değerleri dikkate almadan (göz ardı ederek) hesaplamaları gerçekleştirebilecek algoritmada tasarlanmış bir analiz programının yazılımcılar tarafından hazırlanması önerilmektedir. Bu yapıda konumlanmış veri setlerinin analizi için programlanmış bir uygulama analiz aşamasında işlemlerin sağlıklı olarak gerçekleştirilebilmesi adına özellikle emek ve zaman maliyeti açısından gereksinim duyulan önemli bir araç konumundadır. Yazılımsal anlamda söz konusu ihtiyacın çözümüne yönelik çalışmaların teşvik edilmesinin kaydadeğer katkılar sağlayacağı düşünülmektedir.

- **CAT ve MST Yöntemlerini Birleştiren Yapıdaki Hibrit Tasarımlara Odaklanılması:** MST'nin CAT yöntemine alternatif olarak son zamanlarda popülerliğinin giderek artması her iki yöntemin avantajlarının birleştirilerek karma test sunum yöntemlerinin ortaya çıkmasına yol açmaktadır. Son zamanlarda MST ve CAT test sunum yöntemlerinin avantajlarını birleştiren ve sınırlamalarını ortadan kaldıran "maMST (Mixed Adaptive Multistage Testing)" ve "OMST (On-The-Fly Assembled Multistage Adaptive Testing)'' gibi karma

test sunum yöntemleri (Raborn ve Sarı, 2021; Zheng ve Chang, 2015) gündeme gelmeye başlamıştır. Xu ve arkadaşları (2021) çalışmalarında MST'yi MIRT (Multidimensional Item Response Theory) çerçevesi içinde tasarlamanın çok boyutlu değerlendirmelerin uygulanmasında hem MST'nin hem de MIRT'nin avantajlarından tam olarak yararlanmak için kritik öneme sahip olduğu vurgusunu yapmaktadırlar. Bu durum bahsi geçen karma yöntemler aracılığıyla mevcut MST ve CAT test sunum yöntemlerinin uygulanma biçimlerinin bir adım öteye taşınarak çok daha hassas ölçümlere yönelinmesi ölçme ve değerlendirme alanında test sunum yöntemlerine bakış açısının değişme ve gelişme sürecinde olduğunun önemli bir göstergesidir. Bu şekilde CAT ve MST yöntemlerini birleştiren yapıdaki hibrit tasarımların çeşitlenerek artması gelecekteki test geliştirme gereksinimlerine yanıt verebilecek nitelikte esnekliğe sahip uyarlamalı (adatif) test tasarımları için umut verici bir yön olabilir. Modern test sunum yöntemlerinin bu derece gelişim ve değişim gösterdiği günümüzde uzaktan eğitim sınavlarında da KTK yöntemi yerine daha hassas ölçümler gerçekleştirebilen yöntemlere (MST, CAT vb.) yönelinmesi ölçme ve değerlendirme açısından önemli bir adım niteliğindedir. Bu sebeple gerek uygulayıcılar gerekse araştırmacıların bu yönde çalışmalar yapmasının teşvik edilmesi önerilmektedir.

- **Yeni Teknolojilerle Desteklenmesi:** MST'nin yapay zekâ, makine öğrenmesi, blokzincir uygulamaları gibi yeni teknolojilerle desteklenerek zenginleştirilmesi ve güçlendirilmesi önemlidir. Bu yönde atılacak adımların MST'nin gelişimi ve sürece kazandırılması açısından katkı sağlayabileceği düşünülmektedir.

- **Gelecekte Gerçekleştirilecek Araştırmalarda Kişi Sayısının Az Olduğu Sınavlarda Dikkatli Olunmalı:** Gelecekte gerçekleştirilecek araştırmalarda kişi sayısının az olduğu (Örneğin; 100 kişilik veri setinde IRT 3 PL hata değeri hesaplamasında sağlıklı kestirimler yapılamamıştır.) sınavlarda, madde havuzunda yüksek düzeyde bilgilendirici maddelerin olmamasından dolayı ileride yapılan hesaplamalarda özellikle dikkatli olunması önerilmektedir.

- **Test Geliştirme Sürecinde Dikkatli Olunmalı:** Testlerin oluşturulması, tasarlanması ve değerlendirilmesi süreci önemlidir. Testlerin iyi hazırlanması, test maddeleri ile yetenek arasındaki ilişkilerin doğru bir şekilde belirlenebilmesini sağlamaktadır. Elde edilen bulgular göstermektedir ki MST test sunum yöntemi ile doğrusal testlere kıyasla daha az madde kullanarak daha doğru yetenek tahminleri sağlanabilmektedir. Bu bilgi, testlerin tasarımında ve puanlamasında MST yönteminin kullanımını teşvik edebilir. Araştırma, MST'nin bilgisayar tabanlı doğası ve zengin davranış verileri toplama avantajlarına vurgu yapmaktadır. Örneğin elde edilecek bilgilerle sadece test sorularına verilen yanıtlarla sınırlı kalmayacak başka davranış özelliklerini (Katılımcı; bir soruda ne kadar düşünüyor? işaretlediği seçeneği değiştirmiş mi?) ortaya koyabilme imkânı sunmaktadır. Bu şekilde davranış profillerini çizerek de adayın yeteneği hakkında sıralaması hakkında farklı fikirler sunabilecek kapasiteye sahiptir. Bu sebeple daha zengin bir veri ve yetenek ölçümüne olanak tanımaktadır. Bu avantajlar göz önünde bulundurularak, MST'nin daha fazla veri analizi ve yetenek ölçümü için kullanılmasının teşvik edilmesi somut sonuçlar elde edilmesi açısından önemlidir.

- **Modern Test Sunum Yöntemleri Dikkate Alınmalı:** Ölçme ve değerlendirme aşamasında geleneksel kâğıt ve kalem testlerinin yanı sıra bilgisayarlı uyarlanabilir test (CAT) ve çok aşamalı test (MST) gibi modern test sunum yöntemleri de göz önünde bulundurulmalıdır. Bu yöntemler, test sunumu sürecinde avantajlar sağlayabilir. Araştırmanın bulguları, geleneksel kağıt ve kalem testlerinin (KTK) bazı dezavantajlarının (özellikle yüksek standart hata değerleri ve doğruluk sorunları bağlamında) olduğunu ortaya koymaktadır. Bu nedenle, eğitim sistemindeki test uygulamalarında KTK yönteminin dikkatlice değerlendirilmesi ve bu sorunları azaltacak alternatif yöntemlere geçiş yapılmasının teşvik edilmesi önemlidir.

- **MST Yöntemi ve Avantajları Araştırılmalı:** Araştırmada MST yöntemiyle ilgili olumlu sonuçlar elde edilmiştir. MST'nin test sunumu sürecinde kolaylık sağladığı, test geliştirmek için daha düşük çaba gerektirdiği ve daha fazla kontrole izin verdiği belirtilmiştir. Bu avantajlar göz önünde bulundurularak MST

yönteminin kullanımı tüm paydaşlar (test geliştiriciler, araştırmacılar, yazılımcılar, yöneticiler, karar alıcılar, öğretenler vb.) tarafından daha fazla araştırılmalıdır.

- **Ölçüm Hassasiyeti Göz önünde Bulundurulmalı:** Araştırma bulgularına göre KTK, IRT ve MST yöntemleri arasında ölçüm hassasiyeti bakımından farklar bulunmaktadır. KTK yöntemi, diğer yöntemlere göre daha fazla hata içerebildiğinden ölçüm hassasiyetinin iyileştirilmesi için diğer yöntemlere (IRT, MST) odaklanılması ölçme ve değerlendirme süreçlerine katkı sağlayabilir.

- **Optimum Test Yöntemi Belirlenmeli:** Araştırma sonuçlarına dayanarak, özellikle geniş ölçekli sınav hizmeti veren kurumlara düşük maliyetle optimum test yöntemini belirleme konusunda rehber niteliğinde veriler sunulmuştur. Bu yönde yapılan analizler ve bulgular dikkate alınarak optimum test yöntemi belirlenebilir.

- **Test Güvenliği ve Maliyet Konuları Ele Alınmalı:** MST yöntemi, test güvenliği ve maliyet konularında avantajlar sunabilir. Bu avantajlar göz önünde bulundurularak MST'nin CAT'e alternatif olarak kullanılması değerlendirilebilir. Araştırma, MST yönteminin özellikle açık ve uzaktan öğrenme sistemlerinde kullanılabilirliği ve maliyet etkinliği açısından olumlu sonuçlar verdiğini göstermektedir. Bu nedenle, bu tür eğitim kurumlarında MST'nin kullanımı ve uygulanması değerlendirilebilir.

- **Daha Fazla Araştırma Yapılmalı:** Verilen önerilerin daha fazla araştırma gerektirdiğini unutmamak önemlidir. Ölçme ve değerlendirme alanında yeni gelişmeler ve yöntemler dikkate alınarak, test süreçlerinin iyileştirilmesi için daha fazla araştırma yapılmalıdır. Örneğin adaylar hem geleneksel yöntemle (KTK) hemde MST ile sınanarak sonuçları karşılaştırılabilir. Bu anlamda özellikle deneysel uygulamaların katkısı önemlidir.

- **Değişim Yönetimi Önemli**: Değişimin çoğunlukla dirençle karşılanan bir süreç olması sebebiyle tüm paydaşların (politika yapıcılar, mevzuatları dizayn edenler, karar alıcılar, yöneticiler, öğretenler, öğrenciler, uygulayıcılar vb.) ölçme ve

değerlendirmedeki bu değişimin gerekliliği konusunda ikna edilmesi önemlidir. Bu nedenle en iyi ikna yöntemi olarak somut örneklerin arttırılması ve paydaşlara sunulması gerekmektedir.

- **Farklı Disiplinlerde Sınanmalı:** MST'nin farklı disiplinlerde sınanmasının özellikle sınav içerikleri ve kapsamlarına ilişkin daha net veriler elde edilmesine imkân sağlayacağı düşünülmektedir.

- **İş Dünyasında Sınanmalı:** Sınavların doğasının bu yapıya evrilmesi önemli bir kültürdür. İş dünyasının öğrenme gelişim programlarında iş başı eğitimleri sırasında bu önerilerin verilmesi ve onların deneyimleri ile bu tür çalışmaların daha hız kazanarak desteklenmesi sağlanabilir.

KAYNAKÇA

Adams, W. K. ve Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert like thinking. *International journal of science education, 33*(9), 1289-1312. doi: 10.1080/09500693.2010.512369

Adedoyin, O. ve Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, *3*(4), 992–1011. Erişim adresi: https://archive.aessweb.com/index.php/5007/article/view/2471

Adom, D., Mensah, J. A. ve Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education, 9*(1), 109-119. doi: 10.11591/ijere.v9i1.20457

Akindele, B. P. (2003). *The development of an item bank for selection tests into Nigerian universities: An exploratory study* (Unpublished PhD Thesis) University of Ibadan, Nigeria.

AL-khadher, M. M. A. ve Albursan, I. S. (2017). Accuracy of measurement in the classical and the modern test theory: An empirical study on a children intelligence test. *International Journal of Psychological Studies, 9*(1), 71-80. http://dx.doi.org/10.5539/ijps.v9n1p71

An, X. ve Yung, Y. F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *SAS Institute Inc. SAS364-2014, 10*(4), 1-14. Erişim adresi: https://www.researchgate.net/profile/Steven-Kator-Iorfa/post/What-are-the-advantages-and-disadvantages-of-item-response-theory-compared-to-conventional-models-in-the-psychology-domain/attachment/59d63c6979197b8077999659/AS%3A415567514226689%401476090424959/download/Item+Response+theory.pdf

Anderson, J., Kearney, G. E. ve Everett, A. V. (1968). An evaluation of Rasch's structural model for test *Items*. *British Journal of Mathematical and Statistical Psychology, 21*(2), 231–238. doi:10.1111/j.2044-8317.1968.tb00411.x

Ani, E. N. (2014). *Application of item response theory in the development and validation of multiple choice test in economics* (Unpublished Master Thesis). Department of Science Education, University of Nigeria, Nsukka, Nigeria.

Ariel, A., Veldkamp, B. P. ve Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement, 30*(3), 204–215. doi:10.1177/0146621605284350

Ariel, A., Veldkamp, B. P. ve van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement, 41*(4), 345–359. Erişim adresi: https://www.jstor.org/stable/20461767

Arndt, S., Turvey, C. ve Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearmans r versus Kendalls tau correlation. *Journal of Psychiatric Research, 33*(2), 97–104. doi:10.1016/s0022-3956(98)90046-2

Arnold, M. E. (1996). *Influences on and limitations of classical test theory reliability estimates*. (Report No. 142). New Orleans: ERIC.

Ayala, R. J. (2009). *The theory and practice of item response theory*. New York/London: The Guilford Press.

Ayanwale, M. A. ve Adeleke, J. O. (2020). Efficacy of Item Response Theory in the validation and score ranking of dichotomous response mathematics achievement test. *Bulgarian Journal of Science and Education Policy, 14*(2), 260-285. Erişim adresi:https://www.researchgate.net/publication/349038293_EFFICACY_OF_ITEM_RESPONSE_THEORY_IN_THE_VALIDATION_AND_SCORE_RANKING_OF_DICHOTOMOUS_RESPONSE_MATHEMATICS_ACHIEVEMENT_TEST

Baehr, M. (2005). Overview of Assessment. In *Program assessment handbook* içinde (s. 1-10). Lisle, IL: Pacific Crest, Inc. Erişim adresi: https://pcrest.com/LO/PA/PAI_files/PAI/Kettering_PAS_May_31_June_1_2007/Program%20Assessment%20Handbook.pdf#page=11

Baker, F. B. (2001). *The basics of item response theory* (2.ed). USA: ERIC Clearinghouse on Assessment and Evaluation. Erişim adresi: http://ericae. net/irt/baker

Baker, F. B. ve Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker, Inc.

Baldwin, D. A. (1997). The concept of security. *Review of International Studies, 23*(1), 5-26. Erişim adresi: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/67188B6038200A97C0B0A370FDC9D6B8/S0260210597000053a.pdf/concept_of_security.pdf

Barnard-Brak, L., Lan, W. Y. ve Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: A 4PL item response theory examination. *Studies in Educational Evaluation, 56,* 1-7. https://doi.org/10.1016/j.stueduc.2017.11.002

Barrada, J. R., Olea, J. ve Abad, F. J. (2008). Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. *The Spanish Journal of Psychology, 11*(2), 618–625. Erişim adresi: https://www.redalyc.org/pdf/172/17213016026.pdf

Barton, M. A. ve Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Rapor No: 81-20). New Jersey: Educational Testing Service (ETS). https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Bates, A. W. T. (t.y.). *Dijital çağda öğretim* (M. Adnan ve Y. G. Güven Ed. ve M. Adnan Çev.). Erişim adresi: https://pressbooks.bccampus.ca/tonybates/front-matter/introduction/

Baylis, J. (2020). The globalization of world politics: An introduction to international relations. USA: Oxford university press. Erişim adresi: https://books.google.com.tr/books?hl=tr&lr=&id=Y1S_DwAAQBAJ&oi=fnd&pg=PP1&dq=The+Globalization+of+World+Politics&ots=uLMU0L3VgU&sig=iUfaRrM2LqThR21Rd7ob8QxcAkE&redir_esc=y#v=onepage&q=The%20Globalization%20of%20World%20Politics&f=false

Berger, S., Verschoor, A. J., Eggen, T. J. H. M. ve Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education*, *4*(January). https://doi.org/10.3389/feduc.2019.00001

Betz, N. E. ve Turner, B. M. (2011). Using item response theory and adaptive testing in online career assessment. *Journal of Career Assessment, 19*(3), 274-286. https://doi.org/10.1177/1069072710395534

Bichi, A. A., Embong, R., Mamat, M. ve Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: a review of empirical studies. *Australian Journal of Basic and Applied Sciences, 9*(7), 549-556. doi:10.13140/RG.2.1.1561.

Binh, H. T. ve Duy, B. T. (2016, September). Student ability estimation based on IRT. In 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS) (pp. 56-61). Erişim adresi:https://www.researchgate.net/publication/308361140_Student_ability_estimation_based_on_IRT

Birnbaum, A. (2008). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Ed.). *Statistical theories of mental test score*. USA: Information Age Publishing Inc.

Blass, E. (2007). *Talent management: Maximising talent for business performance*. London/Berkhamstead: CMI Ashridge Consulting. Erişim adresi: https://eoe.leadershipacademy.nhs.uk/wp-content/uploads/sites/6/2019/04/1237115518_RBgM_maximising_talent_for_business_performance.pdf

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and practice, 16*, 21–23. doi:10.1111/j.1745-3992.1997.tb00605.x.

Bollen, K. ve Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Bond, T. G. ve Fox, C. M. (2015). *Applying the rasch model fundamental measurement in the human sciences* (3th ed.). New York and London: Routledge Taylor & Francis Group.

Bond, T. G., Yan, Z. ve Heene, M. (2020). *Applying the rasch model fundamental measurement in the human sciences* (4th ed.). New York and London: Routledge Taylor & Francis Group.

Braun, H., Kanjee, A., Bettinger, E. ve Kremer, M. (2006). *Improving education through assessment, innovation, and evaluation*. Cambridge, MA: American Academy of Arts and Sciences.

Brown, J. D. (2012). Classical test theory. G. Fulcher ve F. Davidson (Editörler), *The routledge handbook of language testing* içinde (s. 323-335). New York/London: Routledge.

Buzan, B. (2008). *People, states & fear: An agenda for international security studies in the post-cold war era* (2nd ed.). UK: ECPR classics press. Erişim adresi: https://books.google.com.tr/books?hl=tr&lr=&id=sURLAQAAQBAJ&oi=fnd&pg=PA1&dq=Barry+Buzan,+People,+States+and+Fear:+An+Agenda+for+International+Security+Studies+in+the+Post-Cold+War+Era,+Boulder:+Lynne+Rienner+Pub,+1991,+s.+7.&ots=_35YFT97vw&sig=IPUBoxV6x6U-_GI3jS2CB7RomAk&redir_esc=y#v=onepage&q&f=false

Cansız Aktaş, M. (2008). *Öğretmenlerin yeni ortaöğretim matematik öğretim programının ölçme değerlendirme boyutuna bakışlarının incelenmesi* (Doktora Tezi). Fen Bilimler Enstitüsü, Karadeniz Teknik Üniversitesi, Trabzon.

Cappelleri, J. C., Lundy, J. J. ve Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648-662. http://dx.doi.org/10.1016/j.clinthera.2014.04.006

Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education, 47*(8), A47. Erişim adresi: https://anadolu.summon.serialssolutions.com/#!/search?ho=t&include.ft.matches=f&l=en&q=ETS%20FINDS%20FLAWS%20IN%20THE%20WAY%20ONLINE%20GRE%20RATES%20SOME%20STUDENTS%20

Cavanaugh, J. E. ve Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics, 11*(3), 1-11. https://doi.org/10.1002/wics.1460

Chang, H. H. (2004). Understanding computerized adaptive testing: from Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences,* (pp. 117-133). Thousand Oaks/London/New Delhi: Sage Publications. Erişim adresi: https://ebookcentral.proquest.com/lib/anadolu/reader.action?docID=1995660).

Chang, H. H. ve Ying, Z. (2007). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*(3), 441-450. doi: 10.1007/S11336-007-9047-7

Chang, H. H. ve Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*(3), 387-398. Erişim adresi: https://link.springer.com/content/pdf/10.1007/BF02294991.pdf

Chen, S. Y., Ankenmann, R. D. ve Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*(2), 129-145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x

Chen, P. Y. ve Popovich, P. M. (2002). Correlation: Parametric and nonparametric measures. Thousand Oaks, CA: Sage Publications.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F. ve Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. Multivariate Behavioral Research, 36(4), 523-562. https://doi.org/10.1207/S15327906MBR3604_03

Chikezie, I. J. ve Joseph, E. U. (2016). Group and item invariance of item difficulty parameter based on item response and classical test theories. 3, 9-21. Erişim adresi: https://www.earnia.org/e4e356ef6f4f9e939e06f04e06f60ef3ef0560f0de6/GROUP%20AND%20ITEM%20INVARIANCE%20OF%20ITEM%20DIFFICULTY%20PARAMETER%20BASED%20ON%20ITEM%20RESPONSE%20AND%20CLASSICAL%20TEST%20THEORIES.pdf

Cohen, J., Chan, T., Jiang, T. ve Seburn, M. (2008). Consistent Estimation of Rasch Item Parameters and Their Standard Errors Under Complex Sample Designs. *Applied Psychological Measurement, 32*(4), 289–310. https://doi.org/10.1177/0146621607300047

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics* (Doktora Tezi). ProQuest Dissertations & Theses Global veri tabanında erişildi. (Order No. 3141396). Erişim adresi: https://www.proquest.com/dissertations-theses/empirical-comparison-item-response-theory/docview/305067822/se-2

Crocker, L. ve Algina, J. (2008). Introduction to classical and modern test theory. In M. Baird, M., Staudt, M. & Strans (Ed.), *Cengage Learning*. USA: Cengage Learning.

Davey, T. ve Nering, M. (2002). Controlling item exposure and maintaining item security. In C. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 165-191). Mahwah: Lawrence Erlbaum Associates.

Dawson, B. ve Trapp, R. G. (2004). Basic&clinical biostatistics (4th Ed.). United States: McGraw-Hill

DeMars, C. E. (2018). Classical test theory and item response theory. P. Irwing, T. Booth ve D. J. Hughes (Ed.). *The Wiley Handbook of Psychometric Testing* (1st ed.) içinde (s. 49–73). doi: 10.1002/9781118489772.ch2

DePiero, F. (2001, October). NetExam: a Web-based assessment tool for ABET2000. In *31st Annual Frontiers in Education Conference. Impact on Engineering and Science Education. Conference Proceedings (Cat. No. 01CH37193)* (Vol. 2, pp. F3A-F13). IEEE. Erişim adresi: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=963726&casa_token=nZ056mUPUmUAAAAA:SGzZJMvDRqjL9en924iYxNgUWLX4KTiS86Zr9rKoy4meE09XRb4__TGxegbRffnlUFAwWlGUuHA

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44*(11), 50–59. Erişim adresi: https://www.jstor.org/stable/41219505

Doğan, N. (2019). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Du, Y. ve Kern, J. L. (2020). The four-parameter normal ogive model with response times. In: Wiberg, M., Molenaar, D., González, J., Böckenholt, U., Kim, JS. (eds) Quantitative Psychology. IMPS 2019. Springer Proceedings in Mathematics & Statistics, vol 322. Cham: Springer. https://doi.org/10.1007/978-3-030-43469-4

Edelen, M. O. ve Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18. doi: 10.1007/s11136-007-9198-0

Eleje, L. İ., Onah, F. E. ve Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational and Social Sciences, 3*(1), 57-75. Erişim adresi: https://dergipark.org.tr/en/pub/ejees/issue/40156/477675

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*(3), 201-212. https://doi.org/10.1177/014662169602000302

Embretson, S. E. ve Reise, S. P. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum Associates, Inc.

Engelhard, G., Wang, J. (2020). Developing a Concept Map for Rasch Measurement Theory. In: Wiberg, M., Molenaar, D., González, J., Böckenholt, U., Kim, JS. (eds) Quantitative Psychology. IMPS 2019. Springer Proceedings in Mathematics & Statistics, vol 322. Cham: Springer. https://doi.org/10.1007/978-3-030-43469-4_2

European Language Portfolio (ELP). (t.y.). Erişim adresi: https://www.coe.int/en/web/portfolio

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika, 7*(1), 19-29. Erişim adresi: https://link.springer.com/article/10.1007/BF02288601

Ferrando, P. J. ve Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT. *Psicologica: International Journal of Methodology and Experimental Psychology, 28*(2), 237-257. Erişim adresi: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://files.eric.ed.gov/fulltext/EJ802623.pdf

Finney, D. J. (1944). The application of probit analysis to the results of mental tests. *Psychometrika, 9*(1), 31-39. Erişim adresi: https://link.springer.com/article/10.1007/BF02288711

Folk, V. G. ve Smith, R. L. (2002). Models for delivery of CBTs. *Computer-based testing: Building the foundation for future assessments* içinde (s. 41-66). Mahwah/New Jersey/London: Lawrence Erlbaum Associates, Publishers

Fotaris, P. ve Mastoras, T. (2014). LMS assessment: using IRT analysis to detect defective multiple-choice test items. *International Journal of Technology Enhanced Learning, 6*(4), 281-296. Erişim adresi: https://www.academia.edu/download/38461842/IJTEL60401_Fotaris___Mastoras.pdf

Fraley, R. C., Waller, N. G. ve Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality And Social Psychology, 78*(2), 350- 365. https://doi.org/10.1037/0022-3514.78.2.350

Galton, F. (1890). Mental tests and measurements. *Mind, 15*(59), 373-381. Erişim adresi: https://www.jstor.org/stable/2247264

Gardner, E. (1989). *Five common misuses of tests* (Rapor No: 108). Washington, DC: ERIC Publications. Erişim Adresi: https://eric.ed.gov/?id=ED315429

Gierl, M., Lai, H. ve Li, J. (2011). Evaluating the performance of CATSIB in a multi-stage adaptive testing Environment. Erişim adresi: https://mcc.ca/wp-content/uploads/Technical-Reports-Gierl-Lai-Li-2011.pdf

Glas, C. A. W. ve van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*(4), 247–261. https://doi.org/10.1177/0146621603027004001

Goldstein, H. ve Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*(2), 139–167. doi:10.1111/j.2044-8317.1989.tb00905.x

Gulliksen, H. (1950). *Theory of mental tests*. New Jersey: John Wiley & Sons Inc. https://doi.org/10.1037/13240-000.

Hambelton, R. K. (1994). Item Response theory: a broad psychometric framework for measurement advances, *Psicothema, 6*(3), 535-556. Erişi adresi: https://www.redalyc.org/pdf/727/72706318.pdf

Hambleton, R. K. ve Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: Issues and Practice, 12*(3), 38-47. Erişim adresi: https://www.academia.edu/download/47596077/ITEMS_Module_16.pdf

Hambleton, R. K. ve Linden, W. J. (1997). *Handbook of modern item response theory* (1st ed.). USA: Springer. https://doi.org/10.1007/978-1-4757-2691-6

Hambleton, R. K. ve Swaminathan, H. (2013). *Item response theory: Principles and applications*. New York: Springer Science & Business Media, LLC.

Hambleton, R. K. ve Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hambleton, R. K., Swaminathan, H. ve Rogers, H. J. (1991). *Fundamentals of item response theory library* (1st ed.; D. Foster, ed.). London: SAGE.

Hambleton, R. K. ve Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education, 19*(3), 221-239. https://doi.org/10.1207/s15324818ame1903_4

Han, K. (Chris) T. (2020). Framework for developing multistage testing with intersectional routing for short-length tests. *Applied Psychological Measurement*, *44*(2), 87-102. https://doi.org/10.1177/0146621619837226

Han, K. T. (2013). MSTGen: Simulated data generator for multistage testing. *Applied Psychological Measurement, 37*, 666–668. https://doi.org/10.1177/0146621613499639

Han, K. T., Dimitrov, D. M. ve Al-Mashary, F. (2019). Developing multistage tests using d-scoring method. *Educational and Psychological Measurement, 79*(5), 988–1008. https://doi.org/10.1177/0013164419841428

Han, K. T. ve Guo, F. (2014). Multistage testing by shaping modules on the fly. D. Yan, A. A. von Davier ve C. Lewis (Ed.), *Computerized multistage testing: Theory and applications* (s. 119-133) içinde. Boca Raton/London/New York: CRC Press Taylor&Francis Group.

Han, K. T. ve Guo, F. (2013). An approach to assembling optimal multistage testing modules on the fly. *GMAC Research Reports* (Report No: RR-13-01). Erişim adresi: https://www.gmac.com/-/media/files/gmac/research/research-report-series/rr-13-01-moduleassemblyonthefly.pdf

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35-41. Erişim adresi: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.1989.tb00313.x

Harvey, R. J. ve Murry, W. D. (1994). Scoring the Myers-Briggs Type Indicator: Empirical comparison of preference score versus latent-trait methods. Journal of Personality Assessment, 62(1), 116-129. https://doi.org/10.1207/s15327752jpa6201_11

Hattie, J., Jaeger, R. M. ve Bond, L. (1999). Persistent methodological questions in educational testing. Review of Research in Education, 24, 393–446. Erişim adresi: https://journals.sagepub.com/doi/pdf/10.3102/0091732X024001393

Hays, R. D., Morales, L. S. ve Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9), 28–42. doi:10.1097/00005650-200009002-00007

Hendrickson, A. (2007). An NCME ınstructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M. ve Varni, J. W. (2007). Practical issues in the application of item response theory: A demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 Generic Core Scales. *Medical Care, 45*(5), 39–47. Erişim adresi: http://www.jstor.org/stable/40221457

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education, 33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

Hori, K., Fukuhara, H. ve Yamada, T. (2020). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. *WIREs Computational Statistics, 2020*(e1531), 1-22. doi:10.1002/wics.1531

Howel, S. L. ve Hricko, M. (2005). *Online assessment and mesaurement: Case studies from higher edication, K-12 and corporate*. USA: Information Science Publishing.

Hwang, D. Y. (2002). *Classical test theory and ıtem response theory: Analytical and empirical comparisons* (Rapor No: ED466779), Austin, TX: ERİC. Erişim adresi: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://files.eric.ed.gov/fulltext/ED466779.pdf

Jabrayilov, R., Emons, W. H. M. ve Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Jodoin, M. G., Zenisky, A. ve Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203-220. https://doi.org/10.1207/s15324818ame1903_3

Journal of Educational Measurement (1977). *14*(2), Erişim adresi: https://onlinelibrary.wiley.com/toc/17453984/1977/14/2

Jung, I. Y. ve Yeom, H. Y. (2009). Enhanced security for online exams using group cryptography. *IEEE transactions on Education, 52*(3), 340-349. Erişim adresi: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4914745

Kalkan, Ö. K. ve Çuhadar, İ. (2020). An evaluation of 4PL IRT and DINA models for estimating pseudo-guessing and slipping parameters. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 131-146. Erişim adresi: https://dergipark.org.tr/en/download/article-file/1137746

Kean, J. ve Reilly, J. (2014). Item response theory. In *Handbook for Clinical Research: Design, Statistics and Implementation* (pp. 195-198). New York: Demos Medical Publishing.

Kearney, C. P. (1983). Uses and Abuses of Assessment and Evaluation Data by Policymakers. *Educational Measurement: Issues and Practice, 2*(3), 9–12. doi:10.1111/j.1745-3992.1983.tb00702.x

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Doctoral dissertation, The University of Texas at Austin).

Kizlik, B. (2014). *Measurement, assessment, and evaluation in education*. Erişim adresi: https://www.cloud.edu/Assets/pdfs/assessment/assessment%20_%20evaluation_ measurement.pdf

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation,* Thousand Oaks, CA: Sage. Publications, Inc., https://www.doi.org/10.4135/9781483385693

Ko, C. C. ve Cheng, C. D. (2004). Secure Internet examination system based on video monitoring. *Internet Research, 14*(1), 48–61. doi:10.1108/10662240410516318

Lawley, D. N. (1943). On Problems connected with Item Selection and Test Construction. Proceedings of the Royal Society of Edinburgh. *Mathematical and Physical Sciences, 61*(3), 273–287. doi:10.1017/s0080454100006282

LeBeau, B., Assouline, S. G., Mahatmya, D. ve Lupkowski-Shoplik, A. (2020). Differentiating among high-achieving learners: a comparison of classical test theory and item response theory on above-level testing. *Gifted Child Quarterly, 64*(3), 219-237. https://doi.org/10.1177/0016986220924050

Li, D. (2020). Constant CSEM achieved through scale transformation and adaptive testing. In: Wiberg, M., Molenaar, D., González, J., Böckenholt, U., Kim, JS. (eds) Quantitative Psychology. IMPS 2019. Springer Proceedings in Mathematics & Statistics, vol 322. Cham: Springer. https://doi.org/10.1007/978-3-030-43469-4_2

Linden, W. J. ve Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In Handbook of modern item response theory (pp. 1-28). New York, NY: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Routledge.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*(3), 147-151. https://doi.org/10.1111/j.1745-3984.1971.tb00918.x

Lord, F. M. (1968). *Some test theory for tailored testing* (Rapor No: 68-38). New Jersey: Educational Testing Service (ETS). https://doi.org/10.1002/j.2333-8504.1968.tb00562.x

Lord, F. M. (1951). *A theory of test scores and their relation to the trait measured* (Rapor No: 51-13). New Jersey: Educational Testing Service (ETS). doi:10.1002/j.2333-8504.1951.tb00922.x

Lord, F. M. ve Novick, M. R. (2008). *Statistical theories of mental test scores*. USA: Information Age Publishing Inc.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2), 0-0. Erişim adresi: https://www.testpublishers.org/assets/documents/Volum%207%20Some%20useful%20cost%20benefit.pdf

Luecht, R. M. (2003). *Exposure control using adaptive multi-stage item bundles* (Rapor No: TM034856). Chicago, IL: Educational Resources Information Center (ERIC). Erişim adresi: https://files.eric.ed.gov/fulltext/ED475831.pdf

Luecht, R. M., Brunfield, T. ve Breithaupt, K. (2006) A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189-202. https://doi.org/10.1207/s15324818ame1903_2

Luecht, R. M. ve Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 239–249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Luecht, R. M. ve Sireci, S. (2012). A review of models for computer-based testing. New York: The College Board. Erişim adresi: https://files.eric.ed.gov/fulltext/ED562580.pdf

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language testing, 18*(4), 351-372. Erişim adresi: https://journals.sagepub.com/doi/pdf/10.1177/026553220101800403

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model* (Doktora Tezi). ProQuest Dissertations and Theses veri tabanından erişildi. (UMI No. 3328282). Erişim adresi: https://www.proquest.com/docview/304482829?pq-origsite=gscholar&fromopenview=true

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304-315. Erişim adresi: https://journals.sagepub.com/doi/pdf/10.1177/0146621613475471

Magis, D., Yan, D. ve von Davier A. A. (2017). *Computerized adaptive and multistage testing with R: Using Packages catR and mstR* (1st ed.). USA: Springer. doi: 10.1007/978-3-319-69218-0

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11. Erişim adresi: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1426043

Maliyet. (t.y.). *Türk Dil Kurumu (TDK) güncel Türkçe sözlük* içinde. Erişim adresi: https://sozluk.gov.tr/

Marcoulides, G. (1999). Generalizability Theory: Picking up Where the Rasch IRT Model Leaves off? S. Embretson, & S. Hershberger (pp. 129-130), In *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. United States: Lawrence Erlbaum Associates Inc.

Masters, G. N. (2014). Getting to the essence of assessment. *CARI (Center for Assessment Reform and Innovation), 1,* 1-6. Erişim adresi: https://www.acer.org/files/uploads/Assessment_Getting_to_the_essence.pdf

McBride, N. L. (2001). *An item response theory analysis of the scales from the international personality item pool and the neo personality inventory-revised* (Doctoral dissertation, Virginia Tech). Erişim adresi: https://vtechworks.lib.vt.edu/handle/10919/34430

McCabe, D. L., Trevino, L. K. ve Butterfield, K. D. (2001). Cheating in Academic Institutions: A Decade of Research. *Ethics & Behavior, 11*(3), 219–232. doi:10.1207/s15327019eb1103_2

McGough, J., Mortensen, J., Johnson, J. Fadali, S. (,2001). A web-based testing system with dynamic question generation. 31st Annual Frontiers in Education Conference. Impact on Engineering and Science Education. Conference Proceedings (Cat. No.01CH37193). doi:10.1109/fie.2001.964056

McSweeney, B. ve McSweeney, W. (1999). *Security, identity and interests: a sociology of international relations*. UK/USA/Australia: Cambridge University Press. Erişim adresi: https://books.google.com.tr/books?hl=tr&lr=&id=VQVTa-CKLjUC&oi=fnd&pg=PP13&dq=Bill+McSweeney,+Security,+Identity+and+Interests:+A+Sociology+of+International+Relations,+Cambridge:+Cambridge+University+Press,+1999,+s.+13.&ots=MvR5g5o57g&sig=0RL4k4J4QEZwqpnRSZUxvh_xR08&redir_esc=y#v=onepage&q=security&f=false

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*(3), 185-187. doi: 10.1207/s15324818ame1903_1

Miller, M. D., Linn, R. L. ve Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). New Jersey: Pearson.

Millman, J., Bishop, C. H. ve Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*(3), 707-726. https://doi.org/10.1177/001316446502500304

Mills, C. N., Potenza, M. T., Fremer, J. J. ve Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates Publishers.

Morizot, J., Ainsworth, A. T. ve Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research In R. W. Robins R. C. Fraley and R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–421). New York/London: The Guilford Press.

Murphy, K. R. ve Davidshofer, C. O. (2005). *Psychological testing: principles and applications (6th ed.)*. New Jersey: Pearson Education Australia PTY, Ltd.

Mutiawani, V., Saputra, K. ve Subianto, M. (2022). Implementing Item Response Theory (IRT) method in quiz assessment system. *TEM Journal, 11*(1), 210-218. doi: 10.18421/TEM111-26

Nering, M. L. ve Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge.

Nunnally, J. C. ve Berstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill Inc.

Obinne, A. D. E. (2012). Using IRT in determining test item prone to guessing. *World Journal of Education, 2*(1), 91-95. Erişim adresi: https://eric.ed.gov/?id=EJ1158955

Obinne, A. D. E. (2008). *Comparison of psychometric properties of West African Examinations Council and national examinations council test items under ıtem response theory* (Doctoral dissertation, UNN). Erişim adresi: http://www.dspace.unn.edu.ng/bitstream/handle/123456789/5729/Obinne%2C%20Amalonye%20Dayalata%20Ethel.pdf?sequence=2&isAllowed=y

Ockey, G. J. (2012). Item Response theory. G. Fulcher ve F. Davidson (Editörler), *The routledge handbook of language testing içinde* (s. 336-349). New York/London: Routledge.

Owen, R. J. (1969). *A bayesian analysis of Rasch's multiplicative poisson model for misreadings* (Rapor No: 69/64). New Jersey: Educational Testing Service (ETS). doi:10.1002/j.2333-8504.1969.tb00742.x

Overton, T. (2012). *Assessing learners with special needs: An applied approach* (7th Ed.). Upper Saddle River, NJ: Merrill Pearson.

Pagano, M. ve Gauvreau. K. (2018). *Principles of biostatistics* (2nd Ed.). London/New York: CRC Press.

Park, R., Kim, J., Chung, H. ve Dodd, B. G. (2017). The development of MST test information for the prediction of test performances. *Educational and Psychological Measurement, 77*(4), 545-715. https://doi.org/10.1177/0013164416662960

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Doktora Tezi). ProQuest Dissertations and Theses veri tabanından erişildi (UMI No: 9950199). Erişim adresi: https://www.proquest.com/docview/304514969?parentSessionId=5hrdMrkXHbysbOeVhRWpNrEId1eQOX%2BB4lUDkPhF6N8%3D

Petrillo, J., Cano, S. J., McLeod, L. D. ve Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value in Health, 18*(1), 25-34. https://doi.org/10.1016/j.jval.2014.10.005

Pohl, S. (2014). Longitudinal multistage testing. *Journal of Educational Measurement, 50*(4), 447-468. https://doi.org/10.1111/jedm.12028

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. *Infectious Disease Modelling,* 5(2020), 111-128. https://doi.org/10.1016/j.idm.2019.12.010

Progar, Š., Sočan, G. ve Peč, M. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology, 17*(3), 5-24. Erişim adresi: http://psiholoska-obzorja.si/arhiv_clanki/2008_3/progar_socan.pdf

Raborn, A. ve Sarı, H. (2021). Mixed Adaptive Multistage Testing: A New Approach. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 358-373. Erişim adresi: https://dergipark.org.tr/en/download/article-file/1543608

Raju, N. S., Price, L. R., Oshima, T. C. ve Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31*(3), 169-180. Erişim adresi: https://journals.sagepub.com/doi/pdf/10.1177/0146621606291569

Reese, L. M., Schnipke, D. L. ve Luebke, S. W. (1999). Incorporating content constraints into a multi-stage adaptive testlet design. (Law School Admissions Council Computerized Testing Report 97-02). Newtown, PA: Law School Admission Council.

Reise, S. P., Ainsworth, A. T. ve Haviland, M.G. (2005). Item response theory. Fundamentals: Applications, and promise in psychological research, *Current Directions in Psychological Science, 14*(2), 95-101. https://doi.org/10.1111/j.0963-7214.2005.00342.x

Reise, S. P. ve Haviland M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*(3), 228-238. https://doi.org/10.1207/s15327752jpa8403_02

Reise, S. P. ve Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93-103. Erişim adresi: https://www.tandfonline.com/doi/epdf/10.1207/S15327752JPA8102_01?needAccess=true&role=button

Reise, S. P. ve Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48. doi: 10.1146/annurev.clinpsy.032408.153553

Reise, S. P., Widaman, K. F. ve Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566. Erişim adresi: https://psycnet.apa.org/fulltext/1994-08191-001.pdf

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika, 1*, 33–49. https://doi.org/10.1007/BF02288003

Rogers, C. F. (2006). Faculty perceptions about e-cheating during online testing. *Journal of Computing Sciences in Colleges, 22*(2), 206-212. Erişim adresi: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.researchgate.net/profile/Camille-Rogers-3/publication/262311152_Faculty_perceptions_about_e-cheating_during_online_testing/links/5751b50f08ae02ac127786b8/Faculty-perceptions-about-e-cheating-during-online-testing.pdf

Rotou, O., Patsula, L., Steffen, M. ve Rizavi, S. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests* (Rapor No: 07-04). New Jersey: Educational Testing Service (ETS). https://doi.org/10.1002/j.2333-8504.2007.tb02046.x

Rupp, A. A. ve Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63–84. https://doi.org/10.1177/0013164404273942

Rusch, T., Lowry, P. B., Mair, P. ve Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management, 54*(2), 189-203. https://doi.org/10.1016/j.im.2016.06.005

Sarı, H. İ. (2020). Testing multistage testing configurations: Post-Hoc vs. hybrid simulations. *International Journal of Psychology and Educational Studies, 7*(1), 27-37. https://doi.org/10.17220/ijpes.2020.01.003

Sarı, H. İ. (2016). *Examining content control in adaptive tests: computerized adaptive testing vs. computerized multistage testing* (Doctoral Thesis). University of Florida.

Sarı, H. İ. ve Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice, 17*(5), 1759-1781. doi: 10.12738/estp.2017.5.0484

Sarı, H. İ., Yahsı-Sarı, H. ve Huggins-Manley, A. C. (2016). Computer adaptive multistage testing: practical ıssues, challenges and principles. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *7*(2), 388–388. https://doi.org/10.21031/epod.280183

Savulescu, C., Polkowski, Z. ve Alexandru, A. I. (2015, June). The online and computer aided assessment. In *2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 25-30). IEEE. Erişim adresi: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7301226&casa_token=ILtEA4MnjI8AAAAA:8zvieTwXmpEjIKtKILwO9sPerGMPYO8mhom3OsV4qCobWZHCm7pNqoIJkQb1_-CzscT39K3arZg&tag=1

Schnipke, D. L. ve Reese, L. M. (1997). *A comparison of testlet-based test designs for computerized adaptive testing*. Chicago, IL: Paper presented at the annual meeting of American Educational Research Association. Erişim adresi: https://files.eric.ed.gov/fulltext/ED409366.pdf

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

Stage, C. (1998). *A comparison between ıtem analysis based on ıtem response theory and on classical test theory: A study of the SweSAT subtest ERC*. Erişim adresi: https://www.umu.se/globalassets/organisation/fakulteter/samfak/institutionen-for-tillampad-utbildningsvetenskap/hogskoleprovet/publications/60608_enr3098sec.pdf

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. Journal of Personality and Social Psychology, 66(2), 341–349. https://doi.org/10.1037/0022-3514.66.2.341

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science, 161*(3844), 849-856. Erişim adresi: https://www.jstor.org/stable/pdf/1724851.pdf

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Rapor No: 23306516). New Jersey: Educational Testing Service (ETS). Erişim adresi: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1994.tb01578.x

Thorpe, G. L. ve Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship, 20*(2), 1-33. Erişim adresi: https://digitalcommons.library.umaine.edu/psy_facpub/20

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*(7), 433-451. Erişim adresi: https://psycnet.apa.org/fulltext/1925-16042-001.pdf

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement, 16*, 8-13. Erişim adresi: https://www.winsteps.com/a/Traub.pdf

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1–13. https://doi.org/10.1007/BF02288894

Turani, A. A., Alkhateeb, J. H. ve Alsewari, A. A. (2020, December). Students online exam proctoring: a case study using 360 degree security cameras. In *2020 emerging technology in computing, communication and electronics (ETCCE),* (pp. 1-5). IEEE. doi: 10.1109/ETCCE51779.2020.9350872

van Rijn, P. W. (2014). Reliability of multistage tests using item response theory. D. Yan, A. A. von Davier ve C. Lewis (Ed.), *Computerized multistage testing: Theory and applications* (s. 251-263) içinde. Boca Raton/London/New York: CRC Press Taylor&Francis Group.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243. https://doi.org/10.1037/a0027127

Wainer H. (1992). *Some practical considerations when converting a linearly administered test to an adaptive format* (Rapor No: 92-21). New Jersey: Educational Testing Service (ETS). Erişim adresi: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1992.tb01445.x

Wainer, H. ve Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement, 24*(3), 185-201. Erişim adresi: https://www.jstor.org/stable/pdf/1434630.pdf

Wallace, C. S. ve Bailey, J. M. (2010). Do concept inventories actually measure anything. *Astronomy Education Review, 9*(1), 010116. Erişim adresi: https://www.as.arizona.edu/cae/download/publications/Wallace_07.pdf

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Doktora Tezi). ProQuest Dissertations & Theses Global veri tabanında erişildi (Order No. 10273809). Erişim adresi: https://www.proquest.com/dissertations-theses/fair-comparison-performance-computerized-adaptive/docview/1901897901/se-2

Wang, C., Chen, P. ve Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement, 57*(1), 3-28. https://doi.org/10.1111/jedm.12241

Wang, C., Zheng, Y. ve Chang, H. H. (2014). Does standard deviation matter? Using "standard deviation" to quantify security of multistage testing. *Psychometrika, 79*(1), 154-174. Erişim adresi: https://link.springer.com/article/10.1007/s11336-013-9356-y

Warm, T. A. (1978). A primer of Item response theory. *Technical report* (Report No: 941078). Oklahoma City: USA Coast Guard Institute. Erişim adresi: https://apps.dtic.mil/sti/citations/ADA063072

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*(4), 17-27. Erişim adresi: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.1998.tb00632.x

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492. https://doi.org/10.1177/0146621682006004

Weiss, D. J. ve Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. Erişim adresi: https://www.jstor.org/stable/pdf/1434587.pdf

Weis, D. J. ve Yoes, M. E. (1990). Item response theory. R. K. Hambleton and J. N. Zaal (edited by). In *Advances in educational and psychological testing* (pp. 69-95). USA: Springer.

Wolfers, A. (1952). "National security" as an ambiguous symbol. *Political Science Quarterly, 67*(4), 481-502. https://doi.org/10.2307/2145138

Wright, B. D. (1967). *Sample-free test calibration and person measurement* (Rapor No: 017-810). Chicago: U.S. Department of Health Education & Welfare Office of Education. Erişim adresi: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://files.eric.ed.gov/fulltext/ED017810.pdf

Wolkowitz, A. A. (2008). *A comparison of classical test theory and item response theory methods for equating number -right scored to formula scored assessments* (Doktora Tezi). ProQuest Dissertations & Theses Global veri tabanında erişildi (Order No. 3297800). Erişim adresi: https://www.proquest.com/dissertations-theses/comparison-classical-test-theory-item-response/docview/304616631/se-2

Xu, L., Wang, S., Cai, Y. ve Tu, D. (2021). The automated test assembly and routing rule for multistage adaptive testing with multidimensional item response theory. *Journal of Educational Measurement, 58*(4), 538-563. https://doi.org/10.1111/jedm.12305

Yan, D., von Davier, A. A. ve Lewis, C. (2014). *Computerized multistage testing: Theory and application* (1st ed.). USA: CRC Press. doi: 10.1201/b16858

Yang, L. ve Reckase, M. D. (2020). The optimal item pool design in multistage computerized adaptive tests with the p-optimality method. *Educational and Psychological Measurement*, *80*(5), 955–974. doi: 10.1177/0013164419901292

Yetenek. (t.y.). *Türk Dil Kurumu (TDK) güncel Türkçe sözlük* içinde. Erişim adresi: https://sozluk.gov.tr/

Yu, C. H. (2013). A simple guide to the item response theory (IRT) and rasch modelling. Erişim adresi: https://www.fisica.net/enem/A-Simple-Guide-to-the-Item-Response-Theory-(IRT)-and-Rasch-Modeling.pdf

Zaman, A., Kashmiri, A. U. R., Mubarak, M. ve Ali, A. (2008, November). Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT. In EDU-COM 2008 International Conference. Sustainability in Higher Education: Directions for Change (pp. 590-599), Edith Cowan University, Perth Western Australia. Erişim adresi: https://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1051&context=ceducom

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Doctoral dissertation), University of Massachusetts Amherst.

Zenisky, A., Hambleton, R. K. ve Luecht, R. M. (2009). Multistage testing: Issues, designs, and research. In *Elements of adaptive testing* (pp. 355-372). New York: Springer.

Zhang, J., Chang, H. H. ve Yi, Q. (2012). Comparing single-pool and multiple-pool designs regarding test security in computerized testing. *Behavior Research Methods, 44,* 742–752. doi: 10.3758/s13428-011-0178-5

Zheng, Y. ve Chang, H. H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. Y. Cheng ve H. H. Chang (Ed.). *Advancing methodologies to support both summative and formative assessments* içinde (s. 21-39). Charlotte, NC: Information Age Publishing Erişim adresi: https://www.researchgate.net/publication/273411237_Multistage_testing_on-the-fly_multistage_testing_and_beyond

Zheng, Y. ve Chang, H. H. (2015). On-the-Fly Assembled Multistage Adaptive Testing. *Applied Psychological Measurement*, *39*(2), 104–118. https://doi.org/10.1177/0146621614544519