HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

AÇIK UÇLU MADDELERDE OTOMATİK PUANLAMANIN GÜVENİRLİĞİ VE TEST EŞİTLEME HATALARINA ETKİSİ

İbrahim UYSAL

Doktora Tezi

Ankara, 2019

Liderlik, araştırma, inovasyon, kaliteli eğitim ve değişim ile

Daha ileriye... En İyiye...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

AÇIK UÇLU MADDELERDE OTOMATİK PUANLAMANIN GÜVENİRLİĞİ VE TEST EŞİTLEME HATALARINA ETKİSİ

THE RELIABILITY OF AUTOMATED ESSAY SCORING AND ITS EFFECT ON TEST EQUATING ERRORS

İbrahim UYSAL

Doktora Tezi

Ankara, 2019

Kabul ve Onay

Eğitim Bilimleri Enstitüsü Müdürlüğüne,

İbrahim UYSAL'ın hazırladığı "Açık Uçlu Maddelerde Otomatik Puanlamanın Güvenirliği ve Test Eşitleme Hatalarına Etkisi" başlıklı bu çalışma jürimiz tarafından **Eğitim Bilimleri Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalında Doktora Tezi** olarak kabul edilmiştir.

Jüri Başkanı	Prof. Dr., Selahattin GELBAL	İmza
Jüri Üyesi	Prof. Dr., Hülya KELECİOĞLU	İmza
Jüri Üyesi (Danışman)	Prof. Dr., Nuri DOĞAN	İmza
Jüri Üyesi	Doç. Dr., Sevilay KİLMEN	İmza
Jüri Üyesi	Doç. Dr., Celal Deha DOĞAN	İmza

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim, Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından 07 / 02 / 2019 tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca ..... / ..... / ..... tarihinde kabul edilmiştir.

Prof. Dr. Ali Ekber ŞAHİN
Eğitim Bilimleri Enstitüsü Müdürü

Öz

Testlerde kullanılan yapılandırılmış cevap maddelerini puanlamak güç, zaman alıcı ve masraflı olabilmektedir. Bilgisayar teknolojisindeki gelişmeler yapılandırılmış cevap maddelerinin otomatik olarak puanlanmasını olanaklı hale getirmiştir. Ancak otomatik puanlamaya geçerlik, güvenirlik ve test eşitleme ile ilgili araştırmalar yapılmadan başvurulması ciddi problemlere yol açabilecektir. Çalışmanın amacı karma testlerde yapılandırılmış cevap maddelerini otomatik puanlamak ve bu durumun güvenirlik ve test eşitleme üzerindeki etkisini araştırmaktır. Araştırmada kullanılan veriler, Türkiye Cumhuriyeti Millî Eğitim Bakanlığı tarafından 2016 yılında uygulanan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) araştırmasına ait 8. sınıf Türkçe testi verileridir. Bahsedilen testler ortak maddeler içermektedir. Otomatik puanlama yöntemi olarak destek vektör makinesi (support vector machine [SVM]), lojistik regresyon (logistic regression [LR]), çok terimli sade bayes (multinominal naive bayes [MNB]), kısa uzun süreli bellek (long-short term memory [LSTM]) ve iki yönlü kısa uzun süreli bellek (bidirectional long-short term memory [BLSTM]) seçilmiştir. Test eşitleme sürecinde Klasik Test Kuramı (KTK) ve Madde Tepki Kuramına (MTK) dayalı yöntemlerden yararlanılmıştır. Araştırma sonucunda gerçek puanlayıcılarla en iyi uyumu gösteren otomatik puanlama yönteminin BLSTM olduğu sonucuna ulaşılmıştır. BLSTM yöntemiyle elde edilen puanlar gerçek puanlayıcıların üzerinde anlaştıkları puanlarla iyi bir uyum göstermektedir. Çoğu eşitleme yönteminde, otomatik puanlama ile gerçekleştirilen eşitleme işlemine ait hataların gerçek puanlayıcılar ile gerçekleştirilen eşitleme işlemine ait hatalara yakın olduğu gözlenmiştir. Hem gerçek puanlayıcılarla uyum hem de eşitleme işlemi açısından otomatik puanlamanın uygulanabileceği sonucuna ulaşılmıştır.

Anahtar sözcükler: Otomatik puanlama, test eşitleme, karma testler, yapılandırılmış cevap maddeleri.

Abstract

It might be difficult, time-consuming and costly to score constructed response items in tests. However, improvements in computer technology have enabled the automated scoring of constructed response items. Yet, the application of automated scoring without making investigation on validity, reliability and test equating can lead to serious problems. In this sense, the aim of this study was to score the constructed response items in mixed format tests automatically and to investigate the effect of this on test equating and reliability. The data examined in this study were the 8th grade Turkish test data of ABİDE research (Education Skills Monitoring and Evaluation) carried out by Ministry of National Education in Turkey in 2016. These tests contained common items. Support vector machine (SVM), logistic regression (LR), multinominal naive bayes (MNB), long-short term memory (LSTM) and bidirectional long-short term memory (BLSTM) were selected as automated scoring methods. During the test equating process, methods based on Classical Test Theory and Item Response Theory were utilized. The results of the study revealed that the most compatible automated scoring method with actual raters is BLSTM. The scores obtained by the BLSTM method were in good agreement with the scores of actual raters. In most of the equating methods, it was observed that errors of equating process done with automated scoring were close to the errors of equating process done by actual raters. It was concluded that automated scoring can be applied since it is compatible with actual raters and convenient in terms of equating.

Keywords: Automated scoring, test equating, mixed format tests, constructed response items.

Teşekkür

Tez çalışmamın başlangıcından sonuna kadar desteklerini esirgemeyen ve bu tezin oluşmasında büyük emekleri olan hem davranışları hem de akademik bilgisi ile bana rehber değerli danışmanım Prof. Dr. Nuri DOĞAN'a,

Doktora eğitimim süresince gelişmemde büyük katkılara sahip olan Prof. Dr. Selahattin GELBAL, Prof. Dr. Hülya KELECİOĞLU ve Doç. Dr. Burcu ATAR'a,

Tez savunma jürimde yer alarak çalışmamı gözden geçiren ve öneriler veren Doç. Dr. Celal Deha DOĞAN'a, tez izleme komitemde yer alarak tezimin tüm aşamalarını gözden geçiren ve öneriler sunan kendisinden çok şey öğrendiğim yüksek lisans tez danışmanım Doç. Dr. Sevilay KİLMEN'e,

Otomatik puanlama yazılımını oluştururken birlikte çalıştığımız Behzad Naderalvojoud'a, hem arkadaşlığı hem de fikirleriyle bu süreçte yanımda olan Arş. Gör. Abdullah Faruk KILIÇ'a,

Tezimde kullandığım veriyi sağlayan Milli Eğitim Bakanlığı'na, bu süreçte yardımlarını esirgemeyen Daire Başkanı Kemal BÜLBÜL ve Milli Eğitim Uzmanı Barış ÖZGÜRLÜK'e,

Görev yaptığım Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi'ndeki çalışma arkadaşlarıma ve bölüm başkanım Prof. Dr. Zekeriya NARTGÜN nezdinde tüm hocalarıma,

Bu süreçte bana hep anlayış gösteren tezimde benim kadar emeği bulunan sevgili eşime, minik oğluma, canım anneme, babama ve ablama, arkadaşım Önder Kamil TÜLEK'e,

Doktora eğitimim boyunca bana yurt içi doktora burs imkânı sağlayan TÜBİTAK'a,

Eğitim hayatım boyunca bugünlere gelmeme katkı sağlamış adını sayamadığım tüm arkadaşlarıma ve hocalarıma teşekkürü bir borç bilirim.

İçindekiler

Tablolar Dizini

Şekiller Dizini

Simgeler ve Kısaltmalar Dizini

AC1: Agreement Coefficient 1 (Uyum Katsayısı 1)

AC2: Agreement Coefficient 2 (Uyum Katsayısı 2)

AERA: American Educational Research Association (Amerika Eğitim Araştırmaları Birliği)

APA: American Pschological Association (Amerika Psikoloji Birliği)

BI: Bias Index (Yanlılık İndeksi)

BLSTM: Bidirectional Long-Short Term Memory (İki Yönlü Kısa-Uzun Süreli Bellek)

EC: Chained Equipercentile Equating (Zincir Eşit Yüzdelikli Eşitleme)

EF: Frequency Estimation Equipercentile Equating (Frekans Eşit Yüzdelikli Eşitleme)

EF (WS=1): Frequency Estimation Equipercentile Equating with Syntetic Population Ratio 1 (Sentetik Evren Oranının 1 Olarak Belirlendiği Frekans Eşit Yüzdelikli Eşitleme)

E-Rater: Electronic Essay Rater (Elektronik Açık Uçlu Puanlayıcı)

FORM A_1: Birinci Örnekleme Uygulanan Yeni Form

FORM B_1: İkinci Örnekleme Uygulanan Temel Form

GRUP 1: A_1 Formuna Cevap Veren Örneklem

GRUP 2: B_1 Formuna Cevap Veren Örneklem

HB: Haebara

IEA: Intelligent Essay Assessor (Yetenekli Açık Uçlu Değerlendirici)

KTK: Klasik Test Kuramı

LC: Chained Linear Equating (Zincir Doğrusal Eşitleme)

LT: Tucker Linear Equating (Tucker Doğrusal Eşitleme)

LT (WS=1): Tucker Linear Equating with Syntetic Population Ratio 1 (Sentetik Evren Oranının 1 Olarak Belirlendiği Tucker Doğrusal Eşitleme)

LR: Logistic Regression (Lojistik Regresyon)

LSTM: Long-Short Term Memory (Kısa-Uzun Süreli Bellek)

MEB: Milli Eğitim Bakanlığı

MM: Mean-Mean (Ortalama-Ortalama)

MNB: Multinominal Naive Bayes (Çok Terimli Sade Bayes)

MS: Mean-Sigma (Ortalama-Standart Sapma)

MTK: Madde Tepki Kuramı

NCME: National Council on Measurement in Education (Ulusal Eğitimde Ölçme Kurulu)

OCR: Optical Character Recognition (Optik Karakter Tanıma)

PEG: Project Essay Grader (Deneme Projesi Değerlendirici)

PI: Prevalence Index (Yaygınlık İndeksi)

RMSE: Root Mean Squared Error (Hata Kareleri Ortalamalasının Karekökü)

PSMEC: Chained Equipercentile Equating with Smoothing (Düzgünleştime ile Zincir Eşit Yüzdelikli Eşitleme)

PSMEF: Frequency Estimation Equipercentile Equating with Smoothing (Düzgünleştirme ile Frekans Eşit Yüzdelikli Eşitleme)

PSMEF (WS=1): Frequency Estimation Equipercentile Equating with Smoothing and Syntetic Population Ratio 1 (Düzgünleştirme Yapılan ve Sentetik Evren Oranının 1 Olarak Belirlendiği Frekans Eşit Yüzdelikli Eşitleme)

SEE: Standard Error of Equating (Eşitlemenin Standart Hatası)

SL: Stocking Lord

SVM: Support Vector Machine (Destek Vektör Makineleri)

QWK: Quadratic Weighted Kappa (Karesel Ağırlıklı Kappa)

UY: Uyum Yüzdesi

Bölüm 1
Giriş

Testlerle ilgili tüm bireyler olumlu ve olumsuz bazı görüşlere sahiptir. Bunun sebebi tüm bireylerin öğrenim yaşamları boyunca çok sayıda testle karşılaşmalarıdır. Bazı bireyler testleri eğitim sürecinde gerekli ve hayati olarak yorumlayarak testlerin öğrenmenin gerçekleştiğine dair bir kanıt, başarıya ilişkin bir geribildirim ve öğrencileri çalışmaya teşvik eden bir motivasyon aracı olduğunu düşünür. Bazı bireyler ise testlerin öğrencilerde sıkılganlık yarattığına, testlerin ölçmeyi iddia ettiği konuyu ölçemediğine, öğrenmeyi doğru şekilde yansıtamadığına inanır. Alanyazında testlerin öğrencilerin kaygılarını arttırabileceğine, öğrencilerin sınıflandırılmasının olumsuz etkiler yaratabileceğine, öğrencilerin öz benliklerine zarar verebileceğine ve kendini gerçekleştiren kehanet yaratabileceğine yönelik düşünceler bulunmaktadır. Aslında testler öğrenciler, programlar ve öğretim yöntemlerine yönelik değerlendirmeler yapılmasını sağlayan araçlardır (Kubiszyn ve Borich, 2013; Miller, Linn ve Gronlund, 2009). Testler bireyleri bilgi, beceri, nitelik ve yetenek temelinde ayırt etmeyi amaçlamaktadır (Geisinger ve Usher-Tate, 2016). Bireylerin ayırt edilmesindeki neden ise bireyler hakkında bazı kararlar alınmasının gerekliliğidir. Bu kararlar öğretimsel olabileceği gibi not verme, seçme, yerleştirme, psikolojik danışmanlık ve rehberliğe yönelik de olabilmektedir. Bu süreçte testler bireyler hakkında öznel öğretmen görüşlerini nesnelleştirmeyi sağlayarak daha doğru ve daha savunulabilir kararlar alınmasına vesile olur (Kubiszyn ve Borich, 2013). Testler sabit bir zaman diliminde tüm öğrencilere benzer koşullarda uygulanan madde setlerini içermektedir (Miller, Linn ve Gronlund, 2009).

Testlerde yer alan madde setleri; cevabın birey tarafından oluşturulması, puanlamadaki tutarlık, güvenirlik ve nesnellik temel alınarak çeşitli sınıflara ayrılmaktadır. Bu doğrultuda öznele karşı nesnel, seçmeye karşı üretme ve sabit cevaba karşı serbest cevap şeklinde sınıflandırmalar oluşturulmuştur. Bu sınıflandırmalardan alanyazında en çok karşılaşılanı öznele karşı nesneldir. Fakat puanlama açısından sınıflandırma yapan bu yaklaşım kısa cevaplı maddeleri tam olarak öznel ya da nesnel olarak sınıflandıramamaktadır. Bunun nedeni ise bazı kısa cevaplı maddelerin objektif olarak puanlanabilmesidir. Dolayısıyla bu sınıflandırmaya paralel olan seçmeye karşı üretme sınıflandırmasının kullanılması daha uygun görülebilir (Kubiszyn ve Borich, 2013; Rodriguez, 2002). Seçmeye karşı

üretme sınıflandırması incelendiğinde seçme maddelerinde bireylerin doğru cevabı alternatif cevaplar arasından seçmesi söz konusu iken üretme maddelerinde ise bireylerin doğru cevapları bir kelime, cümle ya da daha uzun bir metin şeklinde ifade etmesi yani yapılandırması söz konusudur (Osterlind, 2002). Seçme maddelerinde en sık karşılaşılan format doğru-yanlış, çoktan seçmeli ve eşleştirmeli iken üretme maddelerinde ise yanıtı sınırlandırılmış (restricted response) ve sınırlandırılmamış (unrestiricted, extended response) açık uçlu maddelerdir (Crocker ve Algina, 2008; Mehrens ve Lehmann, 1991; Salkind, 2006). Seçme maddelerinden doğru-yanlış madde türünde, katılımcılar bir durum ya da soruda doğru-yanlış, evet-hayır gibi iki seçenekli cevaplardan birisini seçerken çoktan seçmeli maddelerde katılımcılar bir problem ya da soruda bir doğru birden fazla yanlış cevapla (çeldirici) karşılaşmakta ve katılımcılardan doğru cevabı seçmesi beklenmektedir. Diğer bir seçme madde türü olan eşleştirme maddelerinde ise katılımcılar birisi uyarıcı diğeri cevaplardan oluşan iki listeyle karşılaşmakta bu listelerdeki nesneler arasında var olan ilişkileri belirleyerek seçim yapmaktadır (Crocker ve Algina, 2008). Yapılandırılmış cevap maddelerinden yanıtı sınırlandırılmış açık uçlu maddeler öğrencilerin sorulara birkaç kelime, birkaç cümle ya da birkaç paragrafla cevap verdiği madde formatıyken yanıtı sınırlandırılmamış açık uçlu maddeler öğrencilerin sorulara birkaç sayfada cevap verdiği madde türüdür (Downing, 2009).

Testlerde yer alacak maddelerin formatının seçiminde test geliştiriciler çoğunlukla ikileme düşmektedir. Bunun nedenleri arasında bilişsel özelliklerin ölçümüne uygunluk, uygulama ve puanlamadaki masraf, testlerde kullanılan madde türünün öğretime etki etmesi, psikometrik özellikler bulunmaktadır. Pratiklik düşünülerek testler sadece çoktan seçmeli maddeleri içerecek şekilde ya da sadece yapılandırılmış cevap maddelerini, hem çoktan seçmeli hem de yapılandırılmış cevap maddelerini içerecek şekilde tasarlanabilmektedir (Rodriguez, 2002; Martinez, 1999). Tablo 1 çoktan seçmeli ve yapılandırılmış cevap maddelerinin avantajlı ve dezavantajlı yönlerini göstermektedir. Tablo 1 incelenerek çoktan seçmeli maddeler ve yapılandırılmış cevap maddeleriyle ilgili detaylı bilgi edinilebilir. Martinez (1999) testlerde kullanılacak tek bir formatın tüm amaçlar ve durumlar için uygun olmadığını, Messick (1993) ise farklı madde formatlarının birlikte kullanılmasının her bir formatın güçlü yönlerinden faydalanmayı, zayıf yönlerini ise telafi etmeyi sağlayacağını belirtmektedir. Dolayısıyla özellikle geniş ölçekli

testlerde çoktan seçmeli ve yapılandırılmış cevap maddelerinin birlikte kullanılması yani karma testler hazırlanması oldukça önemlidir.

Tablo 1

Çoktan Seçmeli ve Yapılandırılmış Cevap Maddelerinin Avantaj ve Dezavantajları

Madde Formatı	Avantajları	Dezavantajları
Çoktan seçmeli	Objektif puanlanması	Şansla cevap verme olasılığı bulunması
	İçeriğin geniş bir şekilde temsil edilmesi	Maddelerin hatırlanabilmesi
	Puanlama maliyetinin düşük olması	Madde hazırlamanın zor olması
	Uygulamanın kolay olması	Yoruma kapalı olması
	Güvenirliğin yüksek olması	Özgün çözümleri ya da fikirleri sınırlandırması
	Puanlayıcı etkisinin bulunmaması	Okuma becerilerinin sonuca etkide bulunabilmesi
Yapılandırılmış Cevap	Kısmi puanlama imkanı sunması	Subjektif puanlanması
	Madde yazımının çoktan seçmeli maddelere göre kolay olması	İçeriğin sınırlı bir şekilde temsil edilmesi
	Bireyi derinlemesine değerlendirme imkanı sunması	Puanlama maliyetinin yüksek olması ve puanlama zorluğu bulunması
	Öğrencilerin kendi cevaplarını kendi kelimeleriyle sunması için özgürlük tanıması	Test süresinin fazla olması
	Beceri ve yetenek ölçümüne daha elverişli olması	Puanlayıcı etkisinin geçerliği tehdit etmesi
	İpucu barındırmaması	Yazma becerilerinin sonuca etkide bulunabilmesi

Not: Downing (2009), Ebel ve Frisbie (1991), Haladyna ve Rodriguez (2013), Rodriguez (2002) kaynaklarından derlenmiştir.

Tablo 1'de çoktan seçmeli ve yapılandırılmış cevap maddelerinin avantaj ve dezavantajları incelendiğinde yapılandırılmış cevap maddelerinin bireyin derinlemesine değerlendirilmesine imkân sağladığı görülmektedir. Ancak geniş ölçekli test uygulamalarında yapılandırılmış cevap maddelerinin puanlanması olukça güç, zaman alıcı veya masraflıdır. Bu durum nedeniyle test uygulayıcıları bir arayışa girmiş ve otomatik puanlama kavramını ortaya çıkarmışlardır. Otomatik madde puanlama, bilgisayar destekli analizlerle yazılı bir metni değerlendirmektir

(Shermis, 2010). Otomatik madde puanlama fikri puanlama zorluğunu azaltmak üzere yaklaşık 50 yıl önce bir ortaokul öğretmeni olan Page (1966) tarafından ortaya atılmıştır (Ramineni ve Williamson, 2013). Page (1966) Project Essay Grade (PEG) programının yaratıcısı olup geliştirilen bu ilk programda kompozisyonlar puanlanırken kelime uzunluğu, kompozisyon uzunluğu, virgül ve edat sayısı, çoğunlukla kullanılmayan maddeler üzerinden puanları tahmin etme yoluna gidilmiştir (Wang ve Brown, 2007). Şu an ise okullarda yer alan yazma becerisi görevlerinin %90'ının otomatik madde puanlama sistemleri ile değerlendirilebileceği belirtilmektedir (Shermis ve Burnstein, 2003). Sınıf içi uygulamaların yanı sıra otomatik puanlama sistemleri ile geniş ölçekli testlerde de puanlama yapılabilmektedir. Bu testlere GMAT, TOEFL, GRE örnek olarak verilebilir. Otomatik puanlama işleminde özellikler herhangi bir ön puanlama olmadan bilgisayara el ile tanımlanacağı gibi ön puanlamadan yararlanarak puanlama davranışları bilgisayara otomatik olarak haritalandırılabilir. Bahsedilen ilk yöntem alanyazında denetlenmeyen (unsupervised), ikinci yöntem ise denetlenen (supervised) makine öğrenme yöntemleri olarak bilinmektedir. Otomatik puanlama sistemleri kısa cevaplı maddelerden başlayarak kompozisyonlara kadar uygun bir yelpazede işlem yapabilmektedir. Otomatik puanlama sistemlerinin en önemli avantajı bireylere hemen dönüt verebilmesidir (Gierl, Latifi, Lai, Boulais ve Champlain, 2014). Otomatik puanlama sistemleri için alanyazında geliştirilen birçok yazılım bulunmaktadır. Bunlardan sıklıkla kullanılanları Electronic Essay Rater (E-Rater), Intelligent Essay Assessor (IEA), Project Essay Grade (PEG) yazılımlarıdır. Bu programların yanı sıra çevrimiçi otomatik puanlama yapan sistemler de mevcuttur. Buna örnek olarak da MyAccess programı verilebilir (Tsai, 2012).

Birey geniş ölçekli bir test uygulamasına katıldığında ve testten düşük bir puan aldığında, test zor mu yoksa bireyin yeteneği düşük mü karar vermek zordur. Zor bir test formundan düşük puan alan bireylerin yeteneği her zaman düşük olarak yorumlanmamalıdır. Geniş ölçekli testlerde kopya, soruların ifşa olması, aynı maddelerle teste iki kez katılma gibi nedenlerle test geliştiriciler çoğu zaman aynı beceriyi ölçen farklı test formları oluşturmak zorunda kalmaktadır. Bahsedilen test formları her ne kadar birbirine paralel (içerik ve istatistiksel açıdan) olarak hazırlanmaya çalışılsa da testlerin güçlüklerinin bir formdan diğerine farklılaşması kaçınılmaz olmaktadır. Dolayısıyla belirli bir testten düşük puan alan bireylerin farklı

bir test formundan aynı düzeyde düşük puan alan bireylerle denkliği sorgulanmalıdır. Farklı test formlarından elde edilen puanların bireyler, kurumlar ve toplumlar için önemli kararlar alınırken kullanılması psikometristlerin farklı test formlarından bireylerin aldıkları puanları adil ve doğru karşılaştırmasını gerektirmektedir. Nitekim yukarıda bahsedilen durumda bireylerden birisi kolay bir test formundan düşük puan almış olabilecekken diğer birey orta güçlükte bir test formundan düşük puan almış olabilir. Eğer kolay test formundan düşük puan alan birey, orta güçlükteki test formundan düşük puan alan bireyi puan sıralamasında geçmişse bahsedilen iki birey hakkında yanlış karar verilebilecektir. Böyle bir durumda daha yetenekli birey işe giremeyebilecektir. Bahsedilen kararlar bireyler için akademik kabul, burs hibesi, akademik ilerleme, görev yetkinliği; kurumlar için bireylere bir meslek için belge verilmesi, orduya, üniversiteye ya da ortaöğretime öğrenci seçilmesi; toplum için eğitimin ilerleyişinin belirlenmesi ve eğitim uygulamalarının değerlendirilmesi olabilir. Yukarıda bahsedilen türde adaletsizliklerin önüne geçmek için bir test formundan alınan puanları diğer test formuna dönüştürmek yani test eşitleme süreçlerini işe koşmak gerekmektedir (Gonzalez ve Wiberg, 2017; Hambleton ve Swaminathan, 1985; Kolen ve Brennan, 2014; von Davier, 2011; Wu, Tam ve Jen, 2016).

Problem Durumu

Yapılandırılmış cevap maddeleri uygulanması zor olan puanlaması uzun zaman ve çaba gerektiren bir yapıya sahiptir (Gierl, vd., 2014). Puanlanacak birey ve yapılandırılmış cevap maddelerinin miktarı arttıkça da daha fazla puanlayıcıya ihtiyaç duyulmaktadır. Bunun yanı sıra çok sayıda puanlayıcının puanlama konusunda eğitilmesi gerekmektedir. Puanlayıcıların duygularının ve bilişsel yetilerinin puanlamada yanlılığa neden olması ise bir başka sorundur (Adesiji, Agbonifo, Adesuyi ve Olabode, 2016). Puanlamadaki öznellik güvenirliği de düşürmektedir (Ebel ve Frisbie, 1991; Hagge, 2010). Geniş kapsamlı test uygulamaları düşünüldüğünde yapılandırılmış cevap maddelerinin puanlanmasının maliyeti de oldukça arttıracağı göz önünde tutulmalıdır (Cohen, Ben-Simon ve Hovav, 2003). Şüphesiz ki otomatik madde puanlama sistemlerinin kullanılması kaynakların verimli kullanılmasını sağlayacak, puanlama süresini azaltacak, enerji kaybını önleyecektir (Attali ve Burstein, 2006; Chen, Xu ve He, 2014). Bu sistemin kullanımı çok sayıda puanlayıcı kullanma zorunluluğunu ortadan kaldıracaktır.

Benzer şekilde puanlama yanlılığının da önüne geçilebilecektir. Farklı eğitim alan puanlayıcılardan kaynaklanan, güvenirlikle ilgili sorunlar giderilebilecek ve genellenebilirlik konusunun da üstesinden gelinebilecektir (Adesiji, Agbonifo, Adesuyi ve Olabode, 2016). Fakat otomatik puanlama sistemlerinin bireyler farklı test formları aldığında ya da teste farklı dönemlerde katıldığında aralarındaki adaletin sağlanmasında önemli olan test eşitleme gibi uygulamalardaki etkinliği alanyazında yeterince araştırılmamıştır. Bu tür araştırmalar yapılmadan otomatik puanlamaya başvurmak ciddi problemlere neden olabilecektir. Bunun yanı sıra otomatik puanlama sonuçlarının gerçek puanlayıcılara yakınlığı önemlidir. Gerçek puanlayıcılarla uyum sağlamayan otomatik puanlama sonuçları bireyler hakkında yanlış kararlar verilmesine neden olabilir. Otomatik puanlama koşulları değiştiğinde otomatik puanlama ve gerçek puanlayıcılar arasındaki uyumda ve eşitleme hatalarında değişikliklerin olması muhtemeldir. Bu doğrultuda otomatik puanlama ve test eşitlemenin yapılabileceği kabul edilebilir sınırların belirlenmesi gerekmektedir. Araştırma bu problem durumlarından yola çıkılarak tasarlanmıştır.

Araştırmanın Amacı ve Önemi

Araştırmanın üç amacı bulunmaktadır. Bunlar; i) otomatik puanlama sistemleri aracılığıyla elde edilen puanların gerçek puanlayıcılardan elde edilen nihai puanlarla olan uyumunu incelemek, ii) otomatik puanlama sistemlerince puanlanan yapılandırılmış cevap maddelerinin test eşitleme sürecinde yer almasının eşitleme hataları üzerindeki etkisini değerlendirmek, iii) otomatik puanlama sistemlerindeki koşulların değişiminde eşitleme hatalarının ve uyum katsayılarının değişimini incelemektir.

Otomatik puanlama sistemlerinin kullanılabilmesi elde edilen puanların mümkün olduğunca gerçek puanlayıcılara benzemesine bağlıdır. Diğer yandan otomatik puanlamanın test eşitleme hatasını arttırmaması, puanların geçerlik ve güvenirliğini düşürmemesi beklenmektedir. Otomatik puanlamanın güvenirliğinin ve otomatik puanlama sonucu elde edilen eşitleme hatalarının gerçek puanlayıcılarla elde edilen eşitleme hatalarından farklı olup olmadığının belirlenmesi yönüyle bu araştırmanın önemli olduğu düşünülmektedir. Böylece uygun koşullarda otomatik puanlama ve otomatik puanlama sonrası test eşitleme çalışmaları gerçekleştirilebilecektir.

Geniş ölçekli testlerde açık uçlu maddelere daha fazla yer verilmesi ve açık uçlu maddelerin daha kolay puanlanmasına yönelik öneri getirmesi nedeniyle bu araştırmanın önemli olduğu düşünülmektedir. Karma formattaki testlerde özellikle yanıtı sınırlandırılmış açık uçlu maddelerin otomatik puanlanması ve bu puanlar kullanılarak Madde Tepki Kuramı (MTK) test eşitleme yöntemleriyle gerçekleştirilen bir test eşitleme çalışmasına incelenen alanyazında rastlanmamıştır. Ayrıca çok sayıda yapılandırılmış cevap maddesi içeren testlerin otomatik puanlamayla eşitlenmesi ile ilgili bir araştırmaya incelenen alanyazında rastlanmamıştır. Araştırma sonuçlarının bu yönüyle alanyazındaki boşluğu gidermesi amaçlanmaktadır.

Araştırma Problemi

Karma formattaki testlerde otomatik puanlamanın güvenirliği nedir? Otomatik puanlamayla yapılan eşitlemenin test eşitleme hataları üzerindeki etkisi nedir?

Alt problemler. Alt problemler aşağıda sırasıyla belirtilmektedir.

1. Gerçek puanlayıcılar ve otomatik puanlama koşulları aracılığıyla elde edilen puanlar üzerinden hesaplanan uyum katsayıları nasıldır?

2. Gerçek puanlayıcılar ve otomatik puanlama koşulları aracılığı ile elde edilen puanların eşitleme hataları nasıldır ve anlamlı farklılık göstermekte midir?

Sınırlılıklar

Araştırmanın sınırlılıkları aşağıda belirtilmektedir.

1. Araştırma otomatik puanlama açısından araştırmacının da içerisinde yer aldığı bir çalışma kapsamında oluşturulan ve Türk diline özgü olarak geliştirilen yazılım ile sınırlıdır.
2. Araştırma kapsamında kullanılan verilerin bilgisayar ortamına aktarımında optik karakter tanıma (Optical Character Recognition [OCR]) sistemleri (Abbyy Finereader 14, Free OCR, Omnipage 18, PDFelement 6, Rediris 17) kullanılmaya çalışılmış fakat programların el yazısını tanımlama performansları yeterli bulunmadığından öğrenci verileri bilgisayar ortamına elle girilmiş ve kontrol edilmiştir.

Tanımlar

Karma Format Test: Yapılandırılmış cevap maddelerini ve çoktan seçmeli maddeleri bir arada içeren testler.

Otomatik Puanlama Güvenirliği: Otomatik puanlama sonucu elde edilen puanlar ile gerçek puanlayıcıların üzerinde anlaştıkları puanlar arasında hesaplanan uyum katsayılarına dayalı olarak belirlenmektedir.

Puanlayıcı: Yapılandırılmış cevap maddelerine ait cevapları derecelendiren kişiler ya da gruplar.

Yapılandırılmış Cevap Maddeleri: Katılımcıların maddelere bir kelime, bir cümle ya da birkaç cümle ile cevap verdiği madde türü.

Bölüm 2

Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Otomatik Puanlama

Otomatik puanlama işlemi denetlenen ve denetlenmeyen makine öğrenmesine dayalı olarak gerçekleştirilebilmektedir. Denetlenen makine öğrenme yöntemleri işe koşulurken genellikle dört adımdan oluşan bir süreçten yararlanılmaktadır (Powers, 2015). Bu adımlar; 1) bilgisayarı eğitmek üzere nitelikli olduğu bilinen bir puanlamanın metine dayalı bir kitaplık ile tanımlanması, 2) Eğitim verilerindeki yazılardan çeşitli özelliklerin çıkarılması, 3) yazının tüm nitelikleriyle ilgili bir model geliştirilmesi, 4) değerlendirilmemiş yazılara kurulan modeli kullanarak puan ataması yapılması ya da kategorilere ayrılmasıdır. Aşağıda bu aşamalar Berg ve Gopinathan (2017), Gierl, vd. (2014), Jang, Kang, Noh, Kim, Sung ve Seong (2014), Lilja (2018)'ya ait kaynaklardan yararlanarak detaylı olarak açıklanmaktadır.

Aşama 1: Veri hazırlığı. Veri hazırlanırken öncelikle verinin elektronik formda olması gerekmektedir. Eğer bilgisayara dayalı bir test kullanılıyorsa veriler doğrudan ve hızlı bir şekilde elde edilebilir. Fakat eğer kağıt-kalem testleri kullanılıyorsa verilerin bilgisayar ortamına elle aktarılması gerekmektedir. Bu işlem için optik karakter tanıma sistemleri kullanılabilse de bu sistemlerin de belirli bir hata ortaya çıkaracağı unutulmamalıdır. Bu işlemler tamamlandıktan sonra gerçek puanlayıcı puanları için otomatik puanlama sistemi tarafından okunabilecek bir tanımlama yapılır. Veri temizlenerek (örneğin kelime aralarındaki boşluğun düzeltilmesi, yazım hatalarının düzeltilmesi, gereksiz sembollerin silinmesi gibi) ikinci aşamaya geçilmektedir.

Aşama 2: Özellik çıkarma. Makine öğrenmesinde kullanılacak ve yazıyı temsil eden özellikler girdi verilerinden belirlenir. Page (1966) özellik çıkarmaya ilişkin olarak iki kavramdan bahsetmektedir. Bu kavramlardan ilki olan "trin" gerçek puanlayıcıların yazıyı puanlarken kullandığı asıl özellikleri anlatmaktadır. "Trin"; kelime seçim kabiliyeti, akıcılık, söz dizimi, cümle yapısı, noktalama, dil bilgisi, stil, organizasyon, sözcüksel karmaşıklık ve paragraf gelişimi olabilir. İkinci kavram olan "prox" ise bilgisayarın "trin"den yararlanarak çıkardığı yaklaşık değişkenlerdir. "Prox"; kelimelerin toplam sayısı, cümle uzunluğu ortalaması, ortalama kelime

uzunluğu ve yazım hatası olabilir. Günümüzde doğal dil işlemeye dayalı daha karmaşık yöntemler kullanılmaktadır. Bu yöntemler kelime sınıflarını, anahtar ifadeleri, sözdizimsel yapıları tanımlamak için kullanılmaktadır. Bunlara örnek olarak n-gramlar (n-grams), söylem bölümlerinin etiketlenmesi (part-of-speech tagging), terim sıklığı-ters doküman sıklığı (term frequency-inverse document frequency) ve kelime paketi (bag of words) yöntemleri gösterilebilir. Bu araştırmaya konu olan n-gramlara ve terim sıklığı-ters doküman sıklığı ile ilgili bilgiye aşağıda yer verilmektedir.

n gramlar (n grams). **"**n gramlar" n uzunluğunda ardışık bir dizi kelimeyi göstermektedir. n=1 olduğunda n gramlar tek bir terim içeren tekgram (unigram), n=2 olduğunda iki gram (bigram), n=3 olduğunda üç gram (trigram) olarak adlandırılmaktadır. Bu yöntemde temel ilke yazının n tane aynı zamanda ortaya çıkan elementten oluşmasıdır. n gramlar hecelerin kontrolünde, yazı özetlemede, kelime bölmede kullanılabilir. Bu yöntemde sonuç vektörü bir cümlede bir kelimenin kaç kez görüldüğünü göstermektedir. Cümlelerin tamamında bulunan her bir kelimeye bir etiket atanarak çalışmaktadır.

Terim sıklığı-ters doküman sıklığı. Bu metrik bir yazıdaki terimlerin yaygınlığı ve önemini ölçmek için kullanılmaktadır. Sınıflandırmada kullanılan bu yöntem terimlere ağırlık verilmesiyle çalışmaktadır. Bu yöntemde terimler n-gramlar ya da kelimeler olabilmektedir. Bir terim yazıda birçok kez bulunmuşsa bu terime yüksek bir değer atanmaktadır. Bir terim, yazıda az sayıda ya da tüm belgelerde çok sayıda bulunmuşsa bu terime düşük bir değer atanmaktadır. Çünkü bu terimler yazı hakkında önemli bilgilerin çıkarımına olanak sağlamamaktadır. Böylece ayırıcılığı olan ve olmayan terimler tanımlanmış olmaktadır.

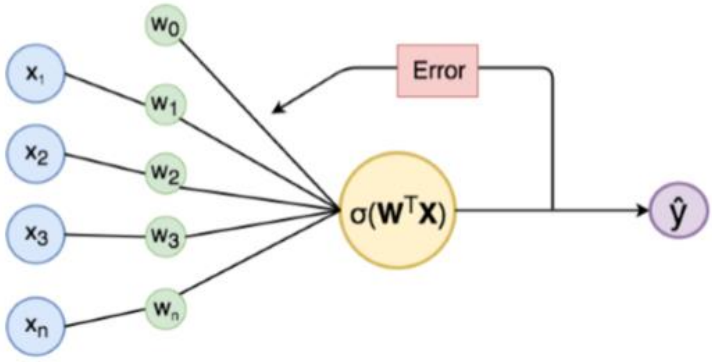
Aşama 3: Makine öğrenmesi. Makine öğrenme algoritmaları girdileri oluşturan yazı özellikleri ile çıktıları oluşturan puanlar arasındaki ilişkileri öğrenmektedir. Bu aşamada gerçek puanlayıcı puanlarından yararlanılmakta ve eğitim gerçekleştirilmektedir. Böylece özellikler puanlarla birlikte bilgisayara haritalandırılmaktadır. Makine öğrenmesi sürecinde kullanılabilecek farklı yöntemler bulunmaktadır. Bu araştırmanın konusu olması nedeniyle üç klasik makine öğrenmesine (lojistik regresyon, çok terimli sade bayes, destek vektör makineleri) dayalı algoritmaya ve yapay sinir ağlarına dayalı iki derin öğrenme (kısa-uzun süreli bellek, iki yönlü kısa-uzun süreli bellek) algoritmasına yer verilmiştir.

Lojistik regresyon (logistic regression). Lojistik regresyon yöntemi doğrusal ikili bir sınıflandırma yapmaktadır. Bu yöntem çoklu sınıflandırmalara kolaylıkla genellenebilmektedir. Model n boyuttaki bir x vektörünü girdi olarak kullanmaktadır. Buradaki boyut sayısı (n), x vektöründeki özelliklerin sayısına karşılık gelmektedir. Girdi özellikleri (x_i, i ∈ 1,2,3,…) lojistik regresyon kullanılarak $y=f(x_n,w_n)$ üzerine haritalandırılır. Bu denklemde yer alan w ağırlıklandırmayı göstermektedir ve en uygun ağırlık değerini bulmak için y fonksiyonu tekrarlanır. En uygun değer bulunurken negatif loglikelihood fonksiyonu kullanılır. Bu fonksiyon değişim ölçüsünün (gradyan) azalmasını kullanarak modelin hatasını bulur. Model ayrıca yanlılığı gösteren bir w_0 değeri içerir. Lojistik regresyon σ(z) ile gösterilen s biçimli (sigmoid) bir fonksiyon aracılığıyla çalışır. σ(z) fonksiyonu eşitlik 1'de gösterilmektedir. Verilen herhangi bir z değeri için s biçimli fonksiyon çıktı değerini 0 ile 1 aralığında verir. İkili puanlandırmada 0'a yakın çıktı değerleri ilk sınıfa, 1'e yakın çıktı değerleri ise ikinci sınıfa atanır. Kayıp fonksiyonu kullanılarak gerçek y değerleri ile tahmin edilen y değerleri ($\hat{y}$) arasındaki farktan hata hesaplanır. Çoklu sınıflandırma için y değeri eşitlik 2 ile elde edilmektedir.

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad (1)$$

$$y = f(x_n;w_n) = \sigma(w^T x) = \sigma(\sum_i^N w_i x_i) \quad (2)$$

Şekil 1 çoklu sınıflandırmaya dayalı bir lojistik regresyon (çok terimli lojistik fonksiyon [multinominal logistic regression]) modelini göstermektedir.



Şekil 1. Çok terimli Lojistik Regresyon modeli

Şekil 1 incelendiğinde x_i girdi özelliklerini; w_i, x_i özelliğine ilişkin ağırlığı, $\sigma(w^T x)$ s biçimli fonksiyonu, $\hat{y}$ tahmin edilen y değerlerini göstermektedir. Hata (error) gerçek y değeri ile $\hat{y}$ değeri arasındaki farktan hesaplanmaktadır.

Çok terimli sade bayes (multinominal naive bayes). Olaslığa dayalı sınıflandırma yapan bu yöntem ön bilgilere dayalı olarak çalışmaktadır. Bu yöntem özellikler arasında şartlı bağımsızlık varsayımına sahiptir. *D*, $F = (f_1, f_2, \ldots, f_n)$ özellikleri için bir vektörü temsil etmek üzere C_j sınıfına ait olasılık, eşitlik 3 ile hesaplanmaktadır. Her bir özellik için C sınıfına ait olasılık eşitlik 4 aracılığıyla bulunmaktadır. Eşitlik 3, eşitlik 4 aracılığıyla dönüştürülerek eşitlik 5 elde edilmektedir. *P(F)* bir sabit olup göz ardı edilebilmektedir. Dolayısıyla bu terim eşitlik 5'te yer almamaktadır.

$$P(C_j \mid D) = P(C_j \mid F) = \frac{\mathrm{P}(C_j)\, P(F \mid C_j)}{\mathrm{P(F)}} \qquad (3)$$

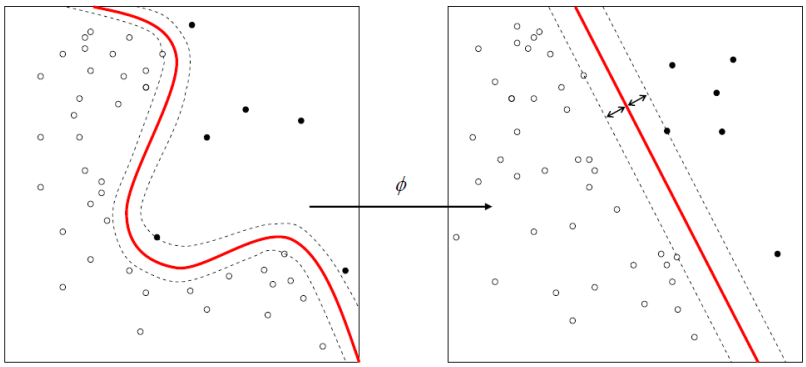
$$P(f_i \mid f_{i+1}, \ldots, f_n, C_j) = P(f_i \mid C_j) \qquad (4)$$

$$P(C_j \mid f_1, f_2, \ldots, f_n) = P(C_j) \prod_{i=1}^{n} P(fi \mid Cj) \qquad (5)$$

Çok terimli sade bayes en yüksek sonsallık (maximum a posteriori) kuralını kullanmaktadır. Bu karar kuralı, olasılığı maksimum yapan seçeneği likelihood fonksiyonuyla seçmektedir. Eşitlik 6 durumsal olasılığı maksimum hale getirerek y sınıfının seçilmesini sağlamaktadır.

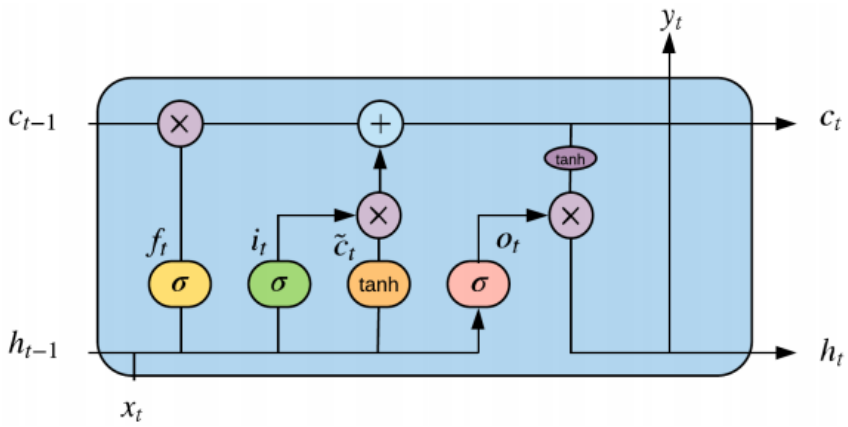
$$y = \operatorname{argmax} P(C_j) \prod_{i=1}^{n} P(fi \mid Cj) \qquad (6)$$

Destek vektör makineleri (support vector machine). Destek vektör makineleri eğitim verisini n boyuttaki bir düzlemde (n özellik sayısını belirtmektedir) tanımlayarak ikili doğrusal bir sınıflandırma yapmaktadır. Her bir veri sahip olduğu özelliklerden oluşturulan bir değerle düzlemde nokta olarak gösterilmektedir. Bu yöntem veriyi ikiye ayırarak en uygun hiper düzlemi bulmaya çalışmaktadır. Hiper düzleme yakın noktalar destek vektörleri olarak, destek vektörleri arasındaki boşluk aralık (margin) olarak adlandırılmaktadır. Yeni bir veri geldiğinde bu veri hiperdüzlemin herhangi bir tarafına yerleşmektedir. Bir dolgu değişkeni tanımlanarak yanlılık ve varyans dengelenmektedir. Eğer dolgu değişkeni küçük tanımlanırsa daha genel bir model elde edilmekte ve aralık artmaktadır. Doğrusallık bulunmadığında Kernel fonksiyonuna bağlı olarak doğrusal olmayan bir hiper düzlem oluşturulabilmektedir. Şekil 2 doğrusal ve doğrusal olmayan düzlemleri göstermektedir. Bu yöntem çoklu sınıflandırmalara da uyarlanabilmektedir.



Şekil 2. Doğrusal ve doğrusal olmayan hiper düzlemi gösteren grafikler

Kısa-uzun süreli bellek (long-short term memory). Yazılar işlenirken düzenli bir ileri besleme ağı kullanılması kelimelerin sayısını belirleyebilmekte fakat kelimelerin sırasının belirlenmesine olanak sağlamamaktadır. Uzun bir vektör olarak tanımlanan tekil yazılar için kelimelerin sıralarının dikkate alınması ve kelimelerin hatırlanması yinelenen sinir ağlarıyla mümkündür. Yinelenen sinir ağları her bir kelime için önceden gördüğü kelimeleri dikkate alarak bir iç döngü uygular. Ancak uzaktaki kelimeler birbirine bağlanırken sorun yaşanabilmektedir. Bu durum yok olan gradyan sorunu olarak tanınmaktadır. Bu sorunu ortadan kaldırmak üzere LSTM modeli kurulmuştur. Bu model yinelenen sinir ağı gibi yeni durumları belirlerken girdiyi (x_t) ve önceki durumları (h_{t-1}) kullanır. Fakat geçitler ile daha karmaşık bir şekilde çalışır. Önceki durum (h_{t-1}) ve şimdiki girdi (x_t), üç geçitten geçer. Bu geçitler girdi (i_t), unutma (f_t) ve çıktı (o_t) geçitleridir. Unutma geçiti (o_t) önceki hafıza hücresinin (c_{t-1}) bir sonraki hafıza hücresinde (c_t) ne kadar kullanılacağına şimdiki girdi (x_t) ile önceki duruma (h_{t-1}) bakarak karar verir. Giriş geçiti, önceki durum (h_{t-1}) ve girdiden (x_t) ne kadarının korunacağını belirler. Girdi (x_t) ve önceki durum (h_{t-1}) üzerinden etkinleştirme fonksiyonu kullanılarak güncel bir aday ($\tilde{c_t}$) belirlenir. Yeni hücre (c_t); güncel aday ($\tilde{c_t}$), girdi (i_t) ve çıktı (o_t) geçitleri kullanılarak oluşturulur. Son olarak yeni durum (h_t ya da çıktı y_t), yeni hücre (c_t) ve çıktı (o_t) geçidi kullanılarak oluşturulur. Şekil 3, LSTM yöntemine ait mimariyi göstermektedir.



Şekil 3. LSTM yöntemine ait mimari

Şekil 3 incelendiğinde LSTM modelinin her bir kelime için yaptığı işlem görülmektedir. LSTM yöntemiyle her bir kelime için bir vektör tanımlanmaktadır ve ard arda gelen vektörler incelenmektedir. LSTM yöntemiyle ardışık kelimelerin anlamsal bütünlüğüne yönelik çıkarım yapılabilmektedir.

İki yönlü kısa-uzun süreli bellek (bidirectional long-short term memory). Bu yöntem LSTM yöntemi ile aynı işlemleri gerçekleştirmektedir. Fakat BLSTM yönteminde yazı ilk kelimeden son kelimeye ve son kelimeden ilk kelimeye doğru incelenmektedir. Bu nedenle de iki LSTM katmanı içerir. Bir yazı LSTM yöntemiyle incelenirken yeni gelen kelime önceki kelimelerin anlamını değiştiremez. Bu durum da sistemin cümleyi hatalı anlamasına yol açabilir. BLSTM yönteminin iki yönden yapılan incelemenin bir kombinasyonu olması daha iyi performans göstermesini sağlayabilmektedir.

Aşama 4: Puan sınıflandırma ve hata analizi. Makine öğrenmesinde oluşturulan puanlama modeliyle bu aşamada yazılar derecelendirilmektedir. Otomatik puanlama sistemlerinde kurulan modeller eğer tek bir duruma yönelikse özgül cevap (prompt-specific), bir grup cevap için hazırlanmışsa ve birbirlerinin yerine kullanılabiliyorsa genel (generic) olarak adlandırılmaktadır. Daha doğru sonuçlar verilmesi nedeniyle genelde özgül cevap modeli üzerinde çalışılmaktadır. Puanlar sınıflandırıldıktan sonra puanlama modeli çeşitli performans ölçütleriyle değerlendirilmektedir. Bunlar uyum yüzdesi, karesel ağırlıklı Kappa değeri olabilir. Ölçütlerle ilgili detaylı bilgiye araştırmanın yöntem bölümünde yer verilmektedir.

Test Eşitleme

Test eşitleme iki ya da daha fazla test formuna ait puanları birbirine uyarlayarak bu puanların değiştirilebilir şekilde kullanılmasını sağlayan istatistiksel bir süreçtir. Eşitlemeye iki durumda ihtiyaç duyulmaktadır. Bunlar; 1) benzer güçlük düzeyine sahip testlerin birbirine denk gruplara uygulanması (yatay eşitleme), 2) farklı güçlük düzeyindeki testlerin farklı yetenek gruplarına uygulanması (dikey ölçekleme) durumlarıdır. Test eşitleme aracılığıyla farklı tarihlerde yapılan testlere katılma ya da bireylerin yıllara göre eğitimlerindeki ilerlemenin incelenmesi durumlarında doğru kararlar almak mümkün olabilmektedir (Hambleton ve Swaminathan, 1985; Kolen ve Brennan, 2014).

Test eşitleme desenleri. Test eşitleme desenleri temelde iki yaklaşım altında incelenebilir. Bu yaklaşımlardan birisi aynı ya da denk katılımcı grubuna test uygulaması yapmak iken (ortak evrene dayalı yöntemler) diğeri her bir test formunda ortak madde seti kullanmaya yöneliktir (ortak maddeye dayalı yöntemler). Bu yaklaşımların dışında ortak kişilerin kullanımına dayalı bir yaklaşım daha bulunmaktadır. Bu yaklaşım testlerdeki maddelerin kamu ile paylaşılması gerektiği durumlarda kullanılabilmektedir. Ancak bu yaklaşıma başvurulduğunda ortak kişilerin farklı özelliklere sahip olacak şekilde dikkatli seçilmesi ve madde konumu etkisinin dikkate alınarak test özelliklerinin benzer şekilde belirlenmesi gerekmektedir (Price, 2017; Wu, Tam ve Jen, 2016).

Ortak madde seti kullanımı yaklaşımında katılımcıların test formlarına seçkisiz olarak atanmaları gerekmezken, ortak evrene dayalı yöntemlerde katılımcılar test formlarına seçkisiz şekilde atanmalıdır. Her iki yaklaşımın kendine özgü güçlü yanları bulunmakla birlikte yaklaşımlara özgü bazı varsayımlar eşitleme sonuçlarına etkide bulunabilmektedir. Seçkisiz grup deseninde ortak bir evrenden seçkisiz olarak belirlenen iki gruba X ve Y formları verilmektedir. Bu desende ölçülen özellik açısından birbiriyle denk grupların oluşturulması oldukça önemlidir. Bu desenin ortak madde bulundurmaması, test sırası etkisi bulundurmaması ve test güvenliğini sağlaması bir avantaj olarak görülse de geniş örnekleme ihtiyaç duyduğu da bir gerçektir. Ayrıca test formlarının aynı özellikleri ölçtüğü gösterilmelidir. Tek grup deseninde aynı katılımcı grubu iki ayrı test formunu cevaplamakta ve bu nedenle kişilerin yeteneklerindeki farklılaşma olasılığı kontrol altına alınabilmektedir. Ancak bireylerin aldığı ilk testten etkilenerek ikinci testi yanıtlaması mümkündür. Bu

durum sıra etkisi problemini ortaya çıkarmaktadır. Böylece test formlarının güçlükleri arasındaki fark doğru bir biçimde belirlenemeyebilmektedir. Dengelenmiş tek grup deseninde, tek grup desenindeki sıra etkisi ortadan kaldırılmaktadır. Bu desende örneklem seçkisiz biçimde ikiye ayrılmakta gruplardan ilki, ilk sırada A testini, ikinci sırada B testini alırken, gruplardan ikincisi ilk sırada B testini, ikinci sırada A testini almaktadır. Dengelenmiş tek grup deseninin seçkisiz grup deseni ve tek grup deseninin bir kombinasyonu olduğu belirtilebilir. Bu desen küçük örneklemlerde doğru eşitleme sonuçları gösterse de işlevsel bir yöntem olmayıp özel bir çaba gerektirmektedir. Bahsedilen bu üç yöntem ortak evrene dayalı desenler olarak tanımlanmaktadır. Denk olmayan gruplarda ortak madde deseninde ise birbirine denk olması gerekmeyen iki örneklemden ilki X formunu, ikincisi Y formunu ve her iki grup da ortak maddeleri temsil eden A formunu almaktadır. Bu desende ortak maddelerin kullanılması grupların yetenekleri arasındaki farkı yansıtabilmektedir ve her iki grubun da aynı testleri yanıtlaması gerekmemektedir. Ancak ortak maddelerin testlerle aynı yapıyı ölçmesi ve testleri temsil etmesi (güçlük açısından) gerekmektedir. Ayrıca ortak maddelerin testlerde aynı konumda yer alması önem taşımaktadır. Denk olmayan gruplarda ortak madde deseninde X formunu yanıtlayan P grubu hiçbir zaman Y formunu, Y formunu yanıtlayan Q grubu ise hiçbir zaman X formunu yanıtlayamayacaktır. Bu yüzden denk olmayan gruplarda ortak madde deseni için birçok yöntem bulunmaktadır. Bu desen bulunduğunda grup farklılıkları ile test farklılıklarının ayrıştırılması ana görevdir. Denk olmayan gruplarda ortak madde deseninde iç ortak ve dış ortak maddeler kullanılabilmektedir. İç ortak maddeler kullanıldığında bu maddelerden alınan puanlar toplam test puanına dâhil edilirken, dış ortak maddeler kullanıldığında bu maddelerden alınan puanlar toplam test puanlarına dâhil edilmemektedir. Dış ortak maddeler genellikle farklı bir zaman kesitinde uygulanmaktadır (AERA, APA ve NCME, 2014; Dorans, Moses ve Eignor, 2010; Kolen ve Brennan, 2014; Price, 2017).

Test eşitleme yöntemleri. Denk olmayan gruplarda ortak madde deseni kullanıldığında klasik eşitleme yöntemleri ve MTK'ya dayalı yöntemler kullanılabilmektedir. Klasik eşitleme yöntemlerinden doğrusal ve eşit yüzdelikli eşitleme yaklaşımlarına, MTK eşitleme yöntemlerinden gözlenen puan eşitleme, eş zamanlı kalibrasyon ve ayrı kalibrasyona dayalı yöntemlere başvurulabilir. Bu araştırmanın konusu olan eşit yüzdelikli (zincir eşit yüzdelikli, frekans tahmini),

doğrusal (Tucker, zincir doğrusal), MTK'ya dayalı ayrı kalibrasyona dayalı (ortalama-ortalama, ortalama-standart sapma, Stocking-Lord, Haebara) yöntemlere ilişkin bilgilere aşağıda yer verilmektedir.

Klasik eşitleme yöntemleri. Klasik eşitleme yöntemleri için bu araştırmada iki doğrusal ve iki eşit yüzdelikli eşitleme yöntemi üzerinde çalışılmıştır. Aşağıda bu yöntemlere ilişkin yer alan bilgiler Albano (2016), Kolen ve Brennan (2013) ile Gonzalez ve Wiberg (2017)'den derlenmiştir.

Tucker doğrusal eşitleme. Bu yöntemde toplam test puanı ve ortak madde puanları regresyon eğimleriyle tanımlanır. γ_P, *P* evreninde *X* testinin *A* ortak testi üzerindeki regresyon eğimi; γ_Q, *Q* evreninde *Y* testinin *A* ortak testi üzerindeki regresyon eğimini göstemek üzere eşitlik 7 ve eşitlik 8 ile değerler hesaplanabilir.

$$\gamma_P = \frac{\sigma X_P, A_P}{\sigma^2_{Ap}} \qquad (7)$$

$$\gamma_Q = \frac{\sigma Y_Q, A_Q}{\sigma^2_{AQ}} \qquad (8)$$

X'in A üzerindeki doğrusal regresyon fonksiyonunun P ve Q evrenleri için aynı olması ve Y'nin A üzerindeki doğrusal regresyon fonksiyonunun P ve Q evrenleri için aynı olması Tucker yönteminin ilk varsayımıdır. İkinci varsayım A'ya bağlı olarak X'in varyansının P ve Q evrenleri, A'ya bağlı olarak Y'nin varyansının P ve Q evrenlerinde aynı olmasıdır.

Zincir doğrusal eşitleme. Bu yöntem P ve Q evreni için test puanlarının (X ve Y) ve ortak maddelere ilişkin puanların (A) ortalama ve standart sapmasının hesaplanmasıyla çalışmaktadır. Bu adım yerine getirildikten sonra X'i A'ya bağlayan ve A'yı Y'ye bağlayan doğrusal dönüşümler bir araya getirilmektedir. Bu yöntem eşitlik 9 aracılığıyla uygulanmaktadır.

$$\varphi = \mu_{(YQ)} + \frac{\sigma_{YQ}}{\sigma_{AQ}} [\mu_{(AP)} - \mu_{(AQ)}] - \frac{\sigma_{YQ}/\sigma_{AQ}}{\sigma_{XP}/\sigma_{AP}} \mu_{(XP)} + \frac{\sigma_{YQ}/\sigma_{AQ}}{\sigma_{XP}/\sigma_{AP}} X \qquad (9)$$

Frekans tahmini eşit yüzdelikli eşitleme. Frekans eşit yüzdelikli eşitleme yöntemi sentetik evrene dayalı olarak çalışmaktadır. Bu yöntemde sentetik evrene dayalı tüm puan dağılımlarının X ve Y formlarını alması gerekmektedir.

P(x) ve P(y) dağılım fonksiyonlarını göstermektedir ve bu fonksiyonlar sırasıyla eşitlik 10 ile eşitlik 11 aracılığıyla bulunabilir. Eşitlik 10 ve 11'in kümülatif

versiyonları için yüzdelik sıralar bulunarak eşit yüzdelikli fonksiyona erişilebilmektedir. Bu amaç için kullanılan denklem eşitlik 12'de gösterilmektedir.

$$P(x)= w_P P_P(x) + w_Q \sum PP(x|v)PQ(v) \qquad (10)$$

$$P(y)= w_Q P_Q(y)+ w_P \sum PQ(y|v)PP(v) \qquad (11)$$

$$equip_Y(x)= G-1[F(x)] \qquad (12)$$

Zincir eşit yüzdelikli eşitleme. Bu yöntemde X ve Y test puanları, ortak test kullanılarak birbirine bağlanmaktadır. Öncelikle test formu X, A testine eşitlenmekte ardından test formu A, Y testine eşitlenmektedir. Hem eşit yüzdelikli hem de doğrusal zincir eşitleme yöntemi diğer yöntemlerden farklı olarak sentetik evreni desteklememektedir. Eşitlik 13 zincir eşit yüzdelikli eşitlemede kullanılan denklemi göstermektedir.

$$\varphi_y(x) = F_{YQ}^{-1}(F_{AQ}(F_{AP}^{-1}(F_{XP}(x)))) = \varphi_{YQ}(\varphi_{AP}(x)) \qquad (13)$$

Düzgünleştirme (smoothing). Evrenden örneklem seçimi örnekleme hatasına dolayısıyla da veride düzensizliğe neden olmaktadır. Bu düzensizlikleri azaltmada eşit yüzdelikli eşitleme yöntemlerinde düzgünleştirme kullanılabilmektedir. Düzgünleştirme iki şekilde yapılabilmektedir. Bunlar; puan dağılımının düzgünleştirildiği ön düzgünleştirme ve eşit yüzdelikli eşitliklerin düzgünleştirildiği son düzgünleştirmedir. Ön düzgünleştirme model parametreleri kullanarak puan dağılımlarını modellemektedir. Denk olmayan gruplarda ortak madde deseni kullanıldığında X testi için puan olasılıkları (p_{jl})=$Pr(X=x_j, A=a_l)$ denklemiyle elde edilmektedir. Aynı denklem Y testi için de kullanılmaktadır (Gonzalez ve Wiberg, 2017; Kolen ve Brennan, 2014). Bu araştırmada alanyazında sıklıkla karşılaşılan polinominal iki değişkenli logaritmik doğrusal fonksiyon kullanılmıştır. Aşağıda bu yönteme ilişkin bilgilere yer verilmektedir.

Polinominal logaritmik doğrusal fonksiyon. Denk olmayan gruplarda ortak madde deseni kullanıldığında her bir test için iki değişkenli bir model kurulmaktadır. Bu iki değişkenden birincisi X testinden alınan puan, ikincisi ise ortak maddelerden alınan puandır (A). Bu doğrultuda aşağıda yer alan eşitlik 14 kullanılarak modelleme işlemi gerçekleştirilmektedir (Moses ve von Davier, 2006). Bu denklemde yer alan p_{jk} puan olasılığını, x_j X testindeki puanları, a_k A testindeki puanları, x_j^i X testinin tek değişkenli dağılımına ilişkin kuvvetini (*i* 1, 2, 3, 4 olabilir bu değerler dağılımda

sırasıyla ortalama, standart sapma, çarpıklık, basıklığı gösterir), a_k^h *A* testi puanlarının tek değişkenli dağılımına ilişkin kuvvetini (h 1, 2, 3, 4 olabilir) ve $x_j^g a_k^f$ terimi çapraz kuvvetleri (xa [kovaryans], g ve f değerleri 1 ve 2 olabilir) ve α normalleştirme sabitini göstermek üzere;

$$log_e^{p_{jk}} = \alpha + \sum_{i=1}^{I} \beta_{xi}(x_j)^i + \sum_{h=1}^{H} \beta_{ah}(a_k)^h + \sum_{g=1}^{G}\sum_{f=1}^{F} \beta_{gf}(x_j)^g(a_k)^f \quad (14)$$

Ön düzgünleştirme için bahsedilen tek değişkenli dağılıma ilişkin kuvvetler ve çapraz kuvvetler kullanılarak çeşitli modeller oluşturulabilmektedir. Bu modellerin en uygun olanının seçilmesi için kullanılacak ölçütlere yöntem bölümünde yer verilmiştir.

MTK'ya dayalı eşitleme yöntemleri. Denk olmayan gruplarda ortak madde deseni kullanıldığında farklı test formlarındaki parametreler aynı ölçek üzerine yerleştirilmelidir. Bunun nedeni X formundaki parametre kestirimlerinin P grubundan, Y formundaki parametre kestirimlerinin Q grubu üzerinden yapılması, P ve Q grubunun da denk olmamasıdır. Her ne kadar yetenek parametreleri kestirilirken 0 ortalama ve 1 standart sapmaya sahip olacak şekilde işlem yapılsa da yetenek parametreleri yine de farklılaşmaktadır. Tüm bu durumlar dikkate alınarak doğrusal bir denklem aracılığıyla ölçek dönüşümü yapılmaktadır (Kolen ve Brennan, 2014). Eşitlik 15, 16 ve 17 sırasıyla *a* parametreleri, *b* parametreleri ve *Θ* (yetenekler) için yapılan dönüşümleri göstermektedir. Θ_i *i* katılımcısına ait yeteneği, Θ_i^* *i* katılımcısının dönüştürülen yeteneğini, a_j *j* maddesine ait ayırıcılık parametresini, a_j^* *j* maddesine ait dönüştürülmüş ayırıcılık parametresini, b_j *j* maddesine ait güçlük parametresini, b_j^* *j* maddesine ait dönüştürülmüş güçlük parametresini, *A* eşitleme denkleminin eğimini, *B* eşitleme denkleminin sabit değerini göstermek üzere;

$$a_j^* = \frac{aj}{A} \quad (15)$$

$$b_j^* = Ab_j + B \quad (16)$$

$$\Theta_i^* = A\Theta_i + B \quad (17)$$

Ölçek dönüştürme işlemi yapılırken yöntemlerin farklı yaklaşımları temel alması nedeniyle *A* (eğim) ve *B* (sabit) değerleri farklılaşmaktadır. Aşağıda ölçek dönüştürmede kullanılan moment (ortalama-standart sapma, ortalama-ortalama) ve

karakteristik eğri (Haebara, Stocking-Lord) yöntemlerine ilişkin bilgilere kısaca yer verilmektedir. Denklemlerin yazılmasında Kolen ve Brennan (2014) kaynağından yararlanılmıştır.

Ortalama-standart sapma. Marco (1977) güçlük parametresinin ortalamasını ve standart sapmasını kullanarak A ve B katsayılarını hesaplamaya yönelik olarak eşitlik 18 ve eşitlik 19'u sunmuştur. Bu eşitliklerde *μ(bi) i* ölçeğinin ortalama güçlüğünü, *μ(bj) j* ölçeğinin ortalama güçlüğünü, *σ(b_i) i* ölçeğinin ortalama standart sapmasını, *σ(b_j) j* ölçeğinin ortalama standart sapmasını göstermek üzere;

$$A = \frac{\sigma(bj)}{\sigma(bi)} \qquad (18)$$

$$B = \mu(bj) - A\,\mu(bi) \qquad (19)$$

Ortalama-ortalama. Loyd ve Hoover (1980) ayırıcılık ve güçlük parametrelerinden yararlanarak A ve B katsayılarını hesaplamaya yönelik olarak eşitlik 20 ve eşitlik 21'i sunmuştur. Bu denklemde; *μ(ai) i* ölçeğinin ortalama ayırıcılığını, *μ(aj) j* ölçeğinin ortalama ayırıcılığını, *μ(bi) i* ölçeğinin ortalama güçlüğünü, *μ(bj) j* ölçeğinin ortalama güçlüğünü göstermek üzere;

$$A = \frac{\mu(ai)}{\mu(aj)} \qquad (20)$$

$$B = \mu(bj) - A\,\mu(bi) \qquad (21)$$

Haebara. Haebara (1980) madde karakteristik eğrilerinden yararlanarak bir eşitlik tanımlamıştır. Eşitlik 22 katılımcıların belirli bir yetenek düzeyi için madde karakteristik eğrileri farkını (*Hdiff*), her bir madde için madde karakteristik eğrisindeki farkın karelerini toplayarak bulmaktadır. (j:V) ortak maddeleri, fark iki ölçek üzerindeki madde karakteristik eğrilerini, *i* kişileri, *j* maddeleri, *a* ayırt edicilik, *b* güçlük, *c* şans parametrelerini temsil etmek üzere;

$$Hdiff = \sum_{j:V}[p_{ij}(\Theta_{JI}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - p_{ij}(\Theta_{JI}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij})]^2 \qquad (22)$$

A ve *B* katsayıları eşitlik 23'ü minimum yapacak şekilde elde edilmektedir.

$$Hcrit = \sum_i Hdiff(\Theta_i) \qquad (23)$$

Stocking-Lord. Stocking ve Lord (1983), Haebara (1980)'a benzer biçimde madde karakteristik eğrilerini kullanarak eşitlik 24'ü tanımlamıştır. Bu eşitlik

katılımcıların belirli bir yetenek düzeyi için madde karakteristik eğrileri farkını (*SLdiff*), her bir madde için madde karakteristik eğrileri toplamı arasındaki farkın karesi ile hesaplamaktadır. Eşitlik 24 içerisinde yer alan terimler HB yöntemi ile aynı olmak üzere;

$$SLdiff = [\sum_{j:V} p_{ij}\left(\Theta_{JI}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}\right) - \sum_{j:V} p_{ij}(\Theta_{JI}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij})]^2 \qquad (24)$$

A ve *B* katsayıları eşitlik 25'i minimum yapacak şekilde elde edilmektedir.

$$SLcrit = \sum_i SLdiff(\Theta_i) \qquad (25)$$

İlgili Araştırmalar

Otomatik puanlama ile ilgili araştırmalar. Otomatik puanlama ile ilişkili alanyazın iki bölüm halinde incelenmiştir. İlk bölümde yer alan araştırmalar otomatik puanlamada farklı yaklaşımları ve yöntemleri kullanan araştırmalara yöneliktir. İkinci bölüm ise Türk dili ile benzer özellik gösteren dillerde geliştirilen otomatik puanlama sistemlerine yöneliktir.

Gierl, Latifi, Lai, Boulais ve Champlain (2014) otomatik madde puanlama sistemleriyle ilgili alanyazını açıklamak ve tıp eğitiminde otomatik puanlama sisteminin nasıl kullanılacağını göstermek üzere bir araştırma gerçekleştirmiştir. 4 adımda gerçekleşecek otomatik puanlama sistemini uygulamak üzere Kanada Tıp Konseyi tarafından 2011, 2012, 2013 yılında yapılan ve yapılandırılmış cevaplardan oluşan maddelerin kullanıldığı testlerden 2013 yılındaki teste otomatik madde puanlama sistemini uygulamışlardır. Öncelikle yanıtlar iyi yapılandırılmış dereceli puanlama anahtarıyla iki uzmana puanlatılmıştır. Bu, otomatik puanlama için ilk aşama olarak açıklanmıştır. Gerçek puanlayıcıların kesin bir puan üzerinde anlaşmaları sağlanmıştır. 2011 ve 2012 yılındaki veriler, sırasıyla özellikleri çıkartmak ve makine öğrenmesi için kullanılmıştır. Bu aşama otomatik madde puanlamanın ikinci ve üçüncü aşaması olarak belirtilmiştir. Özellikleri çıkarma aşamasında n-gram temsilinden yararlanılmıştır. Makine öğrenme yöntemlerinden destek vektör makinesi (SVM) kullanılarak çıkarılan girdi özellikleri, çıktı olan puanlayıcı puanlarına haritalandırılmıştır. 10 kat çapraz geçerlik kullanılarak haritalandırma fonksiyonu geliştirilmiştir. Dördüncü aşamada ise 2013 yılı verileri otomatik puanlama sistemince ve gerçek puanlayıcılarla puanlanarak sınıflama doğruluğu ve hata araştırılmıştır. Bu amaçları gerçekleştirmek için LightSIDE yazılımı kullanılmıştır. Sonuçta maddeler için gerçek puanlayıcılar ve otomatik

puanlama arasındaki uyumun ,95 ile ,98 arasında değiştiği bulunmuştur. Kappa değerlerinin ise ,88 ile ,96 arasında değiştiği bulunmuştur. Araştırma sonuçları gerçek puanlayıcılar ile otomatik puanlama sistemleri arasında çok iyi bir uyum bulunduğunu göstermiştir.

Adesiji, Agbonifo, Adesuyi ve Olabode (2016) araştırmalarında teste dayalı betimleyici bir otomatik puanlama sistemi (automated descriptive test-based scoring system) geliştirmiş olup bu program alan bilgisi, metin gözden geçirici ve puanlama aracı olmak üzere üç modül içermektedir. Alan bilgisi modülünde öğrenciden beklenen cevaplara ilişkin bir anahtar kelimeler kümesi bulunmakta ve öğrencinin verdiği cevaplardaki kelimelerle anahtar kelimeler karşılaştırılarak eşleşen kelime sayısı belirlenmektedir. Metin gözden geçirici modülü öğrenci cümlelerini ve kelimelerini gözden geçirerek cevapları düzenlemekte ve sınıflandırmaktadır. Cümleler nokta ile tespit edilmekte ve her cümle özne, yüklem ve nesne olmak üzere üç parçada değerlendirilmektedir. İnceleme yapılırken isim, sıfat, zamir, zarf, bağlaç, edat, ünlem ve fiil dikkate alınmaktadır. Cümleler incelenirken cümlenin gramer olarak doğruluğu ve noktalama işaretlerinin geçerli olması incelenmektedir. Kelimeler incelenirken isim ve sıfatlar seçilmekte tamlayan ve edatlar göz ardı edilerek kontrol gerçekleştirilmektedir. Yan yana yazılan isim ve sıfat sayısı incelenmekte eğer yan yana yazılan isim ve sıfat sayısı üçün üzerinde ise cümle anlamlı olmayacağından bu durum kurallara aykırı olarak değerlendirilmektedir. Böylece kurnaz yazarlar tespit edilmektedir. Cevapların düzenlenmesinde şablon cevaplar kullanılmakta cümledeki noktalama işaretleri ve boşluklar silinmektedir. Cevapların sınıflandırılması aşamasında cevaplar dört bölümde sınıflandırılmaktadır. Kategoriler sınıfında ikili kelimeler ve grup tipinde beklenen cevaplar bulunmaktadır. Vurgulama sınıfında tanımlar ve kısa açıklamalar yer almaktadır. Listeleme ve vurgulama sınıfında öğrencilerden listeleme yapmaları ve listelediklerini açıklamaları beklenmektedir. Tartışma/açıklama sınıfında detaylı açıklama gerektiren maddelerle ilgili işlem yapılmaktadır. Puanlama aracı cevabın sınıflandırılması aracılığıyla çalışmakta, puanı sınıflandırma (marker class) ve çok değişkenli Bernolli modeli ile yazı sınıflandırması yaparak puan atama işlemini gerçekleştirmektedir. Kategori sınıfında puanların %20'si bulunan kelimelerden, %40'ı kelime çiftlerinde yer alan her bir kelimeye verilen puanlardan oluşur. Vurgulama sınıfında puanların %50'si vurgudan, %50'si ise bulunan kelimelere

verilen puanlardan oluşur. Listeleme ve vurgulama sınıfında puanların %30'u listelemeden, %60'ı vurgulamadan ve %10'u ise bulunan kelimelere verilen puanlardan oluşur. Tartışma/açıklama sınıfında puanların %30'u bulunan kelimelere verilen puanlardan, %70'i yazı gözden geçiriciden gelen puanlardan oluşur. 50 üniversite öğrencisi üzerinde geliştirilen bu sistem %73,3 doğruluk ile çalışmaktadır.

Taghipour ve Tou Ng (2016) yazı ile puan arasındaki ilişkileri öğrenen bir yaklaşımla otomatik puanlama gerçekleştirmiştir. Bu yöntem özelliklerin el ile belirlenmesine dayalı regresyon yöntemlerinden farklılık göstermektedir. Yazının içeriği, grameri, organizasyonu gibi puanlamada önem arz eden tüm faktörlerin düşünülmesine gerek kalmamakta yinelenen sinir ağları ile karmaşık modelleri çözümleyebilmektedir. Bahsedilen bu yaklaşım modelleme ve genelleme konusunda üstünlüklere sahiptir. Araştırmada yinelenen sinir ağlarına dayalı üç yöntem ele alınmıştır. Bunlar; 1) temel yineleyen birimler (basic recurrent units), 2) aralıklı yineleyen birimler (gated recurrent units), 3) kısa-uzun süreli bellek birimleri (long-short term memory units)'dir. Gerçek puanlayıcılar ve otomatik puanlama arasındaki uyum incelenerek yöntemlerin performansları değerlendirilmiştir. Bu amaçla karesel ağırlıklı Kappa (quadratic weighted Kappa-QWK) değeri kullanılmıştır. Araştırmacılar yinelenen sinir ağlarına dayalı bu yöntemlerden en iyi modelin uzun-kısa süreli bellek (long-short term memory-LSTM) sinir ağı olduğunu belirlemişlerdir. Bahsedilen bu yönteme ait QWK değeri ,746'dır. BLSTM yönteminin LSTM yönteminden biraz daha düşük QWK değerine (,699) sahip olduğu bulunmuştur. Sistemin değerlendirilmesinde ise beş kat çapraz doğrulama kullanılmış olup her kata verilerin %60'ı eğitim, %20'si gelişim ve %20'si ise test amaçlı atanmıştır. LSTM yöntemi ile gerçek puanlayıcı grupları arasında bulunan uyum, gerçek puanlayıcıların kendi aralarındaki uyuma yakındır.

Türk diline dair bilgiler. Türk dili 19. yüzyılın başlarına dek Ural-Altay dil ailesinde görülmekte idi. Bunun nedeni ortak kelimeler, erilik-dişilik ayrımının bulunmaması, sondan eklemelilik, kimi yapı benzerlikleri ve çok eski dönemlere dayanan ilişkiler gibidir. Fakat diller arasında genetik bir bağın bulunmayışı, dillerin farklı bölgelerde yer alışı ve etnik kökendeki farklılık bunun yanı sıra temel ses denkliklerinin farklılaşması Ural ve Altay dil ailelerinin ayrı ele alınmasını gerektirmiştir. Bahsedilen ayrım sonrası Türkçe Altay dil ailesinde incelenmeye devam etmiştir. Ural dil ailesinin en tanınan üyeleri Fin, Macar ve Eston dilleridir

(Akar, 2017; Eker, 2017). Bu araştırmada her ne kadar farklı dil ailelerine mensup olsa da dillerin yapı benzerliklerine sahip olması nedeniyle Ural dil ailesine özgü geliştirilen otomatik puanlama sistemlerine de alanyazında yer verilmiştir. İncelenen alanyazında Fin diline ait bir otomatik puanlama sistemine rastlanmış ve bu otomatik puanlama sistemi metnin ilerleyen bölümünde açıklanmıştır. Altay dil ailesinde Moğolca, Mançu-Tunguzca, Korece ve Japonca dilleri yer almaktadır (Akar, 2017; Eker, 2017; Ercilasun, 2016). Bu diller genetik bir bağ göstermekte olup benzer özelliklere sahiptir. Bu özellikler; 1) ses uyumu bulunması, 2) bitişiklik bulunması (kelimeler kök ve gövdeye ek getirilerek yapılmaktadır), 3) yalnızca son ek bulunması (kelime önüne ya da içine ek getirilmemektedir), 4) cümlelerin sırasıyla özne-tümleç-yüklem içermesi, 5) tamlamalarda tamlayanın tamlanandan önce gelmesi, 6) sıfat tamlamalarında tamlayan ile tamlanan arasında hal, cinsiyet ve sayı bakımından farklılıklar bulunmaması, 7) çokluk bildiren sayılardan sonra gelen isimlerin çokluk eki almaması, 8) kelimelerin gramatikal cinsiyeti bulunmamasıdır (Akar, 2017; Eker, 2017). Altay ailesinde yer alan diller için alanyazın tarandığında Japonca ve Korece dilleri için hazırlanan otomatik puanlama sistemleri bulunmuştur.

Türk dili ile benzer özellikler gösteren dillerdeki otomatik puanlama sistemleri. Aşağıda sırasıyla Fin, Kore ve Japon dilleri için hazırlanan otomatik puanlama sistemleri ile ilgili bilgilere yer verilmektedir.

Kakkonen, Myller, Timonen ve Sutinen (2005) yaptıkları araştırmada Fin dilinde yazılmış metinleri puanlayacak bir otomatik puanlama sistemi oluşturmuşlardır. Oluşturdukları sistemde yazıları puanlarken sıkça başvurulan iki yöntemden bir arada yararlanmışlardır. Bu yöntemler; 1) puanlanacak yazının gerçek puanlayıcılar tarafından puanlanması ve benzer bir yazı ile karşılaştırma, 2) puanlanacak yazının başlığıyla ilgili kaynaklardan yararlanma. Sistem öncelikle her bir tekil kelimenin gözlenme sıklığını incelemektedir. Fin dili sondan eklemeli bir dil olduğundan son ekleri ayrıştıran bir modül kullanılmaktadır. Her bir bağlam için öncelikle konu ile ilgili metinlerle karşılaştırma yapılmaktadır. Karşılaştırma yapılırken farklı yöntemlerle (örtük anlamsal [latent semantic], olasılıksal örtük anlamsal [probabilistic latent semantic]) benzerliği belirlemeye dayalı vektörler kullanılmaktadır. 6 veri seti üzerinde gerçekleştirilen araştırma sonuçlarına göre örtük anlamsal yöntemin daha iyi sonuçlar verdiği gözlemlenmiştir. Bu yöntem için

gerçek puanlayıcılar ile otomatik puanlama sistemi arasındaki korelasyon katsayıları ,54 ile ,90 aralığında değişmektedir.

Jang, Kang, Noh, Kim, Sung ve Seong (2014) yaptıkları araştırmada yanıtı sınırlandırılmış ve sınırlandırılmamış açık uçlu maddelerde puanlama yapacak Kore diline dayalı bir otomatik puanlama sistemini oluşturmuşlardır. Araştırmada 7 madde üzerinde çalışılmış olup her bir maddeye yaklaşık 3000 öğrenci yanıt vermiştir. Sistem doğal dil işlemeye dayalı biçimsel özellikler üzerine kurgulanmış olup göstergelere dayalı bir puanlama şablonuna sahiptir. Sistem girdi olarak öğrenci cevaplarını ve gerçek puanlayıcılar tarafından oluşturulan özel bir puanlama rehberi gerektirmektedir. Puanlama rehberi yüksek frekanslı öğrenci cevaplarını, kavramları, anahtar kelimeleri ve puanlama seçeneklerini içermektedir. Cevap şablonu oluşturulurken seçenek adımında yazım yanlışları, kelime boşlukları düzeltilmekte, gereksiz işaretçiler silinmekte ve anahtar kelime uygulaması yapılmaktadır. Bu aşamadan sonra öğrenci cevaplarında yer alan kelimeler frekansa göre sıralanmakta ve gerçek puanlayıcılar tarafından puanlanmaktadır. Gerçek puanlayıcılar bu süreçte bir miktar cevabı daha puanlamakta ve bu şekilde kavramlar elde edilmektedir. Anahtar kelime seçeneği kullanıldığında gerçek puanlayıcılar sisteme anahtar kelimeleri tanımlamakta bu şekilde yanlış cevaplar tespit edilebilmektedir. Öğrenci cevaplarına yazım yanlışları, kelime boşlukları ve gereksiz işaretçiler için normalleştirme uygulandıktan sonra otomatik puanlama aşamasında model cevapları ile eşleşme, yüksek frekanslı cevaplarla eşleşme, kavrama dayalı değerlendirme ve yanlış cevaplar için anahtar kelime incelemesi yapılmaktadır. Kavram değerlendirme aşamasında anlamsal ve gramere dayalı işlem yapılmaktadır. Son adımda işlenmemiş cevaplar için kavramlar birleştirilmektedir. Sistem öğrencilerin yaklaşık %90-95'ine ilişkin cevapları bu şekilde puanlamıştır. Araştırma sonucunda gerçek puanlayıcılarla yaklaşık %95 uyum yüzdesine ulaşılmıştır. Sonuçlar sistemin geniş ölçekli testlerde kullanılabileceğini göstermiştir.

Ishioka ve Kameda (2006) Japon dili üzerinde geliştirdikleri otomatik puanlama sisteminde uzman puanlayıcılar yerine uzman metinlerini kullanmışlardır. Bu sistem üç ölçütle yazıları değerlendirmektedir. Bunlar; 1) etkili yazma, 2) organizasyon, 3) kapsamdır. Etkili yazma evrelerin, yan cümlelerin ve cümlelerin düzenlenmesi ile ilgili sözdizimsel çeşitliliği incelemektedir. Etkili yazma

aşamasında kelimelerin çeşitliliği, uzun kelimelerin yüzdesi, pasif cümlelerin yüzdesi ve okuma kolaylığı değerlendirilmektedir. Organizasyon, fikirlerin düzgün bir şekilde ifade edilmesiyle ilgili özellikleri değerlendirmektedir. Kapsam ise başlıkla alakalı kelimeleri, ilgili bilgileri ve belirli veya özel kelimeleri incelemektedir. Son değerlendirme aşamasında bir gazetede yer alan başyazılar ve makaleler kullanılarak bir öğrenme süreci gerçekleşmekte ve puan ataması yapılmaktadır. Sistemin performansını değerlendirmek üzere sınıflararası korelasyon katsayısı (interclass correlation [ICC]) hesaplanmıştır. Araştırma sonucunda sistemin 800-1600 karakter aralığındaki yapılandırılmış cevap maddelerini puanlamada geçerli olduğu görülmüştür. Yapılan iki denemede de otomatik puanlama sisteminin gerçek puanlayıcılarla olan korelasyonunun, gerçek puanlayıcılar arasında hesaplanan korelasyondan yüksek olduğu bulunmuştur. 143 üniversite öğrencisinin sigara kullanımı konusunda yazdıkları yazılar için otomatik puanlama sistemi ile gerçek puanlayıcılar arasındaki korelasyon ,83, gerçek puanlayıcıların kendi aralarındaki korelasyon ise ,70 bulunmuştur.

Test eşitleme ile ilgili araştırmalar. Alanyazında KTK ve MTK'ya dayalı eşitleme yöntemlerini karma formattaki testlerde karşılaştıran az sayıda araştırmaya rastlanmıştır. Bu araştırmaların bir kısmı gerçek veriye dayalı iken bir kısmı ise benzetim çalışmasıdır. Araştırmaların bir kısmı eşitleme performansını seçkisiz ve sistematik hata açısından karşılaştırırken bir kısmı ise eşitleme doğruluğunu eşitleme özelliklerinin korunumuna dayalı olarak incelemektedir. Alanyazında ulaşılan ve bu araştırmayla bağlantılı olan araştırmalara aşağıda yer verilmektedir.

Hagge ve Kolen (2011) gerçekleştirdikleri araştırmada denk olmayan gruplarda ortak madde deseni kullanarak bir yerleştirme sınavının İspanyolca testini eşitlemişlerdir. Araştırmada dört faktör üzerinde inceleme yapılmıştır. Bunlar; 1) eski ve yeni test formlarını alan grupların beceri farklılıkları, 2) çoktan seçmeli ve yapılandırılmış cevap maddelerinin göreli güçlükleri, 3) ortak maddelerin formatı temsili, 4) eşitleme yöntemleridir. Eşitleme yöntemi olarak KTK'ya dayalı frekans tahmini ve zincir eşit yüzdelikli; MTK'ya dayalı gözlenen puan ve gerçek puan (Haebara) seçilmiştir. Eşit yüzdelikli eşitleme yöntemlerine kübik eğri yöntemi ile son düzgünleştirme uygulanmıştır. Frekans tahmini ve MTK gözlenen puan eşitleme yöntemleri için sentetik evren oranı 1 olarak yani yeni test formunu alan gruba dayalı olarak belirlenmiştir. Ortak maddeler sadece çoktan seçmeli ve çoktan seçmeli

maddelerin yanında yapılandırılmış cevap maddelerini içerecek şekilde tasarlanmıştır. Sonuçlar eşitlemenin standart hatası, yanlılık ve hata kareleri ortalamasının karekökünün ağırlıklandırılmasıyla incelenmiştir. Araştırma sonucunda ortak madde olarak her iki madde türünün kullanıldığı, grupların denk olduğu, çoktan seçmeli maddelerle yapılandırılmış cevap maddelerinin benzer yapıları ölçtüğü koşulda eşitlemenin standart hatasına göre en düşük hatayı MTK gözlenen puan eşitleme, en yüksek hatayı zincir eşit yüzdelikli eşitlemenin verdiği bulunmuştur. En yüksek yanlılığı gösteren yöntemin frekans tahmini, en düşük yanlılığı gösteren yöntemlerin ise MTK gözlenen ve gerçek puan eşitleme olduğu sonucuna ulaşılmıştır. RMSE açısından değerlendirme yapıldığında MTK'ya dayalı yöntemlerin, KTK'ya dayalı yöntemlerden daha düşük değerlere sahip olduğu belirtilmektedir.

Hagge, Liu, He, Powers, Wang ve Kolen (2011) araştırmasında denk olmayan gruplarda ortak madde deseni ve gerçek veri kullanarak frekans tahmini, zincir eşit yüzdelikli, Madde Tepki Kuramına dayalı gerçek puan (Haebara) ve gözlenen puan eşitleme yöntemlerini karma testler üzerinde karşılaştırmıştır. Ek olarak eşit yüzdelikli eşitleme yöntemlerinde (frekans tahmini ve zincir eşit yüzdelikli) ön düzgünleştirme ve son düzgünleştirmenin etkisini incelemişlerdir. Ön düzgünleştirme için iki değişkenli logaritmik doğrusal fonksiyon (bivariate log-linear) kullanılırken son düzgünleştirme için kübik eğri (cubic spline) yöntemini kullanmışlardır. Frekans tahmini ve MTK gözlenen puan eşitleme için sentetik evren oranını 1 olarak belirleyerek yeni formu alan grup üzerine odaklanmışlardır. Sonuçlar değerlendirilirken eşitlemenin standart hatası (seçkisiz hata [SEE]) ortalamalarını kullanmışlardır. Araştırma sonucunda beş farklı gerçek veri üzerinde yapılan eşitleme işlemleri göstermiştir ki MTK'ya dayalı yöntemler KTK'ya dayalı yöntemlerden daha düşük standart hataya sahiptir. MTK'ya dayalı gözlenen puan eşitleme, MTK'ya dayalı gerçek puan eşitlemeden biraz daha düşük; KTK'ya dayalı frekans tahmini, KTK'ya dayalı zincir eşit yüzdelikli eşitlemeden daha düşük standart hataya sahiptir. Ön düzgünleştime ve son düzgünleştirme frekans tahmini ve zincir eşit yüzdelikli eşitleme yöntemlerinde elde edilen standart hatayı azaltmıştır. Ön düzgünleştirme, son düzgünleştirmeden daha düşük standart hatalar göstermiştir.

He (2011) denk olmayan gruplarda ortak madde deseni kullanarak gerçekleştirdiği araştırmada karma formattaki testlerde eşitleme yöntemlerinin, test

özelliklerinin ve form benzerliğinin eşitlemeye etkisini incelemiştir. Çoktan seçmeli ve yapılandırılmış cevap maddeleri arasındaki ilişki, çoktan seçmeli maddelerin puanlarının oranı, ortak madde oranı, formlar arasındaki farklılıklar, eşitleme yöntemleri ortak madde olarak sadece çoktan seçmeli maddelerin kullanıldığı testlerde incelenen faktörlerdir. Gerçek veriler üzerinde, yerleştirme amaçlı bir sınav üzerinde gerçekleştirilen bu araştırmada KTK'ya dayalı frekans tahmini ve zincir eşit yüzdelikli eşitleme, MTK'ya dayalı gerçek puan (Stocking-Lord [SL]) ve gözlenen puan eşitleme yöntemleri kullanılmıştır. Ek olarak frekans tahmini ve zincir eşit yüzdelikli eşitleme yöntemlerine kübik eğri (cubic splines) metodu ile son düzgünleştirme uygulanmıştır. Sonuçlar değerlendirilirken üç kriterden yararlanılmıştır. Bunlar; 1) birincil düzey eşitlik (first order equity), 2) ikincil düzey eşitlik (second order equity) ve benzer dağılım özelliğidir. Birincil düzey eşitlik, eşitleme işlemi sonrası koşullu ortalamaların (conditional means) benzerliğini gerektirirken ve birey açısından test tarafsızlığını gösterirken; ikincil düzey eşitlik, eşitleme işlemi sonrası ölçmenin koşullu standart hatasının (conditional standard error of measurement) eşitliğini gerektirmekte ve ölçmenin doğruluğunu göstermektedir. Benzer dağılım özelliği ise dağılımlar arasındaki mutlak farklılıklara ilişkin olarak hesaplanmakta ve gruplar açısından test tarafsızlığını göstermektedir. Araştırma sonucunda birincil düzey eşitlik özelliğine göre MTK gerçek puan eşitlemenin, MTK gözlenen puan eşitlemeden; ikincil düzey eşitlik ve benzer dağılım özelliğine göre MTK gözlenen puan eşitlemenin, MTK gerçek puan eşitlemeden daha iyi performans gösterdiği belirlenmiştir. Birincil düzey eşitlik ölçütüne göre zincir eşitlemenin, frekans tahmininden daha iyi performans gösterdiği, ikincil düzey eşitlik ve benzer dağılım ölçütüne göre ise bahsedilen iki yöntemin benzer performans gösterdiği sonucuna ulaşılmıştır. Frekans tahmini ve zincir eşit yüzdelikli eşitleme için yapılan son düzgünleştirmenin ikincil düzey eşitlik ve benzer dağılım özelliklerine olumlu yansıdığı sonucuna ulaşılmıştır. Test formları Klasik Test Kuramına göre yapılandırıldığında ve klasik eşitleme yöntemlerine başvurulduğunda zincir eşit yüzdelikli eşitlemenin frekans tahmini yöntemine tercih edilmesi ve son düzgünleştirmenin kullanılması araştırma sonuçlarına ilişkin önerilerdir.

Liu ve Kolen (2011), karma formattaki testleri denk olmayan gruplarda ortak madde deseni kullanarak eşitlemiştir. Gerçek veriden oluşan iki test formunu (tarih

ve biyoloji) ikiye bölerek, ortak maddeler çoktan seçmeli olacak şekilde eşitleme işlemini gerçekleştirmişlerdir. Eşitleme yöntemi olarak MTK'ya dayalı gerçek puan (Stocking-Lord) ve gözlenen puan, KTK'ya dayalı frekans tahmini ve zincir eşit yüzdelikli eşitleme kullanılmıştır. Eşit yüzdelikli eşitleme yöntemlerine logaritmik doğrusal fonksiyon kullanılarak ön düzgünleştirme uygulanmış ve bu şekilde evren değerleri elde edilmiştir. Bu değerler yanlılığı hesaplamada kullanılmıştır. Eşit yüzdelikli eşitleme yöntemlerine ait sonuçlar son düzgünleştirmeyle elde edilmiştir. Sonuçlar değerlendirilirken eşitlemenin standart hatası (SEE), yanlılık ve hata kareleri ortalamasının karekökü (RMSE) kullanılmıştır. Sonuçlar tarih ve biyoloji testi için ayrı ayrı incelenmiştir. Eşitlemenin standart hatası incelendiğinde her iki test (tarih ve biyoloji) için de sırasıyla en düşük SEE değerini MTK gözlenen puan, MTK gerçek puan, frekans tahmini ve zincir eşit yüzdelikli eşitleme göstermiştir. Her iki test için de MTK gözlenen puan eşitleme, MTK gerçek puan eşitlemeden daha az yanlılık ve RMSE değeri göstermiştir. Tarih testinde MTK'ya dayalı yöntemler KTK'ya dayalı yöntemlerden daha az yanlılık ve RMSE değerine sahiptir. Tarih testinde düşük ve yüksek puan aralıklarında frekans tahmini yöntemi zincir eşit yüzdelikli eşitlemeden daha az yanlılık ve RMSE değerine sahiptir. Biyoloji testinde ise yöntemler arasında RMSE ve yanlılık değerleri için kesin bir çıkarım yapılmamıştır.

Lee, Lee ve Brennan (2012) araştırmasında denk olmayan gruplarda ortak madde desenini kullanarak 11 eşitleme yöntemini karşılaştırmıştır. Bu amaçla bir yerleştirme sınavına ait İngiliz, Fransız dillerine ait testler ve biyoloji testi kullanılmıştır. Araştırmada incelenen eşitleme yöntemleri Tucker, Levine gözlenen puan, Levine gerçek puan, düzgünleştirilmemiş zincir eşit yüzdelikli, logaritmik doğrusal fonksiyonla ön düzgünleştirilme uygulanmış zincir eşit yüzdelikli, kübik eğri yöntemiyle son düzgünleştirilme uygulanmış zincir eşit yüzdelikli, düzgünleştirilmemiş frekans tahmini, logaritmik doğrusal fonksiyonla ön düzgünleştirilme uygulanmış frekans tahmini, kübik eğri yöntemiyle son düzgünleştirilme uygulanmış frekans tahmini, MTK gerçek puan (Stocking-Lord) ve MTK gözlenen puandır. Bu yöntemlerden ilk üçü doğrusal, diğerleri eğrisel (curvilinear) modellerdir. Sonuçlar değerlendirilirken birincil ve ikincil düzey eşitlik kriteri kullanılmıştır. Araştırma sonucunda birincil düzey eşitlik açısından diğer yöntemlere göre en iyi performansa sahip yöntemin MTK gerçek puan eşitleme

olduğu bulunmuştur. Birincil düzey eşitlik açısından yöntemler doğrusal, eğrisel ve MTK'ya dayalı olmak üzere ayrışmıştır ve MTK'ya dayalı yöntemlerin KTK'ya dayalı yöntemlerden daha iyi performans gösterdiği belirlenmiştir. İkincil düzey eşitlik açısından MTK gözlenen puan eşitleme, MTK gerçek puan eşitlemeden daha iyi bir performans sergilemiştir. Ancak tüm yöntemler arasında çok büyük farklılıklar olmadığından yöntemlerden birisinin diğerinden daha üstün olduğunu söylemek mümkün değildir.

Wolf (2013) denk olmayan gruplarda ortak madde deseni kullanarak simülatif bir çalışma gerçekleştirmiştir. Araştırmasında test boyutu, yetenek dağılımı, ortak madde seti ve eşitleme yöntemlerinin karma testlerde gerçekleştirilen eşitleme işlemi üzerindeki etkisini incelemiştir. Eşitleme yöntemi olarak Klasik Test Kuramına dayalı frekans tahmini, zincir eşit yüzdelikli eşitleme; MTK'ya dayalı gerçek puan (eş zamanlı ölçekleme [concurrent calibration]) ve gözlenen puan eşitleme yöntemlerini kullanmıştır. Ayrıca klasik kurama dayalı test eşitleme yöntemlerine ön düzgünleştirme uygulanmıştır. Frekans tahmini ve zincir eşit yüzdelikli eşitleme sonuçları sadece ön düzgünleştirmeye dayalıdır. Frekans tahmini yöntemi için sentetik evren testi alan yeni gruba dayalı olarak tanımlanmıştır. Testler tek boyutlu olarak ve karmaşık çok boyutlu olarak iki etkenli model (bifactor model) altında incelenmiştir. Ortak madde setleri sadece çoktan seçmeli, çoktan seçmeli ve 1 yapılandırılmış cevap maddesi ile çoktan seçmeli ve 2 yapılandırılmış cevap maddesinden oluşacak şekilde belirlenmiştir. Yetenek dağılımları, 0 ortalamaya ve 1 standart sapmaya sahip olacak şekilde eşit; ilk grubun yetenek ortalaması 0, standart sapması 1, ikinci grubun yetenek ortalaması ,15 ve ,30, standart sapması 1 olacak şekilde belirlenmiştir. 50 maddelik testlerde 45 madde çoktan seçmeli, 5 madde yapılandırılmış cevap maddesidir ve 10 madde ortak maddedir. Örneklemler 3000'er kişiden oluşmaktadır. Sonuçlar değerlendirilirken birincil düzey eşitlik, ikincil düzey eşitlik ve benzer dağılım özelliği ölçüt olarak kullanılmıştır. Araştırma sonucunda tüm ölçütlere göre MTK'ya dayalı yöntemlerin KTK'ya dayalı yöntemlerden daha iyi performans gösterdiği belirlenmiştir. Birincil düzey eşitlik ölçütü dikkate alındığında en iyi performansı gösteren yöntemin MTK'ya dayalı gerçek puan eşitleme olduğu sırasıyla bu yöntemi MTK gözlenen puan eşitleme, zincir eşit yüzdelikli eşitleme ve frekans tahmini yönteminin takip ettiği bulunmuştur. Tek boyutlu yapılarda ikincil düzey eşitlik ölçütüne göre yöntemlerin benzer

performanslara sahip olduğu sonucuna ulaşılmıştır. Tek boyutlu testlerde benzer dağılım özelliği açısından inceleme yapıldığında denk gruplarda KTK'ya dayalı eşitleme yöntemlerinin benzer sonuçlara sahip olduğu belirlenmiştir. Tek boyutlu testlerde benzer dağılım özelliğine göre MTK'ya dayalı eşitleme yöntemlerinden gözlenen puan eşitlemenin, eş zamanlı kalibrasyondan daha iyi sonuçlar verdiği görülmüştür. Grupların yetenek dağılımları benzer olduğunda klasik test kuramına dayalı yöntemler birbirine benzer sonuçlar göstermektedir.

Otomatik puanlama ve test eşitleme ile ilgili araştırmalar. Alanyazın incelendiğinde otomatik puanlama ile test eşitlemenin birlikte ele alındığı kısıtlı çalışmaya rastlanılmıştır. İncelenen alanyazında ulaşılan iki araştırmaya aşağıda yer verilmektedir. Almond (2014) araştırmasında yalnızca açık uçlu maddeler üzerinde çalışırken Olgar (2015) çoktan seçmeli maddelerle birlikte yanıtı sınırlandırılmamış açık uçlu bir madde üzerinde çalışarak otomatik puanlama sonrası test eşitleme çalışması gerçekleştirmiştir. Aşağıda ilgili çalışmalarla ilgili detaylı bilgilere yer verilmektedir.

Almond (2014) yapılandırılmış cevap maddelerinden oluşan testlerde otomatik madde puanlama sistemleri tarafından puanlanan maddeleri ortak madde olarak kullandığı araştırmasında doğrusal lojistik eşitleme yöntemiyle bir test eşitleme çalışması gerçekleştirmiştir. Bu amacı gerçekleştirmek üzere GRE 2007 ve 2008 yazma becerileri testini kullanmıştır. 500 kişilik iki örneklem kullanılarak gerçekleştirdiği bu araştırmada 500 kişilik örneklemlerden ilkini özellikleri tespit etme amaçlı ikincisini ise puanlama amaçlı kullanmıştır. İki katılımcı grubu aldıkları özgün formların yanında bir de ortak formu yanıtlamıştır. Araştırmada generic e-rater yöntemi kullanılmıştır ve sonuçların bu doğrultuda değerlendirilmesi gerektiği belirtilmiştir. Araştırmada eşitleme sonuçlarının mantıklı olduğu sonucuna ulaşmıştır.

Olgar (2015) karma formattaki testleri eşitlediği araştırmada 30 çoktan seçmeli ve 1 yanıtı sınırlandırılmamış açık uçlu madde üzerinde çalışmıştır. Yanıtı sınırlandırılmamış açık uçlu madde generic e-rater otomatik puanlama sistemince puanlanmıştır. Araştırmada veri olarak öğretmenlere uygulanan sertifika testi kullanılmıştır. İncelenen test İngilizce yazma ve kullanma becerilerine ilişkindir. Araştırmada ortak madde olarak çoktan seçmeli ve çoktan seçmeli ile açık uçlu maddeler bir arada kullanılmıştır. Çoktan seçmeli maddeler için zincir ortalama

(chained mean), frekans tahmini (frequency estimation), zincir eşit yüzdelikli (chained equipercentile), zincir doğrusal (chained linear), Tucker ve Levine gözlenen puan eşitleme yöntemleri kullanılmıştır. Açık uçlu maddelerin eşitlenmesinde ise doğrusal lojistik (linear lojistic) eşitleme yöntemi kullanılmıştır. Sonuçta ortak madde olarak çoktan seçmeli maddeler ile otomatik puanlama sistemince puanlanan açık uçlu maddelerin kullanılması, ortak madde olarak sadece çoktan seçmeli maddelerin kullanıldığı durumla benzer sonuçlar göstermiştir.

Alanyazında yer alan araştırmalar bir bütün olarak ele alındığında test eşitleme açısından MTK'ya dayalı gerçek puan eşitleme yöntemlerinden sadece birkaçı üzerinde çalışıldığı ve MTK gerçek puan eşitleme yöntemlerinin karşılaştırılmadığı görülmektedir. Bu araştırmada MTK gerçek puan eşitleme yöntemlerinden dördü birbirleriyle karşılaştırılmakta bunun yanında KTK'ya dayalı yöntemlerin MTK'ya dayalı yöntemlerle karşılaştırması yapılmaktadır. Alanyazında her ne kadar KTK ve MTK eşitleme yöntemleri birbirleriyle karşılaştırılsa da doğrusal eşitleme yöntemlerine çok fazla yer verilmediği görülmektedir. Bu araştırmada ise doğrusal eşitleme yöntemlerine de yer verilmiştir. Alanyazında otomatik puanlama ile ilgili güvenirlik incelemeleri yapılmış ve yeterli bulunmuşsa da test eşitleme gibi durumlarda otomatik puanlamanın etkisi yeterince değerlendirilmemiştir. Bu araştırma hem önceki araştırmalara benzer şekilde otomatik puanlamanın güvenirliğini ele almakta hem de otomatik puanlamanın test eşitleme sürecinde kullanılmasının etkilerini incelemektedir. Alanyazında otomatik puanlama ile gerçekleştirilen test eşitleme çalışmalarına rastlansa da bu araştırmalarda ya tüm maddelerin yapılandırılmış cevap maddesi olduğu testlerde çalışılmış ya da tek bir yapılandırılmış cevap maddesiyle karma testler üzerinde çalışılmıştır. Bu araştırmada ise çok sayıda yapılandırılmış cevap maddesinden oluşan karma testler kullanılmıştır. Ayrıca alanyazında otomatik puanlama aracılığıyla MTK test eşitleme yöntemleri kullanılarak yapılan bir eşitleme çalışmasına rastlanmamıştır. Bu araştırma ise hem KTK hem MTK yöntemlerini kullanarak otomatik puanlama sonrası eşitleme işlemi gerçekleştirmektedir.

Bölüm 3
Yöntem

Araştırma Modeli

Araştırma, karma formattaki testlerde otomatik açık uçlu madde puanlamanın güvenirliğini ve test eşitleme üzerindeki etkisini gerçek puanlayıcılarla gerçekleştirilen test eşitleme uygulamasıyla karşılaştırarak belirlediğinden ilişkiseldir. Frankel, Wallen ve Hyun (2015) ilişkisel araştırmalara değişkenler arasındaki ilişki olasılığını değerlendirmek için başvurulduğunu, Creswell (2012) ise ilişkisel araştırmalarla bir değişkendeki farklılığın diğer değişkeni nasıl etkilediğini görmenin mümkün olduğunu belirtmektedir. Bu nedenle mevcut araştırmanın ilişkisel olduğu belirtilebilir.

Araştırmanın Veri Kaynağı

Araştırmanın veri kaynağını Türkiye'de 2016 yılında uygulanan Millî Eğitim Bakanlığı (MEB) tarafından uygulanan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) 8. sınıflar araştırması oluşturmaktadır. Bu araştırma Türkiye'nin her ilinden 400 öğrencinin katılımı ile gerçekleştirilmiştir. 2016 yılı ABİDE 8. sınıflar araştırması kapsamındaki 12 test formundan her birini yaklaşık 3000 öğrenci yanıtlamıştır (MEB, 2017a).

Hâlihazırdaki bu araştırmada 2016 yılı ABİDE 8. sınıf uygulamasına katılan A_1 ve B_1 kitapçıklarını yanıtlayan öğrencilerden MEB tarafından kullanımına izin verilen ve seçkisiz olarak belirlenen 1000'er öğrenciden amaca uygun verilerin seçilmesiyle ve verinin temizlenmesiyle A_1 kitapçığından 607 ve B_1 kitapçığından 584 kişiye ait veri kullanılmıştır. Spence (1996) test eşitleme çalışmaları için her bir test formunu en az 500 bireyin yanıtlaması gerektiğini belirtmektedir. Araştırmada kullanılan veri bu kriteri sağlar niteliktedir.

Verilerin Elde Edilmesi

Araştırmada ikincil veri kullanılmıştır. Öncelikle MEB'den 2016 yılı ABİDE 8. sınıf Türkçe testi A_1 ve B_1 kitapçığı verileri talep edilmiştir. EK-T'de görülebileceği gibi MEB'den gerekli izin alınmıştır. Bu doğrultuda iki farklı puanlayıcı grubuna ve nihai puanlara ait puan matrisleri elde edilmiştir. Bunun yanı sıra JPEG formatındaki öğrenci cevap kâğıtlarına erişilerek bilgisayar ortamına girişi yapılmıştır. Verilerin bilgisayara giriş işlemi elle yapılmıştır. Bu durumun nedeni öğrenci yazılarının

okunmasının zor olması ve bitişik el yazısı kullanımı nedeniyle optik karakter tanıma sistemlerinden (OCR) destek alınamaması bunun yanında OCR programlarından kaynaklanacak hataları bertaraf etmektir. Elle girilen verilerin öğrenci cevaplarıyla tamamen eşleşmesi için veriler kontrol edilerek hatalar giderilmiştir. Öğrenci cevapları doğrudan aktarılmış olup herhangi bir düzeltmeye tabi tutulmamıştır.

Veri Özellikleri

Araştırmada 2016 yılında MEB tarafından yürütülen Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) 8. sınıf araştırması kapsamındaki 12 test kitapçığından A_1 ve B_1 kitapçığında yer alan iki Türkçe test formu verisi kullanılmıştır. Bu test formları ve puanlanması ile ilgili detaylı bilgilere bunun yanı sıra 2016 yılı ABİDE 8. sınıf uygulamasında yer alan testlerle ilgili genel bilgilere aşağıda yer verilmektedir.

2016 yılı ABİDE 8. sınıf araştırmasında yer alan testler öğrencilerin öğrendiklerini gerçek yaşam durumlarında kullanabilme becerisine odaklanmaktadır. Farklı madde türleri kullanılarak öğrencilerin üst düzey düşünme becerilerinin incelenmesinin amaçlandığı testlerde çoktan seçmeli ve açık uçlu maddeler birlikte yer almaktadır. Bu araştırmada kullanılacak testlerin pilot uygulaması 2015 yılında ve nihai uygulaması 2016 yılında gerçekleştirilmiştir. 2016 yılı ABİDE 8. sınıf araştırması Türkçe, Matematik, Fen Bilimleri ve Sosyal Bilgiler olmak üzere 4 dersi kapsamaktadır. ABİDE araştırmasının madde yazarlarına alanda yetkin akademisyenler tarafından “Açık Uçlu Soru Yazma” eğitimi verilmiş olup ölçülecek beceriler belirlenmiş ve çoktan seçmeli, açık uçlu maddelerin yazımı sağlanmıştır. Maddeler ölçme ve değerlendirme uzmanları ile dil uzmanları tarafından incelenmiş ve redaksiyona tabi tutulmuştur. Açık uçlu maddelerin puanlama işlemini sağlamak üzere bir yazılım oluşturulmuştur. Bu yazılım her bir maddeye verilen cevapların iki gerçek puanlayıcıya ulaşmasını sağlamaktadır. Puanlar arasında uyum olmadığında ise cevabın bir üst puanlayıcıya gönderilmesine imkân tanımaktadır. Üst puanlayıcılar maddeleri ve puanlama anahtarlarını hazırlayan kişilerden oluşmaktadır. Projenin pilot uygulaması 5000 öğrenci ve 160 puanlayıcı ile gerçekleştirilmiştir (MEB, 2017a; MEB, 2017b).

2016 yılı ABİDE 8. sınıf esas uygulamasında her bir ders için 51 madde belirlenmiştir. Bu maddelerden 27 tanesinin nihai uygulama için, 24 tanesinin ise bir

sonraki uygulamalar için pilot olarak kullanılması planlanmıştır. Nihai uygulamada her bir ders için 20 maddeden oluşan testler kullanılmıştır. Bu testlerde yer alan 2 madde pilot maddedir. Testlerin uygulanmasında çok sayıda kitapçık kullanılmasına rağmen temelde üç test formu bulunmaktadır (A, B ve C formları). Temel test formları altında oluşan test formlarında (örneğin A_1, A_2, A_3 ve A_4) bulunan maddeler pilot maddeler açısından farklılık göstermektedir. Açık uçlu maddelerin puanlanması pilot uygulamada bahsedilen şekilde gerçekleştirilmiştir. 1800 öğretmen öğrencilerin cevaplarını puanlamıştır. Tüm illerden temsilci birer öğretmen ve MEB bünyesindeki madde yazarlarından oluşan 32 kişilik ekip üst puanlayıcı olarak belirlenmiştir. Puanlama yapılırken puanlayıcılar dereceli puanlama anahtarlarından yararlanmıştır. Dereceli puanlama anahtarı oluşturulurken öncelikle yargısal olarak doğru ve kısmi doğru yanıtlar (eğer madde üç kategorili ise) belirlenmiştir. Uygulama sonrası ise beklenmeyen doğru ve kısmi doğru yanıtlar puanlama anahtarına dâhil edilmiştir. Dereceli puanlama anahtarları; doğru yanıt, kısmi doğru yanıt (eğer madde üç kategorili ise), yanlış yanıt ve boş yanıt bölümlerini içermektedir. Bu bölümler ilgili maddenin kabul edilecek, kısmen kabul edilecek ve kabul edilmeyecek cevaplarını belirtmekte bunun yanında örnek cevaplara yer verilmektedir (MEB, 2017a; MEB, 2017b). ABİDE araştırmasında testlerde yer alan her bir madde için puanlayıcılar arası tutarlılık Cramer's V katsayısı ile hesaplanmıştır. Bu araştırmanın konusu olan Türkçe testlerindeki tüm maddelere ilişkin puanlayıcılar arası tutarlık katsayıları (Cramer's V) tablo 2'de gösterilmektedir.

Tablo 2 incelendiğinde ABİDE 2016 yılı 8. sınıf Türkçe testlerinde yer alan açık uçlu maddeler için hesaplanan Cramer's V katsayılarının ,83 ile ,99 arasında değiştiği ve ortalamasının ,93 olduğu belirtilebilir. Araştırmada kullanılan A_1 test kitapçığı için Cramer's V katsayılarının ,83 ile ,98 arasında değiştiği ve A_1 kitapçığı için Cramer's V katsayılarının ortalamasının ,94 olduğu, B_1 test kitapçığı için ise Cramer's V katsayılarının ,87 ile ,99 arasında değiştiği ve B_1 kitapçığı için Cramer's V katsayılarının ortalamasının ,93 olduğu görülmektedir. Test puanlarına ilişkin güvenirlik katsayıları A kitapçığı için ,73, B kitapçığı için ise ,76 olarak belirtilebilir (MEB, 2017a).

Tablo 2

ABİDE 2016 8. Sınıf Türkçe Teslerinde Yer Alan Yapılandırılmış Cevap Maddelerine İlişkin Cramer V Katsayıları

Madde Kodu	A_1 Kitapçığı Madde Numarası	B_1 Kitapçığı Madde Numarası	Cramer's V
T-2016-0002	Madde 2	-	,83
T-2016-0007*	Madde 7	Madde 5	,92
T-2016-0008*	Madde 8	Madde 6	,98
T-2016-0010*	Madde 10	Madde 8	,90
T-2016-0011*	Madde 11	Madde 9	,97
T-2016-0014	Madde 14	-	,92
T-2016-0015*	Madde 15	Madde 17	,94
T-2016-0018	Madde 18	-	,98
T-2016-0019	Madde 19	-	,98
T-2016-0029	-	Madde 3	,87
T-2016-0031	-	Madde 11	,93
T-2016-0032	-	Madde 12	,91
T-2016-0035	-	Madde 18	,99
T-2016-0037	-	Madde 20	,93
Ortalama	,94	,93	,93

* A_1 ve B_1 kitapçıklarında yer alan ortak maddeleri göstermektedir.

Araştırma kapsamında ABİDE 2016 nihai uygulamasında elde edilen Türkçe testi verileri kullanılmıştır. EK-A'da ABİDE 2016 8. sınıf Türkçe testinde yer alan ve MEB tarafından açıklanan örnek maddeler ve dereceli puanlama anahtarları gösterilmektedir. 18 maddelik 2 Türkçe test formundan (A_1 ve B_1) elde edilen veriler araştırmanın odak noktasıdır. A_1 testinde yer alan dokuz madde ve B_1 testinde yer alan 10 madde açık uçludur. Açık uçlu maddeler 0-1 ve 0-1-2 şeklinde puanlanmaktadır. A_1 ve B_1 testlerinde dokuz madde ise ortaktır (MEB, 2017a).

Araştırma denk olmayan gruplar için ortak madde deseni kullanılarak iki test formunu eşitlemeyi amaçlamaktadır. Fakat eşitleme işlemi öncesi bazı kriterlerin sağlanması gerekmektedir. Angoff (1984) ortak madde sayısının testin uzunluğu ne kadar artarsa artsın testteki maddelerin sayısının %20'sinden az olmaması gerektiğini belirtmiştir. Bu uygulamada ise ortak madde oranı %50'dir. Verinin özellikleri dikkate alındığında ortak maddelerde ikili ve çoklu puanlanan madde türlerinin birlikte kullanılması gerekmektedir. Nitekim Tate (2000) karma formattaki

testlerde ortak madde olarak her iki madde türünün birlikte kullanımını önermektedir. Buna gerekçe olarak ortak maddelerin testin tamamını temsil etmesi gerekliliğini göstermektedir. A_1 ve B_1 kitapçıklarında dokuz ortak maddeden beşinin açık uçlu, dördünün ise çoktan seçmeli olduğu belirtilebilir.

Verilerin Analizi

Aşağıda verilerin analize hazırlanması ve analiz sürecinden detaylı olarak bahsedilmektedir. Öncelikle verinin otomatik puanlamaya hazırlanma sürecinden bahsedilmiş ardından sırasıyla otomatik puanlama yazılımına, otomatik puanlama işleminin gerçekleştirilmesine, otomatik puanlama işlemine yönelik hesaplanan uyum katsayılarına, KTK ve MTK'ya dayalı eşitleme işlemine ve eşitleme hatalarının hesaplanmasına ilişkin bilgilere yer verilmiştir.

Verilerin hazırlanması. Araştırma verileri analiz edilmeden önce MEB'den alınan 1000 öğrenciye ait veri incelenmiştir. İncelemeler sonucunda yapılandırılmış cevap maddelerinden alınan puanlara ilişkin kategorilere dengeli dağılım temel alınarak veri girişi yapılmıştır. Bu işlem veride yapılandırılmış cevap maddelerine ilişkin yaygınlık (kategorilere dağılımda dengesizlik [prevalence]) probleminin mümkün olduğunca önüne geçmek için gerçekleştirilmiştir. A_1 kitapçığı için 9, B_1 kitapçığı için 10 maddenin tamamı dikkate alınarak A_1 kitapçığından 697 ve B_1 kitapçığından 701 veri girişi yapılmıştır. Ardından testteki yapılandırılmış cevap maddelerinin yarısına veya yarısından fazlasına cevap veren öğrenciler seçilmiştir. Bu işlem sonrasında yapılandırılmış cevap maddelerinin her birisi için kayıp veri oranı hesaplanmıştır. Kayıp veri oranı %5'in altında kalacak şekilde veri temizlenmiştir. Bu işlem otomatik puanlamada uyum katsayılarının normalden yüksek çıkmasının önlenmesi amacıyla gerçekleştirilmiştir. Veriler temizlenirken kategorilere dağılım dikkate alınmıştır. Bazı kategorilerde az sayıda veri bulunduğu için mümkün olduğunca bu kategorilerde puan alan bireylerin araştırma kapsamından çıkarılmamasına dikkat edilmiştir. Kayıp veri, araştırmacı tarafından belirlenen maddelerin yarısı veya yarısından fazlasına cevap verme kriterleri dikkate alınarak A_1 kitapçığından 84 ve B_1 kitapçığından ise 96 kişiye ait veri temizlenmiştir. Ardından gerçek puanlayıcı grubu 1 ve gerçek puanlayıcı grubu 2'nin öğrencilere verdiği puanlar incelenmiştir. Burada görülen kayıp puanlar nedeniyle bir grup öğrenci de araştırma kapsamı dışına alınmıştır. Bu yönde A_1 kitapçığından ve B_1 kitapçığından toplamda 6 kişi çıkarılmıştır. Son olarak çoktan seçmeli maddelerdeki

kayıp veri sayısı değerlendirilmiş testteki toplam madde sayısının yarısından ve çoktan seçmeli maddelerin yarısından fazlasına cevap vermeyen öğrenciler araştırma kapsamından çıkarılmış ve kayıp veri oranının %5'in altında kalması sağlanmıştır. Bu yönde A_1 kitapçığından veri çıkarılmamış olup B_1 kitapçığından 15 kişi araştırma kapsamı dışında bırakılmıştır. Son durumda A_1 kitapçığından 90 kişi, B_1 kitapçığından ise 117 kişi çıkarılmıştır. Böylece veri hazırlık süreci tamamlanarak A_1 kitapçığından 607, B_1 kitapçığından 584 veri ile otomatik puanlama işlemine geçilmiştir. Aşağıda otomatik puanlama için hazırlanan yazılıma ilişkin bilgilere yer verilmiştir.

Otomatik puanlama yazılımının oluşturulması. Araştırmada, araştırmacının da içerisinde bulunduğu bir ekip tarafından geliştirilen otomatik puanlama yazılımı kullanılmıştır. Yazılım geliştirilirken MEB tarafından uygulanan "Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme Projesi"ne ait Türkçe testi yapılandırılmış cevap maddeleri kullanılmıştır. Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme Projesi Türkçe testi bu araştırmada kullanılan testlerden (ABİDE 2016 Araştırması) bağımsızdır. Bu test beşinci sınıf öğrencilerine yönelik olup 5 yapılandırılmış cevap maddesi, 15 çoktan seçmeli madde içermektedir. Yazılım hazırlanırken bahsedilen 5 açık uçlu maddeden yararlanılmıştır. 5 açık uçlu maddeden üçü 0-1 şeklinde puanlanırken, ikisi 0-1-2 şeklinde puanlanmaktadır. Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme projesi kapsamında her bir öğrenciye ait cevaplar iki puanlayıcı tarafından puanlanmış ve gerektiğinde üst puanlayıcıya ulaşılarak nihai bir puan elde edilmiştir. Puanlama işlemlerinde dereceli puanlama anahtarlarından yararlanılmıştır. Şekil 4'te ikili puanlanan örnek bir maddeye yer verilmektedir.

Şekil 4'de "Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme" projesi Türkçe testi 16. maddesi görülmektedir. Öğretmenlerden oluşan puanlayıcı grupları madde 16'ya ilişkin yanıtları puanlarken tablo 3'te yer alan dereceli puanlama anahtarından yararlanmıştır. Tablo 3, madde 16'ya ilişkin dereceli puanlama anahtarı yanında örnek yanıtları da göstermektedir.

GÜZEL ATLAR ÜLKESİ: KAPADOKYA

Kapadokya neresidir? Bir şehir, bir ülke yoksa bir bölge midir? Neden her yıl binlerce insan orayı ziyaret eder, yüzlerce kilometre öteden görmeye gelir, dağları geçer, denizleri aşar? Peki, Kapadokya'da ilk önce nereyi ziyaret etmek gerekir? Ne güzel sorular bunlar değil mi! İnsan, öğrenmeye merak etmekle başlar. Sorular sorar, araştırır, bulur, öğrenir. Öğrendikçe de daha bilgili, daha cesur, daha güvenli olur.

Kapadokya, Anadolu ya da Mezopotamya gibi bir bölgenin adı. Nevşehir ilinin sınırları içinde, çok geniş bir alan. 25.000 kilometrekare. Yalnız, oldukça ilginç bir bölge. Bu sebeple binlerce insan her yıl oraya geliyor. Öyle bir bölge ki tarihi "Yontma Taş Devri"ne kadar uzanıyor. Sırasıyla Hititler, Persler, Bizanslılar, Selçuklular ve Osmanlılar yaşamış Kapadokya'da.

Birinci paragraftaki soruların hangisinin cevabı ikinci paragrafta yoktur?

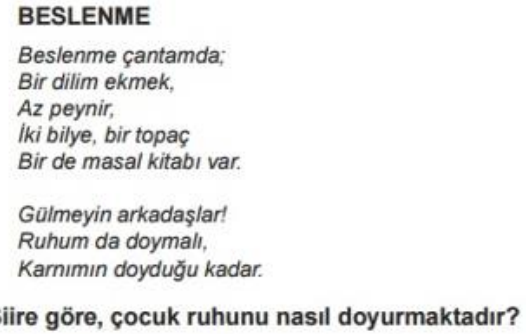
Şekil 4. Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme projesi Türkçe testi madde 16

Tablo 3

Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme Projesi Türkçe testi Madde 16'ya İlişkin Dereceli Puanlama Anahtarı ve Örnek Yanıtlar

Madde No	16
Bağlam Adı	Güzel Atlar Ülkesi: Kapadokya
Doğru Yanıt (1 Puan) Açıklama	"Kapadokya'da ilk önce nereyi ziyaret etmek gerekir?" sorusuna atıfta bulunan cevaplar doğru cevap olarak kabul edilecektir.
Yanlış Yanıt (0 Puan) Açıklama	Boş cevap ve "Kapadokya'da ilk önce nereyi ziyaret etmek gerekir?" sorusuna atıfta bulunan cevapların haricindeki tüm cevaplar yanlış olarak kabul edilecektir.
Örnek Doğru Yanıtlar	- Peki Kapadoyada en önce nereyi ziyaret etmek gerekir - Kapadokya'da ilk önce nereyi ziyaret etmek gerekir? sorusunun cevabı yoktu? - Kapadokyayı ziyarete gelen ilk önce nereye gider?
Örnek Yanlış Yanıtlar	- Kapadokya neresidir? Sorusunun cevabı yok - NEDEN Binlerce insan orayı ziyaret eder? Peki Kapa dokyada ilk nereyi ziyaret etmek gerekir? - Bir şehirmi yoksa bir ülkemidir

Üçlü puanlanan örnek bir maddeye şekil 5'de yer verilmektedir.



BESLENME

Beslenme çantamda;
Bir dilim ekmek,
Az peynir,
İki bilye, bir topaç
Bir de masal kitabı var.

Gülmeyin arkadaşlar!
Ruhum da doymalı,
Karnımın doyduğu kadar.

Şiire göre, çocuk ruhunu nasıl doyurmaktadır?

Şekil 5. Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme projesi Türkçe testi madde 20

Şekil 5'de "Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme" projesi Türkçe testi 20. maddesi görülmektedir. Öğretmenlerden oluşan puanlayıcı grupları madde 20'ye ilişkin yanıtları puanlarken tablo 4'te yer alan dereceli puanlama anahtarından yararlanmıştır. Tablo 4, madde 20'ye ilişkin dereceli puanlama anahtarı yanında örnek yanıtları da göstermektedir.

Tablo 4

Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme Projesi Türkçe testi Madde 20'ye İlişkin Dereceli Puanlama Anahtarı ve Örnek Yanıtlar

Madde No	20
Bağlam Adı	Beslenme
Doğru Yanıt (2 Puan) Açıklama	Çocuğun ruhunu; oyun oynayarak ve kitap okuyarak doyurduğunu ifade eden tüm cevaplar doğru kabul edilir.
Kısmi Doğru Yanıt (1 Puan) Açıklama	Oyun oynar ve kitap okur ifadelerinden sadece birini içeren cevaplar kısmi cevap olarak kabul edilir.
Yanlış Yanıt (0 Puan) Açıklama	Yanlış, ilgisiz ve metinden aynen alınan ifadeler.
Örnek Doğru Yanıtlar	- İki bilyeyi ve bir tane topacı oynayıp, bir masal kitabı okuyarak doyurmaktadır. - 1 bilye bir topaç birde masal kitab okuyup oyunayı Ruhudoyar - Beslenerek, eğlenerek ve okuyarak. - okuyarak ruhunu doyurma isteğiyle
Örnek Kısmi Doğru Yanıtlar	- eğlenerek doyuruyo - Kitap okuyarak, kendini kitabın içine koyarak, ruhunu geliştirip, hissederek. -. iki bilye bir topaç birde masal Kitabi ruhunu doyurmuştur
Örnek Yanlış Yanıtlar	- bir dilim ekmek ,az peynir, iki bilye, bir topaç birde masal kitapı var. - Çocuk ruhunu masal kitabıyla doyurur.

Otomatik puanlama yazılımı hazırlanırken “Ölçme ve Değerlendirme Uygulamalarını İzleme, Araştırma ve Geliştirme Projesi” Türkçe testine ilişkin 303 ve 637 öğrenciye ait veriden yararlanılmıştır. Madde 16 için 303 veri ile deneme yapılırken, madde 20 için 637 veri ile deneme yapılmıştır. Madde 20 üç kategorili puanlandığı için daha fazla veri üzerinde deneme yapılması uygun bulunmuştur. Linux işletim sistemi üzerinde Python programı kullanılarak otomatik puanlama sistemi oluşturulmuş ve denemeler yapılmıştır. Denetlenen (supervised) ve denetlenmeyen (unsupervised) makine öğrenme yöntemleri ile yapılan denemeler sonucunda denetlenen makine öğrenme yönteminin özellikle cevaplar özgünleştikçe daha uygun olduğu bulunmuştur. Bu nedenle çalışmaya denetlenen makine öğrenme yöntemleriyle devam edilmiştir. Gerçek puanlayıcılar aracılığıyla puanlama özellikleri bilgisayara haritalandırılarak otomatik puanlama yapılması sağlanmıştır. Otomatik puanlamada SVM, LR, MNB, LSTM ve BLSTM olmak üzere beş yöntem kullanılmıştır. Python aracılığıyla hazırlanan yazılımda iki Türkçe kütüphanesinden yararlanılmıştır. Verinin %90’ı sistemi eğitmek %10’u ise sistemi test etmek amacıyla kullanılmıştır. Çapraz geçerlik ile rastgele örnekleme yöntemine başvurulmuştur. 10 kat çapraz geçerleme ile test verileri 10 kez birbirinden farklı olacak şekilde değiştirilerek veri sayısı kadar otomatik puanlama yapılmış bu puanlar üzerinden uyum yüzdeleri hesaplanarak sonuçlar değerlendirilmiştir. Böylece 303 veri üzerinde yapılan denemede 303 puanlanma sonucuna, 637 veri üzerinde yapılan denemede 637 puanlama sonucuna ulaşılmıştır. Tablo 5’te ikili (0-1) puanlanan madde 16 ve üçlü (0-1-2) puanlanan madde 20’ye ait örnek sonuçlara yer verilmektedir.

Tablo 5

Otomatik Puanlama ile Gerçek Puanlayıcılar Arasındaki Uyum Yüzdeleri

	Veri Sayısı	Kategori Sayısı	SVM(%)	LR(%)	MNB(%)	LSTM(%)	BLSTM(%)
Madde 16	303	2	98,0	98,3	96,1	99,0	99,0
Madde 20	637	3	85,5	82,4	75,1	87,3	88,7

Not: Uyum yüzdelerinin %80’in üzerinde olması kabul edilebilir bir uyumu göstermektedir.

Tablo 5 incelendiğinde madde 16 için elde edilen uyum yüzdelerinin oldukça yüksek olduğu görülmektedir. Bu madde için en yüksek uyum yüzdesini gösteren

yöntemler LSTM ve BLSTM olarak bulunmuştur. Madde 20 için elde edilen uyum yüzdelerinin yeterli düzeyde olduğu sonucuna ulaşılmıştır. Madde 20'de en iyi uyumu gösteren yöntemin BLSTM yöntemi olduğu görülmektedir. Tüm yöntemler için elde edilen uyum yüzdelerinin beklenen düzeyde olması ile oluşturulan sistemin yapılandırılmış cevap maddelerini puanlamada yeterli olacağına kanaat getirilmiştir. Bu yönde hâlihazırdaki araştırma için otomatik puanlama işlemine geçilmiştir. Detaylar aşağıda belirtilmektedir.

ABİDE verilerinin otomatik puanlanması. Otomatik puanlama aşamasında gerçek puanlayıcı grubu 1, gerçek puanlayıcı grubu 2 ve üst puanlayıcılar (gerektiğinde) aracılığıyla belirlenen nihai puanlardan bir kısmı kullanılarak otomatik puanlama sistemi eğitilmiştir. Bu şekilde otomatik puanlama sisteminin gerçek puanlayıcılardan nasıl puanlama yapıldığını öğrenmesi sağlanmış ve sisteme puanlama özellikleri haritalandırılmıştır. Ardından sistemin eğitilmesinde kullanılmayan veriler otomatik olarak puanlanmıştır. Sistemin test edilmesinde kullanılan veri sayısı araştırmada incelenen bir faktördür. Test için kullanılan veri oranları %10, %20 ve %33 olarak belirlenmiştir. Dolayısıyla sistemin eğitilmesinde kullanılan veri sayısı sırasıyla %90, %80 ve %67'dir. Bu değerler A_1 kitapçığı için 607 verinin sırasıyla 61, 121 ve 200'ünün sistemi test etmek amacıyla kullanıldığını; sırasıyla 546, 486 ve 407'sinin ise sistemi eğitmek amacıyla kullanıldığını göstermektedir. B_1 kitapçığı için 584 verinin sırasıyla 58, 117 ve 193'ünün sistemi test etmek amacıyla kullanıldığını; sırasıyla 526, 467 ve 391'inin ise sistemi eğitmek amacıyla kullanıldığını göstermektedir. Araştırmada eğitim için kullanılacak veri sayısı mümkün olduğunca azaltılarak bu durumun etkisi incelenmeye çalışılmıştır. Sonuçlar hesaplanırken %10 test veri oranı için 10 kat, %20 test veri oranı için 5 kat ve %33 test veri oranı için 3 kat çapraz geçerlik kullanılmıştır. Bu şekilde eğitim ve test verileri farklılaştırılarak A_1 kitapçığı için 607 verinin tümü, B_1 kitapçığı için 584 verinin tümü test verisi haline getirilmiştir. Sonuçta sistem tarafından A_1 kitapçığı için puanlanmış 607 veri, B_1 kitapçığı için puanlanmış 584 veri elde edilmiştir. Son adımda sistemce oluşturulmuş puanların gerçek puanlayıcıların üzerinde anlaştıkları puanlarla olan uyumu hesaplanmıştır. Karşılaştırma imkânı sunması açısından gerçek puanlayıcı grubu 1 ve gerçek puanlayıcı grubu 2'nin nihai puanlarla olan uyumu da hesaplanmıştır. Bu işlemler her bir madde için

gerçekleştirilmiştir. Aşağıda uyum katsayılarına ilişkin detaylı bilgilere yer verilmektedir.

Uyum katsayıları. Puanlayıcılar arası uyum incelenirken uyum yüzdesi (percentage of agreement), otomatik puanlama araştırmalarında sıklıkla kullanılan karesel ağırlıklı Kappa (quadratic weighted Kappa [QWK]) katsayısı ve verideki yaygınlık (prevalence) sorunundan etkilenmeyen Gwet'in AC1 (Gwet's AC1) katsayısından yararlanılmıştır. Aşağıda bu katsayılara ilişkin detaylı bilgilere yer verilmektedir.

Uyum yüzdesi (percentage of agreement). Uyum yüzdesi basit ve hızlı bir şekilde hesaplanabilen anlaşılması ve yorumlanması kolay bir katsayıdır. Bu yöntemde katılımcıların birinci ve ikinci puanlayıcıdan aldıkları puan dizileri karşılaştırılmakta, puanlayıcıların üzerinde tam olarak anlaştıkları derecelendirme sayısının tüm derecelendirmelerin sayısına oranı hesaplanmakta ve sonuç yüzde cinsinden ifade edilmektedir. Elde edilen sonuçlar %0 ile %100 aralığında değişmektedir. Bu katsayı şans eseri oluşabilecek anlaşmaları hesaba katmadığından eleştirilmektedir. Çünkü bu durum uyumun olduğundan fazla bulunmasına yol açabilmektedir. Tüm ölçek düzeyleri (sınıflama, sıralama, eşit aralıklı ve oran) için kullanılabilecek bu yöntem, aynı puanlayıcının birden fazla puanlama yaptığı durumlarda ve ikiden fazla puanlayıcı bulunduğunda da kullanılabilmektedir. Puan kategorisi sayısı iki ya da daha fazla olduğunda da kullanılabilen bu yöntemde puan kategorisi sayısı arttıkça hesaplama yapmak zorlaşmaktadır. Ek olarak bu yöntem bir ya da iki ve bir ya da daha fazla derecedeki anlaşmazlıklar arasındaki farkı belirleyememektedir (Araujo ve Born, 1985; Goodwin, 2001; Graham, Milanowski ve Miller, 2012; Meyer, 1999). Araştırmacılar kesin bir kural olmamakla beraber uyum yüzdesinin %80'in üzerinde olması gerektiği konusunda görüş birliğine sahiptir (Hartmann, 1977).

Karesel ağırlıklı Kappa (quadratic weighted Kappa). Kappa katsayısı en sık kullanılan uyum katsayılarından birisidir. Kappa katsayısı puanlayıcılar arasındaki şans eseri anlaşma olasılığını dikkate alan bir uyum katsayısıdır. Fakat Kappa katsayısı puanlayıcıların anlaşmama olasılığını dikkate almamaktadır. Bu nedenle Kappa katsayısı ağırlıklandırılma yoluna gidilmiştir. Kappa katsayısı ağırlıklandırılırken uyuşmazlığın derecesine göre ağırlıklar kullanılmaktadır. En sık kullanılan iki ağırlıklandırma tekniği doğrusal (linear) ve kareseldir (quadratic).

Doğrusal ağırlıklandırmada ağırlıklar puanların standart sapması ile orantılı iken karesel ağırlıklandırmada ağırlıklar puanların standart sapmasının karesi ile orantılıdır. Yorumlanması kolay olduğundan uygulamada karesel ağırlıklı Kappa kullanımı oldukça fazladır. İki puan kategorisi bulunduğunda kullanılabilen bu katsayı ikiden fazla puan kategorisi bulunduğunda da kullanılabilmektedir. Puan kategorisi arttıkça Kappa değerini hesaplamak ve yorumlamak zorlaşmaktadır. Ek olarak puanlardan birisi diğerinden veya diğerlerinden çok daha fazla sayıda ise bu katsayı yanıltıcı bir şekilde düşük olabilmektedir. Bu durum alanyazında yaygınlık (prevalence) sorunu olarak tanımlanmaktadır ve Kappa katsayısıyla ilgili en çok raporlanan sorundur. Yaygınlığın yanı sıra yanlılık (bias) ve puanlamadaki bağımlılık (nonindependence of ratings) Kappa değeri üzerinde etkilidir.

Karesel ağırlıklı Kappa uyum katsayısı iki puanlayıcıya ait puanlar arasındaki uyumu değerlendirmede kullanılabileceği gibi otomatik puanlama sistemi puanları ile üzerinde karara varılmış gerçek puanlayıcı puanları arasındaki uyumu değerlendirmekte de kullanılabilir ve 0 ile 1 aralığında değişen değerler alır. 0 katsayısı puanlayıcılar arasında uyum olmadığını gösterirken, 1 katsayısı puanlayıcılar arasındaki tam uyumu göstermektedir. Değerlendiriciler arasında şans eseri ortaya çıkacak değerden az uyuma rastlandığında bu değer 0'ın altında düşebilmektedir. Kategori sayısı arttıkça karesel ağırlıklı Kappa değeri de artmaktadır (Altman, 1991; Brenner ve Kliebsch, 1996; Graham, Milanowski ve Miller, 2012; Preston ve Goodman, 2012; Sim ve Wright, 2005; Vanbelle, 2016). Landis ve Koch (1977) Kappa katsayısının yorumlanması için bir ölçüt belirtmiştir. Altman (1991) ise bu ölçütün uyarlamasını yapmıştır. Tablo 6, bu iki çalışmayı dikkate alarak oluşturulan Kappa aralıklarını ve onlara karşılık gelen uyum gücü ifadelerini göstermektedir. Altman (1991) ağırlıklandırılmış Kappa değerlerinin ağırlıklandırılmamış Kappa değerlerinden yüksek olacağını belirtmekte ve Williamson, Xi ve Breyer (2012) gerçek puanlayıcılar ve otomatik puanlama sistemleri arasındaki uyumun ,70'in üzerinde olmasını önermektedir.

Tablo 6

Kappa Katsayısı Ölçütü

Kappa Değeri	Uyum Gücü
<0,20	Zayıf
0,21-0,40	Kayda değer
0,41-0,60	Orta
0,61-0,80	İyi
0,81-1,00	Çok iyi

Karesel ağırlıklı Kappa değeri hesaplanırken Wang, Wei, Zhou ve Huang (2018) ile Preston ve Goodman (2012) tarafından kullanılan denklemlerden yararlanılmıştır. Aşağıda hesaplamalara ilişkin bilgilere yer verilmektedir. Karesel ağırlıklı Kappa değerini hesaplamak için ilk olarak açık uçlu madde derecelendirmelerinden oluşan NxN matrisi oluşturulmaktadır. Bu matris O ile gösterilmektedir. N ise muhtemel cevap sayısını göstermektedir. Bu doğrultuda puanlayıcıların derecelendirmeleri arasındaki farka dayalı W ağırlık matrisi hesaplanmaktadır. Eşitlik 26, W ağırlık matrisini oluşturmak için kullanılan formülü göstermektedir.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \qquad (26)$$

Bu denklemde yer alan i gerçek puanlayıcılar tarafından belirlenen derecelendirmeyi, j ise otomatik puanlama sisteminin derecelendirmesini göstermektedir. Ardından derecelendirilen puanlar arasında hiçbir korelasyon bulunmadığını varsayan E beklenen derecelendirme matrisi oluşturulmaktadır. Bu üç matris kullanılarak (O, W ve E) karesel ağırlıklı Kappa değeri eşitlik 27 ile hesaplanmaktadır.

$$k = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \qquad (27)$$

Elde edilen değerlere Fisher dönüşümü uygulanmaktadır. Bu amaçla kullanılacak denkleme eşitlik 28'de yer verilmektedir. Fisher dönüşümü, varyansı dengeleyen bir yapıya sahiptir.

$$z = \frac{1}{2} \ln \frac{1+k}{1-k} \qquad (28)$$

Bu işlemden sonra dönüştürülmüş Kappa değerlerinin z uzayında ortalaması alınmaktadır. Her bir Kappa ortalaması açık uçlu madde sayısına göre ağırlıklandırılmaktadır. Ortalama Kappa değerini elde etmek için son olarak ters dönüşüm uygulanmaktadır. Bu amaçla kullanılacak denkleme eşitlik 29'da yer verilmektedir.

$$k = \frac{e^{2x}-1}{e^{2x}+1} \qquad (29)$$

Gwet'in AC1 katsayısı (Gwet's AC1 index). Gwet'in AC1 katsayısı (Gwet, 2008) Cohen'in Kappa katsayısında karşılaşılan paradokslar doğrultusunda ortaya çıkmıştır. Verilerin kategorilere dağılımındaki (yaygınlık [prevalence]) çarpıklık, puanlayıcılardan kaynaklanan yanlılık (bias), puanlayıcıların duyarlık (sensitivity) ve özgüllüğünün (specificity) farklılaşması Kappa değerinin puanlayıcılar arasındaki uyumu tespit etme yeteneğini düşürmektedir (Eugenio ve Glass, 2004; Gwet, 2008). AC1 katsayısı Kappa katsayısından her bir kategori için marjinal olasılık ortalamaları ile şans uyumunun beklenen oranı üzerinde yaptığı düzeltme ile ayrılmaktadır. Böylece Kappa değerine göre paradokslardan daha az etkilenmekte, kategoriler arasındaki çarpıklık yani kategoriler arasındaki değişkenliğe karşı daha kararlı olmaktadır (Hoek ve Scholman, 2017).

AC1 katsayısının kategorilerde dengesizlik ve simetri eksikliği bulunduğunda puanlayıcılar arası uyumu tespit etme yeteneği fazladır (Shankar ve Bangdiwala, 2014). Gwet'in AC1 katsayısı kategorik veride ve herhangi bir sayıda puanlayıcı bulunması durumunda kullanılabilmektedir (Wongpakaran, Wongpakaran, Wedding ve Gwet, 2013). Eğer veri sıralı ya da aralık ölçeğinde ise Gwet tarafından oluşturulan AC2 katsayısı kullanılmalıdır (Gwet, 2014). AC1 katsayısı; uyum yüzdesinden daha düşük, Kappa katsayısından ise daha yüksek değerler almaktadır (Lacy, Watson, Riffe ve Lovejoy, 2015). Gwet'in AC1 katsayısı Landis ve Koch (1977) tarafından Kappa katsayısı için sunulan kriterler aracılığıyla yorumlanabilmektedir (Senay, Delisle, Raynauld, Morin ve Fernandes, 2015; Siriwardhana, Walters, Rait, Bazo-Alvarez ve Weerasinghe, 2018). Dolayısıyla tablo 6'daki değerler kullanılabilir. Hoek ve Scholman (2017) araştırmacılara, araştırmalarında Kappa değeri ile birlikte AC1 değerinin kullanımını önermektedir. Bunun yanı sıra Haley (2007) AC1 katsayısının otomatik puanlama sistemlerini değerlendirmede iyi bir çözüm olduğunu belirtmektedir.

Gwet'in AC1 katsayısını hesaplamak için kullanılan denklem, eşitlik 30'da gösterilmektedir (Gwet, 2016). Bu denklemde P_a uyum yüzdesi, P_e şans uyum yüzdesini göstermektedir.

$$K_G^{\wedge} = \frac{P_a - P_e}{1 - P_e} \qquad (30)$$

Uyum yüzdesi P_a'yı hesaplamak için kullanılan denklem eşitlik 31 ile gösterilmektedir. Bu denklemde n kişi sayısını göstermektedir, denklemde yer alan $P_a|i$'yi hesaplamak için gerekli olan denklem de eşitlik 32'de gösterilmektedir. $P_a|i$'yi hesaplamak için kullanılan denklemde yer alan q kategori sayısını, r puanlayıcı sayısını göstermektedir.

$$P_a = \frac{1}{n} \sum_{i=1}^{n} P_a|i \qquad (31)$$

$$P_a|i = \sum_{k=1}^{q} \frac{r_{ik}\,(r_{ik}-1)}{r\,(r-1)} \qquad (32)$$

Şans uyum yüzdesi P_e'yi hesaplamak için kullanılan denkleme eşitlik 33'de yer verilmiştir. Bu denklemde yer alan q olası kategori sayısını, π_k seçkisiz puanlayıcının seçkisiz kişiyi k kategorisine sınıflandırma olasılığını göstermektedir ve hesaplanması için gerekli denkleme eşitlik 34'de yer verilmektedir.

$$P_e = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k (1 - \pi_k) \qquad (33)$$

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ik}}{r} \qquad (34)$$

Uyum katsayıları yorumlanırken puanların yaygınlığı ve puanlayıcıların yanlılığı önem taşımaktadır. Yaygınlık sorunu puanların kategorilere dağılımında çarpıklık olduğunda ortaya çıkmakta iken yanlılık sorunu puanlayıcıların bir duruma ilişkin değerlendirmelerine ait frekanslar arasında farklılık olduğunda ortaya çıkmaktadır (Bryt, Bishop ve Carlin, 1993; Eugenio ve Glass, 2004). Bu sorunların var olup olmadığını incelemek üzere yaygınlık ve yanlılık indeksleri hesaplanabilmektedir. Özellikle Kappa katsayısı, yaygınlık ve yanlılık katsayılarından oldukça etkilenmektedir. Bryt, Bishop ve Carlin (1993) sadece Kappa katsayısının kullanımının yanıltıcı olabileceğini, Kappa katsayısı kullanıldığında yaygınlık ve yanlılık indekslerinin de tartışılması gerektiğini belirtmektedir.

İki puanlayıcı tarafından iki kategorili puanlanan bir maddenin yaygınlık ve yanlılık indekslerini hesaplamak için öncelikle denklemde kullanılacak değerler belirlenmelidir. Bu değerler tablo 7 temel alınarak bulunabilir.

Tablo 7

İki Puanlayıcı ve İki Kategorili Puanlamaya Ait Matris

		Puanlayıcı 1	
		1	0
Puanlayıcı 2	1	a	b
	0	c	d

Tablo 7'deki a değeri her iki puanlayıcının da 1 puanı verdiği kişi sayısını, d değeri her iki puanlayıcının da 0 puan verdiği kişi sayısını göstermektedir. N ise tüm kişilerin sayısını göstermektedir. Bu yönde iki puanlayıcı ve iki puan kategorisi için yaygınlık indeksi (prevalence index-PI) eşitlik 35'de yer alan denklemle hesaplanmaktadır (Byrt, Bishop ve Carlin, 1993). Yaygınlık indeksi hesaplandıktan sonra bu değerlerin mutlak değeri alınmış ve katsayının büyüklüğüne odaklanılmıştır.

$$PI = (a-d)/N \qquad (35)$$

İki puanlayıcı ve iki puan kategorisi için yanlılık indeksi hesaplanırken puanlayıcı birin 0 ve puanlayıcı ikinin 1 puanı verdiği kişi sayısından puanlayıcı birin 1 puan ve puanlayıcı ikinin 0 puanı verdiği kişi sayısı çıkarılarak bu değer N yani toplam kişi sayısına bölünür. Eşitlik 36 aracılığıyla iki puanlayıcı ve iki puan kategorisi için yanlılık indeksi (bias index-BI) hesaplanabilmektedir (Byrt, Bishop ve Carlin, 1993). Yanlılık indeksleri için mutlak değerler alınarak sadece yanlılığın büyüklüğüne odaklanılmıştır.

$$BI = [(a+b)/N] - [(a+c)/N] = (b-c)/N \qquad (36)$$

Üç kategorili ve iki puanlayıcının yaptığı değerlendirmeler için yaygınlık ve yanlılık indeksleri hesaplanırken üç durum üzerinde (durum 1:0, durum 2:1 ve durum 3:2 temel alınarak) çalışılarak bu durumların her biri için yaygınlık ve yanlılık katsayıları hesaplanmış ve bu indekslerin mutlak değeri ardından ise ortalamaları alınmıştır. Üç farklı durum, üç farklı puan için 2x2 matisleri aracılığıyla yapılacak hesaplamaları anlatmaktadır. Tablo 8, örnek olarak 2 puanı için yapılacak

hesaplamalarda kullanılacak değerleri elde edecek matrisi göstermektedir. 2 puanı dikkate alınacak olursa a her iki puanlayıcının 2 puanı verdiği kişi sayısını, b ilk puanlayıcının 0 ya da 1 puanı verip ikinci puanlayıcının 2 puanı verdiği kişi sayısını, c birinci puanlayıcının 2 puanı verip ikinci puanlayıcının 0 ya da 1 puanı verdiği kişi sayısını, d ise puanlayıcıların ikisinin de 1, ikisinin de 0, birincinin 0 ve ikincinin 1, birincinin 1 ve ikincinin 0 puanı verdiği kişi sayısını göstermektedir. N toplam kişi sayısını göstermek üzere yaygınlık ve yanlılık indeksi sırasıyla eşitlik 35 ve eşitlik 36 aracılığıyla hesaplanabilir. 0 ve 1 puanları için oluşturulan tablolara sırasıyla EK-B ve EK-C'de yer verilmiştir.

Tablo 8

İki Puanlayıcı ve Üç Kategorili Puanlamada Durum 3'e İlişkin Matris

		Puanlayıcı 1	
		Doğru (2)	Yanlış (0 ve 1)
Puanlayıcı 2	Doğru (2)	a	b
	Yanlış (0 ve 1)	c	d

Tablo 8'deki değerlerden yararlanılarak her bir kitapçıkta yer alan tüm yapılandırılmış cevap maddelerine ilişkin yaygınlık ve yanlılık katsayıları incelenmiştir. Yaygınlık indeksi -1 ile 1 aralığında değişen katsayılar almakta olup mutlak değerler kullanıldığından elde edilen katsayıların 1'e yaklaşmasının Kappa değerini düşüreceği belirtilebilir. Yanlılık indeksinin mutlak değerleri ise 0 ile 1 aralığında değişmekte olup yanlılık katsayılarının yükselmesinin Kappa değerini yükselteceği belirtilebilir (Bryt, Bishop ve Carlin, 1993). Tablo 9, A_1 kitapçığında yer alan yapılandırılmış cevap maddelerine ilişkin yaygınlık ve yanlılık katsayılarını göstermektedir. Tablo 9'da yüksek bulunan yaygınlık ve yanlılık katsayıları koyu, düşük bulunan yaygınlık ve yanlılık katsayıları ise italik olarak gösterilmiştir. Bu katsayıların yanı sıra tablo 9'da kategorilere dağılıma ilişkin frekans ve yüzde değerlerine de yer verilmiştir.

Tablo 9

A_1 Kitapçığı Yapılandırılmış Cevap Maddelerine İlişkin Yaygınlık (PI) ve Yanlılık (BI) Katsayıları

Madde No	Kategori	f	%	PI	BI
Madde 2	0	54	,089	**,827**	*,012*
	1	553	,911		
Madde 7	0	114	,188	**,426**	*,009*
	1	144	,237		
	2	349	,575		
Madde 8	0	129	,213	,333	*,002*
	1	193	,318		
	2	285	,470		
Madde 10	0	260	,428	*,163*	*,020*
	1	347	,572		
Madde 11	0	222	,366	*,259*	*,007*
	1	385	,634		
Madde 14	0	163	,269	**,456**	*,003*
	1	444	,731		
Madde 15	0	111	,183	,363	*,008*
	1	183	,301		
	2	313	,516		
Madde 18	0	318	,524	,363	*,002*
	1	203	,334		
	2	86	,142		
Madde 19	0	405	,667	**,554**	*,002*
	1	80	,132		
	2	122	,201		

Tablo 9 incelendiğinde A1 kitapçığında yer alan yapılandırılmış cevap maddelerinin yaygınlık katsayılarının ,163 ile ,827 aralığında değiştiği görülmektedir. En yüksek yaygınlık katsayısı madde 2'de gözlenmiştir. Madde 2'ye ait yaygınlık katsayısının (,827) oldukça yüksek olduğu belirtilebilir. Bu katsayıyı sırasıyla ,554 yaygınlık katsayısı ile madde 19, ,456 yaygınlık katsayısı ile madde 14 ve ,426 yaygınlık katsayısı ile madde 7 takip etmektedir. Bu maddelerde de yaygınlık katsayıları nispeten yüksektir. En düşük yaygınlık katsayısına sahip maddeler madde 10 (,163) ve madde 11 (,259)'dir. Bu maddelerin yaygınlık

katsayıları düşüktür. A1 kitapçığında yer alan yapılandırılmış cevap maddelerinin yanlılık katsayılarının ise ,002 ile ,020 aralığında değiştiği belirtilebilir. Yanlılık katsayılarının tamamının 0'a yakın olduğu söylenebilir. Bu durum tüm maddelerde puanlayıcı yanlılığının neredeyse bulunmadığı anlamına gelmektedir. İki kategorili bazı maddelerde (madde 10 ve madde 11) kategorilere dağılım kısmen yakın olsa da bazı maddelerde (madde 2 ve madde 14) kategorilere dağılımda oldukça fark bulunmaktadır. Üç kategorili maddeler incelendiğinde ise puanların kategorilere dağılımında genellikle fark olduğu görülmektedir.

Tablo 10, B_1 kitapçığında yer alan yapılandırılmış cevap maddelerine ilişkin yaygınlık ve yanlılık katsayılarını bunun yanı sıra kategorilere dağılıma ilişkin frekans ve yüzde değerlerini göstermektedir. Tablo 10'da yüksek bulunan yaygınlık ve yanlılık katsayıları koyu, düşük bulunan yaygınlık ve yanlılık katsayıları ise italik olarak gösterilmiştir.

Tablo 10 incelendiğinde B_1 kitapçığında yer alan yapılandırılmış cevap maddelerinin yaygınlık katsayılarının ,070 ile ,682 aralığında değiştiği görülmektedir. B_1 kitapçığında en yüksek yaygınlık katsayısı (,682) madde 3'te ortaya çıkmıştır. Bu katsayıyı ,495 yaygınlık katsayısı ile madde 5 takip etmektedir. Bu maddeye ait yaygınlık katsayısının da yüksek olduğu belirtilebilir. En düşük yaygınlık katsayısına sahip maddeler madde 18 (,070), madde 8 (,147) ve madde 9 (,269)'dur. Bu maddelere ait yaygınlık katsayıları düşüktür. B_1 kitapçığında yer alan yapılandırılmış cevap maddelerinin yanlılık katsayılarının ise ,000 ile ,021 aralığında değiştiği belirtilebilir. Yanlılık katsayılarının tamamının 0'a yakın olduğu görülmektedir. Bu durum tüm maddelerde puanlayıcı yanlılığının neredeyse bulunmadığı anlamına gelmektedir. İkili puanlanan maddelerde kategorilere dağılım dikkate alındığında maddelerin bir kısmında (madde 8 ve madde 18) frekanslar yakın iken bir kısmında (madde 3, madde 9 ve madde 20) ise frekanslar arasında oldukça fark vardır. Üç kategorili puanlanan maddeler incelendiğinde madde 5, madde 6, madde 11 ve madde 17'ye ait frekansların farklılaştığı; madde 12'de ise diğer maddelere göre frekansların daha yakın olduğu görülmektedir.

Tablo 10

B_1 Kitapçığı Yapılandırılmış Cevap Maddelerine İlişkin Yaygınlık (PI) ve Yanlılık (BI) Katsayıları

Madde No	Kategori	f	%	PI	BI
Madde 3	0	101	,173	**,682**	*,007*
	1	483	,827		
Madde 5	0	108	,185	**,495**	*,008*
	1	113	,193		
	2	363	,622		
Madde 6	0	130	,223	,333	*,007*
	1	166	,284		
	2	288	,493		
Madde 8	0	257	,440	*,147*	*,003*
	1	327	,560		
Madde 9	0	211	,361	*,269*	*,002*
	1	373	,639		
Madde 11	0	270	,462	,333	*,000*
	1	50	,086		
	2	264	,452		
Madde 12	0	200	,342	,333	*,021*
	1	168	,288		
	2	216	,370		
Madde 17	0	92	,158	,333	*,006*
	1	206	,353		
	2	286	,490		
Madde 18	0	312	,534	*,070*	*,002*
	1	272	,466		
Madde 20	0	394	,675	,377	*,007*
	1	190	,325		

A_1 ve B_1 kitapçığında yer alan maddelerin tamamının yaygınlık katsayıları değerlendirildiğinde A_1 kitapçığında yer alan madde 2, madde 7, madde 14 ve madde 19'un, B_1 kitapçığında yer alan madde 3 ve madde 5'in yaygınlık katsayısının yüksek olduğu ve bu nedenle bu maddelerde QWK değerinin olduğundan düşük olabileceği öngörülmektedir. A_1 kitapçığında en düşük yaygınlık katsayısına sahip maddeler madde 10 ve madde 11, B_1 kitapçığında en düşük yaygınlık katsayısına sahip maddeler madde 8, madde 9 ve madde 18'in QWK değerinin gerçekte olan

uyuma daha yakın olacağı öngörülmektedir. A_1 ve B_1 kitapçığında yer alan maddelerin tamamının yanlılık değerleri ele alındığında tüm maddelerin yanlılık değerinin çok düşük olduğu bu nedenle de QWK değerinin olduğundan yüksek bulunma ihtimalinin düşük olduğu sonucuna ulaşılmaktadır.

Uyum katsayıları hesaplanırken R programında bulunan "irr" (Gamer, Fellows, Lemon ve Singh, 2010), "rel" (Martire, 2017) ve "Metrics" (Frasco, 2018) paketlerinden yararlanılmıştır. Uyum yüzdesi için "irr", karesel ağırlıklı Kappa katsayısı için "rel" ve Gwet's AC1 katsayısı için "Metrics" paketi kullanılmıştır.

Uyum katsayıları için tüm maddelerin ortalaması alınarak sistemin performansı değerlendirilmiştir. Her bir yöntem ve test veri oranı için ortalamalar hesaplandıktan sonra yöntemlerin ortalama performansları belirlenmiştir.

Tüm koşullar dikkate alınarak otomatik puanlamada en iyi uyumu gösteren üç koşul belirlenmiş (BLSTM %10, BLSTM %20 ve BLSTM %33) ve eşitleme işlemine geçilmiştir. Karşılaştırma yapabilmek için ise her test formuna ilişkin gerçek puanlayıcıların nihai puanları kullanılarak test formları eşitlenmiştir. Eşitleme işleminde KTK ve MTK'ya dayalı yöntemler kullanılmıştır. Eşitleme işlemi öncesinde bu araştırmaya konu olan test verilerine ilişkin istatistiklere, güvenirlik değerlerine ve KTK'ya dayalı madde istatistiklerine, yer verilmiştir. Gerçek puanlayıcılar için A_1 ve B_1 kitapçığına ilişkin istatistikler ve güvenirlik değerleri tablo 11'de bulunmaktadır. Otomatik puanlama aracılığıyla (BLSTM %10, BLSTM %20 ve BLSTM %33) elde edilen test istatistikleri ve güvenirlik katsayılarına EK-Ç'de yer verilmiştir. Güvenirlik katsayısı iki şekilde incelenmiştir. İlk durumda güvenirlik Cronbach'ın alfa (Cronbach, 1951) katsayısı ile ikinci durumda ise faktör analizine dayalı olarak McDonald'ın omega (McDonald, 1999) katsayısı ile belirlenmiştir. Alfa katsayısı güvenirliğin alt sınırını vermesi nedeniyle kullanılırken omega katsayısı daha az ve daha gerçekçi varsayımlara sahip olması nedeniyle seçilmiştir (Bendermacher, 2010; Dunn, Baguley ve Brunsden, 2014).

Tablo 11

A1 ve B1 Kitapçıklarına İlişkin İstatistikler

	Kitapçık	
	A_1	B_1
Madde Sayısı	18	18
Örneklem Sayısı	607	584
Ortalama	13,152	14,101
Standart Sapma	4,530	4,964
Medyan (Ortanca)	14	15
Minimum Değer	1	0
Maksimum Değer	23	23
Çarpıklık	-,249	-,466
Güvenirlik (Alfa)	,766	,797
Güvenirlik (Omega)	,868	,893

Tablo 11 incelendiğinde her iki kitapçıkta 18'er maddenin bulunduğu, A_1 kitapçığını 607 kişinin, B_1 kitapçığını ise 584 kişinin yanıtladığı görülmektedir. A_1 kitapçığına ait ortalamanın (13,152), B_1 kitapçığına ait ortalamadan (14,101) biraz daha düşük olduğu görülmektedir. A_1 kitapçığının medyan (ortanca) değeri (14), B_1 kitapçığının medyan (ortanca) değerinden (15) düşüktür. Test kitapçıkları standart sapma açısından karşılaştırıldığında A_1 kitapçığının standart sapmasının (4,530), B_1 kitapçığının standart sapmasından (4,964) düşük olduğu görülmektedir. A_1 test formundan katılımcıların aldıkları en düşük puan 1 iken B_1 kitapçığından katılımcıların aldıkları en düşük puan 0'dır. Her iki test kitapçığından alınan en yüksek puan 23'dür. Çarpıklık değeri için karşılaştırma yapıldığında her iki test kitapçığına ait verinin sola çarpık olduğu A_1 kitapçığının çarpıklık katsayısının (-,249), B_1 kitapçığının çarpıklık katsayısından (-,466) düşük olduğu görülmektedir. Buna göre B_1 kitapçığından elde edilen verilerin daha çarpık olduğu belirtilebilir. Güvenirlik katsayıları açısından karşılaştırma yapıldığında alfa katsayılarının her iki test kitapçığının da güvenilir olduğunu gösterdiği bunun yanında B_1 kitapçığından elde edilen verilerin güvenirliğinin (,797), A_1 kitapçığından elde edilen verilerin güvenirliğinden (,766) yüksek olduğu belirtilebilir. Nitekim Cortina (1993) Cronbach'ın alfa katsayısının ,70 üzerinde olmasının güvenirliği göstermede yeterli olacağını belirtmektedir. Güvenirlik katsayıları McDonald'ın omega katsayısına göre karşılaştırıldığında B_1 kitapçığına ait omega katsayısının (,893), A_1 kitapçığına ait

omega katsayısından (,868) yüksek olduğu ve her iki kitapçığa ait güvenirlik değerlerinin yüksek olduğu görülmektedir.

A_1 ve B_1 kitapçığında yer alan maddelerin KTK'ya dayalı madde istatistikleri incelenmiştir. Tablo 12, A_1 ve B_1 kitapçığında yer alan maddelere ilişkin güçlük, ayırt edicilik değerlerini bunun yanı sıra madde çıkarıldığında elde edilecek alfa değerini göstermektedir. Tablo 12'de yer alan değerler gerçek puanlayıcıların bulunduğu koşula ilişkindir. Otomatik puanlama için KTK'ya dayalı madde istatistiklerine BLSTM yöntemi %10 test veri oranı için EK-D, BLSTM yöntemi %20 test veri oranı için EK-E ve BLSTM yöntemi %33 test veri oranı için EK-F'de yer verilmektedir. Ayırt edicilik değerleri iki kategorili puanlanan maddeler için nokta çift serili korelasyon katsayısıyla, üçlü puanlanan maddeler için eta katsayısıyla elde edilmiştir. Üçlü puanlanan maddeler için güçlük katsayısı hesaplanırken eşitlik 37'den yararlanılmıştır (Brookhart ve Nitko, 2015).

$$P \text{ (Güçlük)} = \frac{Madde\ ortalaması - Olası\ minimum\ puan}{Olası\ maksimum\ puan - Olası\ minimum\ puan} \qquad (37)$$

Tablo 12 incelendiğinde A_1 kitapçığında yer alan maddelerin güçlük indekslerinin ,257 (madde 17) ile ,911 (madde 2) aralığında değiştiği görülmektedir. B_1 kitapçığında yer alan maddelerin güçlük indekslerinin ise ,260 (madde 14) ile ,909 (madde 1) aralığında değiştiği görülmektedir. Her iki kitapçıkta da çok kolay (P>,80) ve çok zora (P<,25) yakın maddelerin bulunduğu görülmektedir (Brookhart ve Nitko, 2015). A_1 kitapçığındaki en kolay madde 2, B_1 kitapçığındaki en kolay maddeler ise madde 1 ve madde 3'tür. A_1 kitapçığındaki en zor madde madde 17 ve madde 19 iken B_1 kitapçığında ise madde 14'tür. A_1 kitapçığında maddeler genel olarak kolay ve orta güçlükte iken zor maddeler de bulunmaktadır. B_1 kitapçığında maddeler genel olarak orta güçlükte iken kolay ve zor maddeler de bulunmaktadır. Kitapçıklardaki ortak maddelerin güçlük indeksleri birbirleriyle karşılaştırıldığında bu değerlerin yakın olduğu görülmektedir. Madde ayırt edicilikleri incelendiğinde A_1 kitapçığında değerlerin ,055 (madde 17) ile ,627 (madde 19) aralığında değiştiği görülmektedir. A_1 kitapçığında en yüksek ayırt edicilik 19. maddede elde edilmiştir. B_1 kitapçığında ise değerlerin ise ,014 (madde 14) ile ,617 (madde 11) aralığında değiştiği görülmektedir. B_1 kitapçığında en yüksek ayırt edicilik 11. maddede elde edilmiştir. Her iki kitapçıkta da birer maddenin (A_1 kitapçığı madde 17 ve B_1 kitapçığı madde 14) oldukça düşük, A_1 kitapçığındaki bir maddenin (madde 1) düşük ayırt

ediciliğe sahip olduğu bulunmuştur. A_1 ve B_1 kitapçıklarında kalan tüm maddelerin ,300 ayırt edicilik değerine çok yakın ya da bu değerden yüksek olduğu bulunmuştur. Testlerdeki ortak maddelerin ayırt edicilikleri incelendiğinde bazı madddelerde değerlerin farklılaştığına rastlanmıştır. Bu durum KTK'nın testi alan örnekleme bağlı olmasıyla açıklanabilir. Maddeler çıkarıldığında elde edilecek alfa değerleri incelendiğinde A_1 kitapçığında madde 17'nin testten çıkarılmasının alfa katsayısını arttıracağı görülmektedir. Bu duruma maddenin çok zor ve düşük ayırt ediciliğe sahip olması neden olmuş olabilir. A_1 kitapçığında madde 1'in testten çıkarılmasının da alfa katsayısını arttıracağı görülmektedir. Ancak alfa katsayısında oluşacak değişim çok düşüktür. B_1 kitapçığında sadece 14. maddenin testten çıkarılmasının güvenirliği arttıracağı görülmektedir. Bu duruma da maddenin çok zor ve ayırt ediciliğin çok düşük olması neden olmuş olabilir.

Tablo 12

A_1 ve B_1 Kitapçığında Yer Alan Maddelerin KTK'ya Dayalı İstatistikleri

Kitapçık A_1				Kitapçık B_1			
Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)	Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)
1	,774	,142	,767	1	,909	,370	,792
2	,911	,351	,758	2	,630	,363	,789
5	,550	,371	,754	3	,827	,476	,785
6	,341	,315	,758	4	,726	,475	,784
7*	,694	,529	,749	5*	,718	,532	,784
8*	,628	,486	,754	6*	,636	,548	,783
9	,717	,279	,760	7	,716	,388	,788
10	,572	,439	,749	8	,560	,389	,788
11	,634	,320	,757	9	,639	,345	,790
12	,530	,296	,759	10	,570	,297	,793
13	,746	,444	,750	11*	,495	,617	,784
14	,731	,538	,744	12*	,514	,583	,783
15*	,666	,357	,764	13	,738	,399	,788
16	,717	,331	,757	14	,260	,014	,806
17	,257	,055	,773	17*	,666	,411	,793
18*	,309	,537	,752	18	,466	,545	,779
19*	,267	,627	,744	19	,678	,378	,788
20	,542	,481	,746	20	,325	,436	,785

* Üçlü puanlanan maddeleri göstermektedir.

Not: A_1 ve B_1 kitapçıklarında ortak maddeler bulunmaktadır. İlk değer A_1, ikinci değer B_1 kitapçığındaki madde numaralarını göstermek üzere ortak maddeler şu şekildedir: 7-5, 8-6, 9-7, 10-8, 11-9, 12-10, 15-17, 16-13, 17-14.

KTK'ya dayalı eşitleme. KTK'ya dayalı eşitleme yöntemlerinden zincir doğrusal (chained linear [LC]), Tucker doğrusal (Tucker linear [LT]), zincir eşit yüzdelikli (chained equipercentile [EC]), frekans eşit yüzdelikli (frequency estimation equipercentile [EF]) eşitleme yöntemleri seçilmiştir. Sentetik evren değeri $w_1=1$ (WS=1) olarak değiştirilerek bu durumun etkisi değerlendirilmiştir. Sentetik evrenin $w_1=1$ olarak belirlenmesi denk olmayan gruplarda ortak madde deseninde yeni test formunu alan grubun sentetik evren olarak belirlenmesi anlamına gelmektedir (Kolen ve Brennan, 2014). Sentetik evren değeri değiştirilmediğinde gruplarda yer alan örneklem sayılarına oranla ($w_1+w_2=1$ olacak şekilde) sentetik evren belirlenmektedir. Fakat zincir eşitleme sentetik evreni desteklemediğinden zincir eşitlemenin kullanıldığı yöntemlerde sentetik evren oranları değiştirilmemiştir (Kolen ve Brennan, 2014). Bunun yanı sıra eşit yüzdelikli eşitleme yöntemleri için ön düzgünleştirme (presmoothing [PSM]) yapılmıştır. Frekans eşit yüzdelikli eşitleme yöntemi için hem ön düzgünleştirme yapılmış hem de sentetik evren oranı değiştirilmiştir. Yapılan bu değişikliklerle sentetik evren parametrelerinin ve/veya ön düzgünleştirmenin eşitleme sonuçlarına etkisi de değerlendirilmiştir. KTK yöntemlerine göre eşitleme yapılırken R (R Development Core Team, 2018) programında yer alan "equate" (Albano, 2016) paketi kullanılmıştır. Ön düzgünleştirme işlemi SAS 9.4 (SAS Institute, 2015) programında PROC IML (Moses ve von Davier, 2006) kodu kullanılarak gerçekleştirilmiştir. Bu işlemin R programı dışında gerçekleştirilmesinin sebebi A_1 testinden ya da B_1 testinden alınan toplam puanlarla ortak formdan alınan toplam puanların puan kombinasyonlarına ilişkin frekansların bazılarının sıfır olması nedeniyle çıkarılması gerekliliğidir (Moses, von Davier ve Casabianca, 2004). Ön düzgünleştirme işlemi denk olmayan gruplarda ortak madde deseni kullanılması nedeniyle polinominal iki değişkenli doğrusal logaritmik fonksiyon (polynominal bivariate loglinear) dağılımı kullanılarak gerçekleştirilmiştir. Her bir form için polinominal iki değişkenli doğrusal logaritmik fonksiyon dağılımındaki 11 farklı model karşılaştırılarak en iyi model seçilmiştir. Modeller test puanları (x) ve ortak madde puanları (a) için içerdikleri değişkenlerle birbirinden ayrılmaktadır. x ve a ortalama, x^2 ve a^2 standart sapma, x^3 ve a^3 çarpıklık,

x^4 ve a^4 basıklık değişkenlerini göstermektedir. xa test puanları ve ortak madde puanlarının ortalamalası, x^2a test puanlarının standart sapması ile ortak madde puanlarının ortalaması, xa^2 test puanlarının ortalaması ile ortak madde puanlarının standart sapması ve x^2a^2 test puanlarının standart sapması ile ortak madde puanlarının standart sapması arasındaki etkileşimi göstermektedir. Benzer yorumlar Y formu için yapılabilir. Modeller seçilirken ki kare değerleri arasındaki fark serbestlik derecesi farkına göre %5'lik hata payı için belirlenen ki kare değeriyle karşılaştırılmış bunun yanında AIC ve BIC değerleri dikkate alınmıştır. Ki kare tablosuna EK-G'de yer verilmiştir. Tablo 13, gerçek puanlayıcıların kullanıldığı durumdaki ki kare, serbestlik derecesi, AIC ve BIC değerlerini göstermektedir. Otomatik puanlama için ki kare, serbestlik derecesi, AIC ve BIC değerleri BLSTM yöntemi %10 test veri oranı için EK-H, BLSTM yöntemi %20 test veri oranı için EK-I ve BLSTM yöntemi %33 test veri oranı için EK-İ'de gösterilmektedir. Eşitleme işlemleri bootstrap tekniği kullanılarak gerçekleştirilmiş olup 10000 tekrar üzerinde çalışılmıştır. Seçilen modeller tablo 13, EK-H, EK-I ve EK-İ'de koyu şekilde gösterilmektedir.

Tablo 13 incelendiğinde A_1 ve B_1 kitapçıklarına yönelik modellere ilişkin ki kare, serbestlik derecesi, AIC, BIC değerleri görülmektedir. Modeller arasında seçim yapılırken modeller ikişerli karşılaştırılmış olup öncelikle serbestlik derecesi farkları ve ki kare farkları hesaplanmıştır. Ardından ki kare anlamlılık tablosu ile karşılaştırma yapılarak %5 hata payı ile anlamlı bulunan modeller için bir üst modele geçiş yapılmıştır. Aynı serbestlik derecesine sahip modeller arasında seçim yapılırken ise AIC ve BIC değerleri incelenmiştir. Sonuçta gerçek puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemlerinde A_1 kitapçığı için en uygun modelin model 6, B_1 kitapçığı için en uygun modelin model 11 olduğu bulunmuştur.

Tablo 13

Gerçek Puanlayıcıların Kullanıldığı Durumlarda Ön Düzgünleştirme Modelinin Belirlenmesi

		A_1 Kitapçığı				B_1 Kitapçığı			
		ki kare	sd	AIC	BIC	ki kare	sd	AIC	BIC
Model 1	x a	373,033	95	681,916	689,671	294,652	107	656,015	664,117
Model 2	x a xa	297,935	94	633,498	643,838	282,036	106	643,985	654,787
Model 3	x x^2 a a^2	263,622	93	605,812	618,737	220,896	105	580,463	593,965
Model 4	x x^2 a a^2 xa	90,920	92	434,883	450,393	130,725	104	474,350	490,553
Model 5	x x^2 x^3 a a^2 a^3	250,985	91	600,757	618,852	198,675	103	557,348	576,251
Model 6	x x^2 x^3 a a^2 a^3 xa	**78,186**	**90**	**424,830**	**445,510**	109,043	102	459,616	481,220
Model 7	x x^2 x^3 x^4 a a^2 a^3 a^4	247,933	89	602,269	625,534	179,619	101	547,812	572,116
Model 8	x x^2 a a^2 xa x^2a xa^2 x^2a^2	76,250	89	424,222	447,487	122,767	101	456,827	481,132
Model 9	x x^2 x^3 x^4 a a^2 a^3 a^4 xa	75,736	88	426,174	452,024	89,606	100	451,781	478,785
Model 10	x x^2 x^3 a a^2 a^3 xa x^2a xa^2 x^2a^2	71,564	87	423,470	451,905	83,617	99	455,158	484,863
Model 11	x x^2 x^3 x^4 a a^2 a^3 a^4 xa x^2a xa^2 x^2a^2	69,176	85	426,011	459,615	**77,403**	**97**	**454,277**	**489,383**

MTK'ya dayalı eşitleme. MTK'ya dayalı eşitleme yöntemlerinden gerçek puan eşitlemede kullanılan ayrı kalibrasyona dayalı ortalama-ortalama (mean-mean [MM]), ortalama-standart sapma (mean-sigma [MS]), Haebara (HB) ve Stocking Lord (SL) yöntemleri kullanılmıştır. MTK ile eşitleme işlemine geçilmeden önce MTK varsayımları incelenmiştir. İncelenen ilk varsayım tek boyutluluktur. Her bir test formu için karma testlere yönelik faktör analizi MPLUS (Muthén ve Muthén, 2012) programı ile gerçekleştirilmiştir. Faktör analizi gerçekleştirilmeden önce verilerin faktör analizine uygunluğunu saptamak üzere örneklemin yeterliğine yönelik Bartlett küresellik testi (Bartlett, 1950) yapılmış ve Kaiser-Meyer-Olkin (Kaiser, 1970; Kaiser ve Rice, 1974) katsayısı hesaplanmıştır. Bartlett küresellik testi tüm değişken çiftleri arasındaki korelasyonları, korelasyon matrisi temelinde inceleyen istatistiksel bir

anlamlılık testidir (Hair, Black, Babin ve Anderson, 2014). Kaiser-Meyer-Olkin (KMO) katsayısı ise araştırılan faktörler için maddelerin ortak varyansını göstermektedir (Beavers, Lounsbury, Richards, Huck, Skolits ve Esquivel, 2013; Tabachnick ve Fidell, 2014). Tablo 14, gerçek puanlayıcılar için A_1 ve B_1 kitapçıklarına yönelik Bartlett küresellik testi ve KMO değerini göstermektedir. EK-J ise %10, %20 ve %33 test veri oranlarında BLSTM yöntemiyle gerçekleştirilen otomatik puanlama sonucunda A_1 ve B_1 kitapçığına ilişkin test verilerinin faktör analizine uygunluğuna ilişkin katsayıları göstermektedir.

Tablo 14

A_1 ve B_1 Kitapçıklarının Faktör Analizine Uygunluğuna Yönelik Bartlett Testi ve KMO Değeri

	A_1 Kitapçığı			B_1 Kitapçığı		
Kaiser-Meyer-Olkin	.864			.904		
Bartlett	χ^2	sd	p	χ^2	sd	p
	1459,4	153	,000*	1615,5	153	,000*

* p<.05

Tablo 14 incelendiğinde KMO katsayısının A_1 kitapçığı için ,864 ve B_1 kitapçığı için ,904 olarak bulunduğu görülmektedir. Beavers, vd. (2013) tarafından sunulan kriterlere göre değerlendirme yapıldığında A_1 kitapçığı için elde edilen değerin yüksek düzeyde, B_1 kitapçığından elde edilen değerin ise çok yüksek düzeyde ortak varyansa işaret ettiği görülmektedir. Bartlett küresellik testi sonuçları her iki kitapçık için anlamlı bulunmuştur (p<,05). Bartlett testi sonuçlarının anlamlı olması beklenmekte bu durum gözlenen matrisin birim (identity) matristen farklı olduğu anlamına gelmekte ve faktörleşebildiğini göstermektedir (Beavers vd., 2013; Pett, Lackey ve Sullivan, 2003). Karma test kullanılması nedeniyle faktör analizinde polikorik ve tetrakorik korelasyonlardan yararlanılmıştır. Faktör analizinde tahmin yöntemi olarak ortalama ve varyansın düzeltildiği ağırlıklandırılmış en küçük kareler (weighted least square mean and variance adjusted [WLSMV]) kullanılmıştır. WLSMV tahmin yöntemi polikorik ve tetrakorik korelasyonlar kullanıldığında en uygun yöntemlerden birisi olarak bilinmektedir (Barendse, Oort ve Timmerman, 2015). Ayrıca boyut sayısına karar verebilmek için Factor 10.5 (Lorenzo-Seva ve Fernando, 2006) programı aracılığıyla paralel analiz (Timmerman ve Lorenzo-Seva,

2011) gerçekleştirilmiştir. Paralel analiz sonuçları hem otomatik puanlama (%10, %20 ve %33 test veri oranlarıyla) hem de gerçek puanlayıcılar için her bir test formunun tek faktörlü yapıya sahip olduğunu göstermektedir. Tablo 15, gerçek puanlayıcılar için A_1 ve B_1 formundaki maddelerin tek faktörlü yapıdaki faktör yüklerini göstermektedir. %10, %20 ve %33 test veri oranlarıyla BLSTM yöntemi kullanılarak gerçekleştirilen otomatik puanlama için A_1 ve B_1 kitapçığındaki maddelerin tek faktörlü yapılardaki yükleri sırasıyla EK-K, EK-L ve EK-M'de gösterilmektedir.

Tablo 15

A_1 ve B_1 Kitapçığındaki Maddelerin Faktör Yükleri

Kitapçık A_1				Kitapçık B_1			
Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1
1	,226	12	,425	1	,700	10	,420
2	,701	13	,662	2	,503	11	,656
5	,523	14	,819	3	,745	12	,585
6	,443	15	,349	4	,685	13	,590
7	,555	16	,487	5	,591	14	,022
8	,502	17	,083	6	,570	17	,416
9	,443	18	,537	7	,567	18	,768
10	,637	19	,693	8	,544	19	,569
11	,501	20	,685	9	,484	20	,664

Tablo 15 incelendiğinde A_1 kitapçığında yer alan madde 17 ve B_1 kitapçığında yer alan madde 14'ün oldukça düşük faktör yüküne sahip olduğu görülmektedir. A_1 kitapçığında yer alan maddelerin faktör yükleri ,083 ile ,701 aralığında değişmektedir. B_1 kitapçığında yer alan maddelerin faktör yükleri ise ,022 ile ,768 aralığında değişmektedir. Stevens (2009)'a göre madde faktör yük değerleri ,40 üzerinde olmalıdır. A_1 kitapçığında çoğu madde bu koşulu sağlarken madde 1, madde 15 ve madde 17 bu koşulu sağlamamaktadır. B_1 kitapçığında madde 14 dışındaki tüm maddelerin faktör yükleri ,40'ın üzerindedir. Testlerde yer alan madde sayısının az olması ve kapsam geçerliği dikkate alınarak faktör yükü düşük maddeler testten çıkarılmadan araştırmaya devam edilmiştir.

Her bir test formu için verinin hangi MTK modeline uyum sağladığını belirlemek amacıyla beş model karşılaştırılmıştır. İki kategorili puanlanan

yapılandırılmış cevap maddeleri bulunduğu ve bu maddelere şansla cevap verme olasılığı bulunmadığı için tüm iki kategorili maddeler bir parametreli model (one parameter model-1PLM) ve iki parametreli model (two parameter model-2PLM) doğrultusunda incelenmiştir. İncelenen modeller 1) 1PLM ve kısmi puan modeli (partial credit model-PCM), 2) 1PLM ve genelleştirilmiş kısmi puan modeli (generalized partial credit model-GPCM), 3) 1PLM ve kademeli cevap modeli (graded response model-GRM), 4) 2PLM ve GPCM, 5) 2PLM ve GRM'dir. Modeller karşılaştırılırken -2loglikelihood ve serbestlik derecesi arasındaki farklar hesaplanarak bu değerler ki kare tablosu ile karşılaştırılmıştır. Karşılaştırmada EK-H'de bulunan tablo kullanılmıştır. Elde edilen değer ki kare tablosunda %5'lik hata payı için belirlenen değerden büyükse bir üst modele geçilmiştir. Aynı serbestlik derecesine sahip modeller karşılaştırılırken ise yeteneklerin (theta) kestirimine ilişkin standart hata ortalamaları kullanılmıştır. Buna göre standart hatası düşük modeller yetenek ve madde parametrelerini kestirmede kullanılmıştır. Gerçek puanlayıcılara ait nihai puanların ve otomatik puanlama sistemlerinin yaptığı puanlamaların tamamı için model karşılaştırılmasına gidilmiş ve tümünde 2PLM ve GPCM yönteminin daha uygun olduğu sonucuna ulaşılmıştır. Tablo 16, gerçek puanlayıcıların puanlama yaptığı koşulda A_1 ve B_1 kitapçıklarına yönelik model veri uyumunu belirleyen değerleri göstermektedir. A_1 ve B_1 kitapçıkları için otomatik puanlama model karşılaştırmalarına ilişkin bilgiler BLSTM yöntemi %10 test veri oranı için EK-N, BLSTM yöntemi %20 test veri oranı için EK-O, BLSTM yöntemi %33 test veri oranı için EK-Ö'de yer almaktadır. Seçilen modeller tablo 16, EK-N, EK-O ve EK-Ö'de koyu bir şekilde gösterilmektedir.

Tablo 16 incelendiğinde modellere ilişkin -2loglikelihood, serbestlik derecesi ve ki kare değerleri görülmektedir. Modeller karşılaştırılırken her iki kitapçıkta da model 4 ve model 5'in daha iyi uyum gösterdiği belirlenmiştir. Bu doğrultuda yetenek kestirimlerine ilişkin standart hata ortalamaları hesaplanarak seçim yapılmıştır. A_1 kitapçığında model 4'e ilişkin standart hata ortalaması ,463 bulunurken model 5'e ilişkin standart hata ortalaması ,468 bulunmuştur. B_1 kitapçığında model 4'e ilişkin standart hata ortalaması ,441 bulunurken model 5'e ilişkin standart hata ortalaması ,443 bulunmuştur. Her iki kitapçıkta model 4 ve model 5 için elde edilen standart hata ortalamaları birbirine oldukça yakındır. Daha düşük standart hata ortalamasına sahip olması nedeniyle her iki kitapçıkta da model 4 seçilmiştir.

Tablo 16

MTK Model Veri Uyumunu Belirlemede Kullanılan Değerler

		A_1 Kitapçığı				B_1 Kitapçığı			
		-2LL	ki kare	sd	SH	-2LL	ki kare	sd	SH
Model 1	1PLM ve PCM	13531	1170,741	322	,431	12795	1207,447	322	,438
Model 2	1PLM ve GPCM	13516	1138,041	317	,459	12796	1196,437	317	,456
Model 3	1PLM ve GRM	13483	1052,055	317	,490	12742	1091,238	317	,489
Model 4	2PLM ve GPCM	**12744**	**466,377**	**304**	**,463**	**12018**	**419,853**	**304**	**,441**
Model 5	2PLM ve GRM	12761	474,613	304	,468	12014	386,126	304	,443

Not: -2LL: -2LogLikelihood, SH: Standart hata, sd: Serbestlik derecesini göstermektedir.

Yetenek ve madde parametreleri XCALİBRE 4.1 (Yoes, 1996) programı kullanılarak kestirilmiştir. XCALİBRE programı ayırt edicilik ve güçlük parametrelerini BILOG (Mislevy ve Bock, 1997) programından daha düşük hata (RMSE) ile kestirmektedir (Weiss ve Minden, 2012). Bu nedenle kestirim yapılırken XCALIBRE programı kullanılmıştır. A_1 ve B_1 kitapçıkları için gerçek puanlayıcılar aracılığıyla kestirilen madde parametreleri tablo 17'de gösterilmektedir. BLSTM yöntemiyle %10, %20 ve %33 test veri oranlarıyla gerçekleştirilen otomatik puanlama işlemi sonrası kestirilen madde parametreleri ise sırasıyla EK-P, EK-R ve EK-S'de gösterilmektedir.

Tablo 17

A_1 ve B_1 Kitapçığında Yer Alan Maddelerin Parametreleri

Kitapçık A_1				Kitapçık B_1			
Madde No	a	b_1	b_2	Madde No	a	b_1	b_2
1	,209	-3,578		1	,925	-2,105	
2	,713	-2,464		2	,556	-0,691	
5	,536	-0,274		3	,952	-1,444	
6	,486	0,928		4	,835	-0,972	
7	,446	-0,930	-1,090	5	,461	-0,710	-1,450
8	,395	-1,090	-0,420	6	,469	-0,850	-0,530
9	,490	-1,298		7	,643	-1,070	
10	,728	-0,314		8	,607	-0,291	
11	,532	-0,730		9	,534	-0,750	
12	,472	-0,177		10	,469	-0,409	
13	,812	-1,096		11	,474	1,790	-1,740
14	1,184	-0,863		12	,484	-0,150	0,030
15	,284	-1,470	-1,030	13	,680	-1,157	
16	,548	-1,199		14	,268	2,351	
17	,293	2,217		17	,353	-1,850	-0,440
18	,494	0,450	1,700	18	1,040	0,131	
19	,622	1,540	0,380	19	,643	-0,865	
20	,842	-0,168		20	,835	0,733	

Tablo 17 incelendiğinde iki ve üç kategorili maddeler için kestirilen ayırt edicilik ve güçlük parametreleri görülmektedir. A_1 kitapçığı incelendiğinde en yüksek ayırt ediciliğe sahip maddenin 14. madde, B_1 kitapçığı incelendiğine en yüksek ayırt ediciliğe sahip maddenin 18. madde olduğu görülmektedir. Güçlük parametreleri ikili puanlanan maddeler için incelendiğine en kolay maddenin A_1 kitapçığında 1. Madde ve B_1 kitapçığında da 1. Madde olduğu belirtilebilir. İkili puanlanan maddeler için en zor madde A_1 kitapçığı için 17. madde, B_1 kitapçığı için 14. maddedir. Üçlü puanlanan maddeler için güçlük parametresi için inceleme yapıldığında ilk eşik değer için A_1 kitapçığında en yüksek değere sahip madde 19. madde, B_1 kitapçığında en yüksek değere sahip madde 11. maddedir. Yani bireyin %50 olasılıkla ilk eşiği aşması için en yüksek yeteneği bu maddeler gerektirmektedir. Üçlü puanlanan maddeler için güçlük parametresi için inceleme yapıldığında ilk eşik değer için A_1 kitapçığında en düşük değere sahip madde 15. madde, B_1 kitapçığında

en düşük değere sahip madde 17. maddedir. Yani bireyin %50 olasılıkla ilk eşiği aşması için en düşük yeteneği bu maddeler gerektirmektedir. Üçlü puanlanan maddeler için güçlük parametresi için inceleme yapıldığında ikinci eşik değer için A_1 kitapçığında en yüksek değere sahip madde 18. madde, B_1 kitapçığında en yüksek değere sahip madde 12. maddedir. Yani bireyin %50 olasılıkla ikinci eşiği aşması için en yüksek yeteneği bu maddeler gerektirmektedir. Üçlü puanlanan maddeler için güçlük parametresi için inceleme yapıldığında ikinci eşik değer için A_1 kitapçığında en düşük değere sahip madde 7. madde, B_1 kitapçığında en düşük değere sahip madde 11. maddedir. Yani bireyin %50 olasılıkla ikinci eşiği aşması için en düşük yeteneği bu maddeler gerektirmektedir.

XCALIBRE programı ile gerçek puanlayıcılar ve %10, %20 ve %33 test veri oranları ile BLSTM yöntemi kullanılarak gerçekleştirilen otomatik puanlama işlemi sonrası A_1 ve B_1 testlerine ilişkin test karakteristik eğrileri ve test bilgi fonksiyonu grafikleri elde edilmiştir. Bu grafiklerin tamamı EK-Ş'de gösterilmektedir. XCALIBRE programında kestirilen yetenek parametreleri ve madde parametreleri IRTEQ programına aktarılarak test eşitleme işlemi gerçekleştirilmiştir.

KTK ve MTK eşitlemeleri sonrası sonuçlar eşitlemenin standart hatası (standart error of equating-SEE), yanlılık (bias-BIAS) ve hata kareleri ortalamasının karekökü (root mean squared error-RMSE) kullanılarak değerlendirilmiştir. Eşitlemenin standart hatası yani seçkisiz hata eşitlenmiş puanların standart sapması üzerine tasarlanmış olup örneklemden kaynaklanmaktadır. Yanlılık yani sistematik hata tahmin edilen eşitleme ve kriter (gerçek) eşitleme ilişkisi arasındaki farka dayalıdır. Yanlılık, denk olmayan gruplarda ortak madde deseninde ortak maddelerin içerik ve istatistiksel özellikler bakımından test formunu temsil etmemesinden, gruplar arasındaki ciddi farklılıklardan ve ortak maddelerin bir uygulamadan diğerine fark etmesinden kaynaklanmaktadır. Yanlılık örneklemden doğrudan etkilenen bir katsayı değildir. Hata kareleri ortalamasının karekökü ise yanlılık ve standart hatanın birleşimidir (Kolen ve Brennan, 2014; LaFlair, Isbell, May, Arvizu ve Jamieson, 2017). SEE, BIAS ve RMSE değerleri; KTK'daki eşitleme işlemi sonrası "equate" (Albano, 2016) paketi aracılığıyla, MTK'daki eşitleme işlemi sonrası Microsoft Office programı içerisinde bulunan MSEXCEL modülü aracılığıyla hesaplanmıştır. Aynı hata katsayıları seçilerek KTK ve MTK yöntemlerinin karşılaştırılması sağlanmıştır. KTK ile karşılaştırmanın daha kolay olabilmesi için

MTK'daki hataların hesaplanmasında yetenekler (theta) kullanılmıştır. Aşağıda, BIAS (eşitlik 38), RMSE (eşitlik 39) ve SEE'nin (eşitlik 40) KTK'da hesaplanmasında kullanılan denklemlere yer verilmektedir (Gonzalez ve Wiberg, 2017). *L* gerçekleştirilen bootstrap sayısını, *l* örneklemi, $\hat{\varphi}(x_i)$ tahmin edilen eşitlenmiş puanları, $\varphi(x_i)$ gerçek eşitlenmiş puanları, $\bar{\hat{\varphi}}(x_i)$ tahmin edilen eşitlenmiş puan ortalamalarını göstermek üzere;

$$BIAS(x_i) = \frac{1}{L}\sum_{l=1}^{L}[(\hat{\varphi}_1(x_i) - \varphi_1(x_i)] \quad (38)$$

$$RMSE(x_i) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}[(\hat{\varphi}_1(x_i) - \varphi_1(x_i)]^2} \quad (39)$$

$$SEE(x_i) = \sqrt{RMSE(x_i)^2 - BIAS(x_i)^2} \quad (40)$$

MTK'ya dayalı olarak SEE (eşitlik 41), BIAS (eşitlik 42) ve RMSE (eşitlik 43) değerlerini hesaplarken aşağıdaki denklemlerden yararlanılabilir. Denklemler yazılırken Deng ve Monfils (2017), Keller ve Keller (2011) kaynaklarından yararlanılmıştır. Θ_i i bireyinin sahip olduğu yeteneği, $\hat{\Theta}_i$ i bireyinin kullanılan eşitleme yöntemi ile kestirilen yeteneğini, *N* kişi sayısını göstermek üzere;

$$SEE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\Theta}_i - \Theta_i - BIAS)^2} \quad (41)$$

$$BIAS = \frac{1}{N}\sum_{i=1}^{N}(\hat{\Theta}_i - \Theta_i) \quad (42)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\Theta}_i - \Theta_i)^2} \quad (43)$$

Eşitleme hataları üç otomatik puanlama koşulu için elde edildikten sonra gerçek puanlayıcılar ile karşılaştırılmıştır. Ardından puanlayıcı türünün eşitleme işlemine ait hatalarda (RMSE) anlamlı farklılık oluşturma durumunu belirlemek üzere fark testi gerçekleştirilmesine karar verilmiştir. Bu doğrultuda otomatik puanlamaya ilişkin üç koşulun ortalaması alınmıştır. Ardından her bir grup için normallik test edilmiştir. Normallik test edilirken Shapiro Wilks testinden yararlanılmıştır. Normallik testi sonuçları gerçek puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemine ait RMSE değerlerinin (W(sd=13)=,860, p=,038<,05) normal dağılmadığını, otomatik puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemine ait RMSE değerlerinin ise (W(sd=13)=,914, p=,210>,05) normal dağıldığını göstermiştir. Sonuçta gruplardan birisi için

normalliğin sağlanmaması nedeniyle fark testi nonparametrik bir teknik olan Mann Whitney U testi ile gerçekleştirilmiştir. Puanlama türünün RMSE üzerindeki etkisini belirlemek amacıyla Cohen (1988) tarafından sunulan *r* katsayısı aracılığıyla etki büyüklüğü hesaplanmıştır. Aşağıda r katsayısını hesaplamak için kullanılacak denkleme eşitlik 44'de yer verilmektedir. Bu denklemde Z katsayısı Mann Whitney U testi sonucunda elde edilen Z değerini, n örneklemde bulunan kişi sayısını göstermektedir (Fritz, Morris ve Richler, 2012).

$$r = \frac{Z}{\sqrt{n}} \qquad (44)$$

Etki büyüklüğü hesaplandıktan sonra gerçek puanlayıcılar ve otomatik puanlama aracılığıyla gerçekleştirilen eşitleme işlemine ait hatalar (RMSE) arasındaki korelasyon incelenmiştir. Normallik testleri sonucuna göre değişkenlerden birisinin normallik şartını sağlamaması nedeniyle ilişki Spearman sıra farkları korelasyonu kullanılarak incelenmiştir.

Bölüm 4
Bulgular ve Yorumlar

Araştırmanın bu bölümünde sırasıyla;

i) gerçek puanlayıcılar ile gerçek puanlayıcıların üzerinde anlaştıkları puanlar arasındaki uyuma ve otomatik puanlama yöntemleri ile gerçek puanlayıcıların üzerinde anlaştıkları puanlar arasındaki uyum katsayılarına,

ii) otomatik puanlama koşulları aracılığıyla elde edilen puanların eşitlenmesine ilişkin hatalara ve gerçek puanlayıcıların üzerinde anlaştıkları puanlar aracılığıyla puanların eşitlenmesine ilişkin hatalara, otomatik puanlama koşullarına ait eşitleme hataları ortalamalarının gerçek puanlayıcı eşitleme hatalarından anlamlı farklılık gösterme durumlarına yer verilmiştir.

1. Gerçek Puanlayıcılar ve Otomatik Puanlamaya Yönelik Uyum Katsayıları

Tablo 18, A_1 kitapçığında yer alan yapılandırılmış cevap maddelerine ilişkin gerçek puanlayıcılar ve gerçek puanlayıcıların üzerinde anlaştıkları puanlar arasındaki uyumu, eğitim ve test veri oranının değişmesi durumunda gerçek puanlayıcıların üzerinde anlaştıkları puanlar ile otomatik puanlama yöntemleri arasındaki uyumun değişimini göstermektedir. Tablo 18 incelendiğinde A_1 kitapçığında yer alan yapılandırılmış cevap maddelerine ilişkin olarak hesaplanan üç uyum katsayısı görülmektedir. Uyum katsayıları öncelikle gerçek puanlayıcı grubu 1 ile nihai puanlar ve gerçek puanlayıcı grubu 2 ile nihai puanlar arasında hesaplanmıştır. Ardından otomatik puanlama sürecinde kullanılan beş farklı yönteme göre elde edilen puanlar ile nihai puanlar arasındaki uyum test veri oranlarına göre bulunmuştur. Çapraz geçerlik kullanılmasıyla uyum katsayıları A_1 kitapçığı için 607 veri üzerinden elde edilmiştir. A_1 kitapçığında yer alan maddelerden örnek iki maddenin yorumuna metnin ilerleyen bölümünde yer verilmiştir. A_1 kitapçığında bulunan diğer maddelere ilişkin sonuçlar tablo 18 üzerinden incelenebilir. Tablo 18'de her bir uyum katsayısı türüne göre en yüksek uyum değerine sahip üç katsayı koyu, en düşük uyum değerine sahip üç katsayı ise italik olarak gösterilmiştir.

Tablo 18

A_1 Kitapçığındaki Yapılandırılmış Cevap Maddelerine Yönelik Gerçek Puanlayıcılar ve Otomatik Puanlama Yöntemleri ile Nihai Puanlar Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcılar ile Nihai Puanlar Arası Uyum			Test verisi seçim yöntemi	Otomatik Puanlama Yöntemleri ile Nihai (Gerçek Puanlaycıların Üzerinde Anlaştıkları) Puanlar Arası Uyum														
					SVM			LR			MNB			LSTM			BLSTM		
		UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK
Madde 2	P_1-P_N	,980	,976	,880	ÇG %10	,914	*,904*	,226	,919	,910	,223	,921	,908	,448	*,913*	*,904*	*,061*	**,941**	**,931**	**,569**
	P_2-P_N	,979	,975	,862	ÇG %20	,916	,906	,212	,923	,914	,273	*,913*	*,899*	,347	,916	,907	,147	**,942**	**,933**	**,593**
					ÇG %33	*,909*	*,899*	*,128*	,921	,912	,208	,918	,906	,337	*,911*	*,903*	*,000*	**,934**	**,924**	**,522**
Madde 7*	P_1-P_N	,979	,970	,974	ÇG %10	,845	,782	,862	,822	,752	,836	,735	,642	*,720*	*,720*	*,629*	*,683*	**,881**	**,833**	**,884**
					ÇG %20	**,855**	**,796**	,859	,815	,743	,832	,731	,639	*,720*	,735	,647	,744	**,881**	**,833**	**,892**
	P_2-P_N	,970	,958	,971	ÇG %33	,827	,756	,825	,822	,752	,832	*,722*	*,625*	*,705*	*,728*	*,638*	,726	**,875**	**,823**	**,877**
Madde 8*	P_1-P_N	,997	,995	,997	ÇG %10	,928	,894	,910	,936	,906	**,915**	,896	,849	,859	*,779*	*,687*	*,701*	**,957**	**,937**	**,937**
					ÇG %20	,936	,906	**,917**	,931	,899	,911	,901	,856	,868	*,776*	*,683*	*,684*	**,946**	**,921**	,899
	P_2-P_N	,987	,981	,985	ÇG %33	,931	,899	,909	,931	,899	,896	,875	,819	,839	*,771*	*,676*	*,672*	**,942**	**,916**	,912
Madde 10*	P_1-P_N	,944	,891	,885	ÇG %10	,837	,682	,665	**,845**	**,699**	**,681**	,827	,667	*,641*	,840	,688	,672	**,863**	**,733**	**,720**
					ÇG %20	,840	,689	,672	**,842**	,693	**,675**	,835	,681	,660	,829	,662	,652	**,842**	**,695**	,673
	P_2-P_N	,947	,897	,892	ÇG %33	*,817*	*,642*	*,626*	*,819*	*,649*	*,626*	,830	,673	,648	*,824*	*,657*	*,637*	,835	,680	,660
Madde 11*	P_1-P_N	,985	,972	,968	ÇG %10	,870	,755	,723	,875	,769	,726	*,843*	*,720*	*,648*	,924	,860	,835	**,956**	**,917**	**,904**
					ÇG %20	,873	,761	,730	,881	,779	,744	*,835*	*,708*	*,626*	,934	,879	,855	**,962**	**,929**	**,918**
	P_2-P_N	,985	,972	,968	ÇG %33	,871	,757	,727	,865	,748	,708	*,825*	*,693*	*,600*	,870	,759	,717	**,946**	**,898**	**,883**

Tablo 18 (Devam)

A_1 Kitapçığındaki Yapılandırılmış Cevap Maddelerine Yönelik Gerçek Puanlayıcılar ve Otomatik Puanlama Yöntemleri ile Nihai Puanlar Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcılar ile Nihai Puanlar Arası Uyum			Test verisi seçim yöntemi	Otomatik Puanlama Yöntemleri ile Nihai (Gerçek Puanlaycıların Üzerinde Anlaştıkları) Puanlar Arası Uyum														
					SVM			LR			MNB			LSTM			BLSTM		
		UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK
Madde 14	P_1-P_N	,975	,959	,937	ÇG %10	,901	,839	,744	,911	,857	,764	,890	,828	,695	*,792*	*,709*	*,318*	**,929**	**,884**	**,818**
					ÇG %20	,895	,829	,724	,904	,847	,747	,881	,817	,667	*,873*	*,807*	*,635*	**,928**	**,880**	**,816**
	P_2-P_N	,969	,948	,921	ÇG %33	,893	,825	,725	,906	,849	,752	,876	,811	,646	*,792*	*,710*	*,315*	**,916**	**,864**	**,781**
Madde 15*	P_1-P_N	,972	,960	,971	ÇG %10	,708	,585	,683	,720	**,603**	,686	,687	,563	,613	*,560*	*,428*	*,224*	**,766**	**,666**	**,714**
					ÇG %20	,717	,595	,678	,712	,593	,664	,672	,544	,589	*,539*	*,415*	*,137*	**,740**	**,628**	**,707**
	P_2-P_N	,960	,943	,943	ÇG %33	,677	,539	,656	,690	,562	,625	,680	,557	,564	*,516*	*,397*	*,000*	**,741**	**,628**	**,711**
Madde 18	P_1-P_N	,997	,995	,997	ÇG %10	,956	,937	**,952**	,924	,893	,914	,867	,811	,790	*,718*	*,616*	*,517*	**,970**	**,958**	**,961**
					ÇG %20	,941	,916	,937	,921	,888	,904	,868	,813	,796	*,761*	*,672*	*,599*	**,965**	**,951**	**,952**
	P_2-P_N	,998	,998	,994	ÇG %33	,924	,893	,912	,923	,891	,906	,863	,807	,756	*,671*	*,544*	*,515*	**,960**	**,944**	**,947**
Madde 19	P_1-P_N	,997	,996	,997	ÇG %10	,919	,892	,900	**,936**	**,915**	**,918**	,815	,752	,807	*,802*	*,739*	*,749*	**,939**	**,918**	**,922**
					ÇG %20	,914	,886	,897	,931	,908	,909	,822	,762	,820	*,797*	*,736*	*,720*	**,937**	**,916**	**,936**
	P_2-P_N	,995	,993	,996	ÇG %33	,918	,890	,904	,921	,895	,899	,820	,760	,800	*,778*	*,719*	*,624*	,919	,891	**,918**

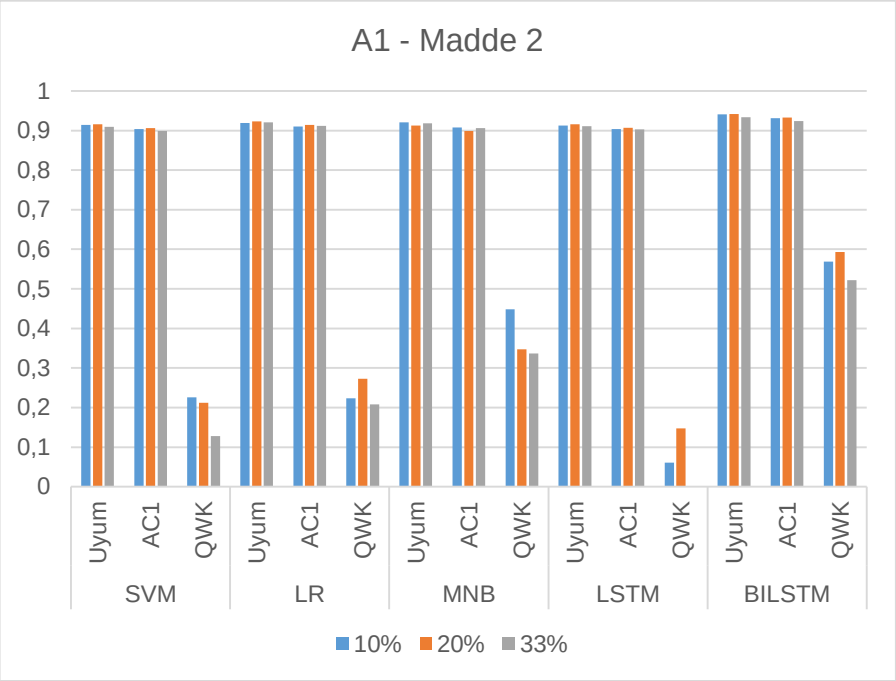
* A_1 ve B_1 kitapçıklarında yer alan ortak maddeleri göstermektedir.

Not 1: P_1: İlk puanlayıcı grubu puanları, P_2: İkinci puanlayıcı grubu puanları, Pn: Nihai puanlar anlamına gelmektedir.

Not 2: UY uyum yüzdesini, AC1 Gwet'in AC1 katsayısını, QWK karesel ağırlıklı Kappa değerini göstermektedir.

Not 3: ÇG: Çapraz geçerlik anlamına gelmekte, %10, %20 ve %33 ise test veri oranlarını göstermektedir.

Şekil 6 otomatik puanlama yöntemleri ve test veri oranlarına göre madde 2 için elde edilen uyum değerlerini göstermektedir. Şekil 6 ve tablo 18 dikkate alınarak madde 2 için yapılan yorumlara şekil 6'nın ardından yer verilmektedir.



Şekil 6. Otomatik puanlama yöntemleri ve test veri oranlarına göre A1 kitapçığı madde 2'ye ilişkin uyum değerlerini gösteren grafik

Tablo 18'de yer alan madde 2'ye ait değerler incelendiğinde gerçek puanlayıcılardan oluşan ilk puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesinin ,980, AC1 indeksinin ,976, QWK değerinin ise ,880 olduğu görülmektedir. Gerçek puanlayıcılardan oluşan ikinci puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesi ,979, AC1 indeksi ,975 ve QWK değeri ise ,862 olarak bulunmuştur. Otomatik puanlama ile nihai puanlar arasındaki uyum incelendiğinde %10 test veri oranı için en yüksek uyum yüzdesi BLSTM yöntemi ile ,941 olarak elde edilmiştir ve bu değeri ,921 ile MNB yöntemi izlemektedir. En düşük uyum yüzdesi ise ,913 ile LSTM yönteminde elde edilmiştir. Uyum yüzdeleri genel olarak incelendiğinde değerlerin birbirine yakın ve kabul edilebilir düzeyde (>,80) olduğu sonucuna ulaşılmaktadır. AC1 indeksi incelendiğinde en yüksek uyumun elde edildiği yöntem ,931 ile BLSTM yöntemi olup bu yöntemi ,910 ile LR yöntemi izlemektedir. En düşük AC1 değeri ise ,904 ile SVM ve LSTM yöntemlerine aittir. AC1 değerlerinin tüm yöntemler için birbirine yakın olduğu ve çok iyi uyum gösterdiği (>,80) gözlemlenmiştir. %10 test veri oranında QWK değeri en yüksek BLSTM

yöntemi ile ,569 olarak bulunmuş olup bu değeri ,448 ile MNB yöntemi izlemektedir. En düşük QWK değeri ,061 ile LSTM yöntemine aittir ve bu değeri ,223 ile LR yöntemi izlemektedir. QWK değerlerinin %10 test veri oranında yöntemler arasında oldukça değişkenlik gösterdiği, ranjının ,508 olduğu ve AC1 indeksi ile uyum yüzdesinden ayrıştığı sonucuna ulaşılmıştır. QWK değeri genel olarak değerlendirildiğinde BLSTM ve MNB yönteminin orta düzeyde (<,60 Λ >,40), LR ve SVM yönteminin (<,40 Λ >,20) kayda değer, LSTM yönteminin ise zayıf (<,20) uyum gösterdiği belirtilebilir.

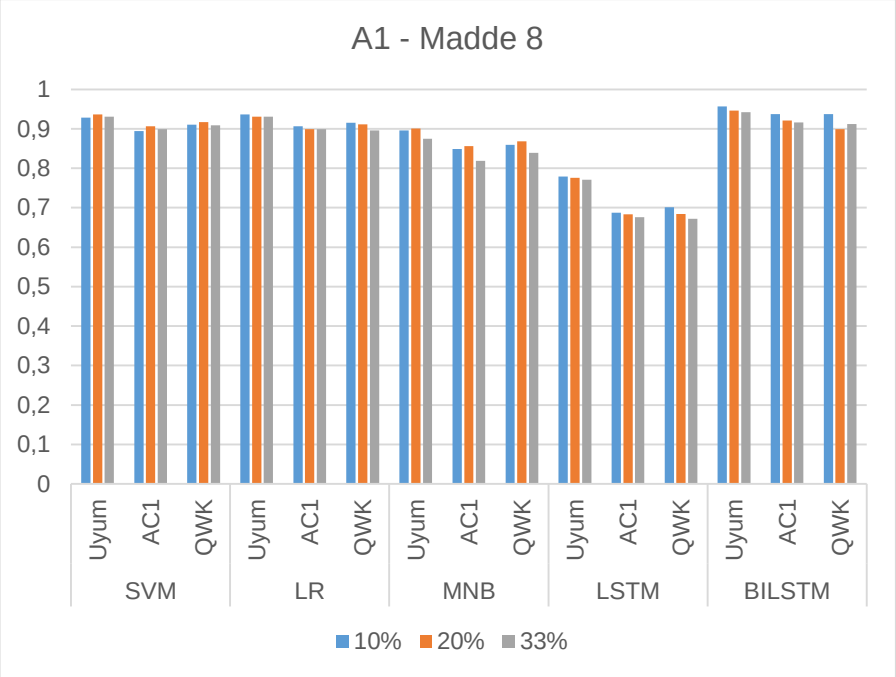
%20 test veri oranında en yüksek uyum yüzdesini ,942 ile BLSTM yöntemi göstermiş olup en düşük uyum yüzdesini ise ,913 ile MNB yöntemi göstermiştir. Tüm yöntemlerde uyum yüzdelerinin birbirine oldukça yakın ve kabul edilebilir düzeyde (>,80) olduğu görülmektedir. AC1 indeksi açısından uyum değerlendirildiğinde en yüksek uyuma ,933 ile BLSTM yönteminde rastlanmış olup en düşük uyuma ise ,899 ile MNB yönteminde rastlanmıştır. AC1 indeksi değerlerinin %20 test veri oranında genel olarak yakın olduğu ve tamamının çok iyi uyum gösterdiği (>,80) belirtilebilir. %20 test veri oranı için QWK değerleri incelendiğinde en yüksek uyuma sahip yöntemin ,593 ile BLSTM yöntemi olduğu en düşük uyuma sahip yöntemin ise ,147 ile LSTM yöntemi olduğu belirtilebilir. En düşük uyuma sahip ikinci yöntem ise ,212 ile SVM'dir. Görüldüğü üzere %20 test veri oranında, %10 test veri oranına benzer şekilde QWK değerleri düşük ve yöntemler arasında farklılık olacak şekilde elde edilmiştir. %20 test veri oranında QWK değerlerinin ranjı ,446'dır. QWK değerleri genel olarak incelendiğinde BLSTM yönteminin orta düzeyde uyuma (<,60 Λ >,40), MNB, LR ve SVM yöntemlerinin kayda değer uyuma (<,40 Λ >,20), LSTM yönteminin ise zayıf uyuma (<,20) işaret ettiği görülmektedir.

%33 test veri oranı için uyum yüzdesi en yüksek yöntem ,934 ile BLSTM yöntemidir. En düşük uyum yüzdesine sahip yöntem ise ,909 ile SVM yöntemidir. %33 test veri oranında genel olarak uyum yüzdeleri yüksek, yakın ve kabul edilebilir (>,80) bulunmuştur. AC1 indeksleri genel olarak yüksek bulunmakla birlikte en yüksek uyum ,924 ile BLSTM, en düşük uyum ,899 ile SVM yöntemine aittir. Tüm yöntemler için elde edilen değerler birbirine yakın olup çok iyi uyum (>,80) göstermektedir. QWK değerleri değerlendirildiğinde en yüksek uyum ,522 ile BLSTM yönteminde en düşük iki uyum ise ,128 ile SVM, ,000 ile LSTM yönteminde elde edilmiştir. %33 test veri oranı için elde edilen tüm QWK değerleri düşüktür ve

yöntemler arasında fazla değişkenlik göstermektedir. %33 test veri oranında QWK değerlerinin ranjı ,522'dir. Elde edilen değerler ele alındığında BLSTM yönteminin orta düzeyde uyuma (<,60 Λ >,40), MNB ve LR yöntemlerinin kayda değer uyuma (<,40 Λ >,20), LSTM ve SVM yöntemlerinin ise zayıf uyuma (<,20) sahip olduğu görülmektedir.

Şekil 6 incelendiğinde görülebileceği gibi tüm koşullarda (test veri oranları ve yöntemler) QWK değerleri AC1 ve uyum yüzdesi değerlerinden oldukça düşüktür. QWK değerlerinin tamamında görülen bu düşüklüğün ve değerin bazı yöntemler ve test veri oranlarında ,000'a kadar yaklaşmasının sebebi yaygınlık katsayısı olmakla birlikte elde edilen QWK katsayıları yöntemler arasında da oldukça farklılık göstermiştir. Bu nedenle araştırma sonuçlarında QWK değerleri değerlendirilmemiştir. QWK değerinin madde 2 için düşük bulunması araştırmada öngörülen durumlardan birisidir. %10, %20 ve %33 test veri oranı için tüm otomatik puanlama yöntemleri dikkate alınarak karşılaştırma yapıldığında uyum değerlerinin %20 test veri oranında biraz daha yüksek, %33 test veri oranında ise biraz daha düşük olduğu gözlemlenmiştir. Ancak aralarındaki farklar oldukça küçüktür. Uyum yüzdeleri tüm koşullarda kabul edilebilir sınır olan ,80'in üzerindedir. AC1 indeksi ise tüm koşullarda (>,80) çok iyi uyuma işaret etmektedir. AC1 değerlerinin Kappa katsayısı ile aynı yönde değerlendirildiği dikkate alındığında elde edilen tüm AC1 katsayıları Williamson, Xi ve Breyer (2012) tarafından otomatik puanlama ile gerçek puanlayıcılar arasında bulunması gereken uyum değerinin ,70'den yüksek olması kriterine de uymaktadır. Tablo 18 incelendiğinde görülebileceği gibi madde 2 için tüm test veri oranları ve yöntemler dikkate alındığında en yüksek uyum yüzdesi (,942) ve en yüksek AC1 değeri (,933) BLSTM yönteminde ve %20 test veri oranında elde edilmiştir. Elde edilen bu değerler gerçek puanlayıcı grupları ile nihai puanlar arasındaki uyum yüzdesi ve AC1 değerine yakındır. Madde 2'de karşılaşılan yaygınlık sorunu nedeniyle gerçek puanlayıcılar ile nihai puanlar arasında hesaplanan QWK değerleri de düşüktür. Bu durum makine öğrenmesine daha olumsuz şekilde yansımıştır.

Şekil 7 otomatik puanlama yöntemleri ve test veri oranlarına göre madde 8 için elde edilen uyum değerlerini göstermektedir. Şekil 7 ve tablo 18 dikkate alınarak madde 8 için yapılan yorumlara şekil 7'nin ardından yer verilmektedir.



Şekil 7. Otomatik puanlama yöntemleri ve test veri oranlarına göre A_1 kitapçığı madde 8'e ilişkin uyum değerlerini gösteren grafik

Tablo 18'de yer alan madde 8'e ait değerler incelendiğinde gerçek puanlayıcılardan oluşan ilk puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesinin ,997, AC1 indeksinin ,995, QWK değerinin ise ,997 olduğu görülmektedir. Gerçek puanlayıcılardan oluşan ikinci puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesi ,987, AC1 indeksi ,981 ve QWK değeri ise ,985 olarak bulunmuştur. Otomatik puanlama ile nihai puanlar arasındaki uyum incelendiğinde %10 test veri oranı için en yüksek uyum yüzdesi BLSTM yöntemi ile ,957 olarak elde edilmiştir. Bu uyum yüzdesi değerini ,936 ile LR yöntemi takip etmektedir. En düşük uyum yüzdesi ise ,779 ile LSTM yönteminde elde edilmiş olup bu yöntemi ,896 ile MNB yöntemi izlemektedir. Uyum yüzdeleri genel olarak incelendiğinde SVM, LR, MNB ve BLSTM yöntemleri için kabul edilebilir değerlere (>,80) ulaşıldığı görülmektedir. AC1 indeksi incelendiğinde en yüksek uyumun elde edildiği yöntem ,937 ile BLSTM yöntemidir. En düşük AC1 değeri ise ,687 ile LSTM yöntemine ait olup bu yöntemi ,849 ile MNB yöntemi izlemektedir. AC1 değerlerinin BLSTM, LR, MNB ve SVM yöntemleri için çok iyi (>,80), LSTM yöntemi için iyi uyuma (>,60 Λ <,80) işaret ettiği bulunmuştur. %10 test veri oranında QWK değeri en yüksek BLSTM yöntemi ile ,937 olarak bulunmuş olup bu değeri ,915 ile LR yöntemi izlemektedir. En düşük QWK değeri ,701 ile LSTM yöntemine aittir. QWK

değerlerinin %10 test veri oranında AC1 indekslerinden büyük olduğu sonucuna ulaşılmıştır. QWK değeri SVM, LR, MNB ve BLSTM yöntemleri için çok iyi (>,80), LSTM yöntemi için iyi uyum (>,60 Λ <,80) göstermektedir.

%20 test veri oranında en yüksek uyum yüzdesini ,946 ile BLSTM yöntemi göstermiş olup en düşük uyum yüzdesini ise ,776 ile LSTM yöntemi göstermiştir. Uyum yüzdesine göre BLSTM, LR, MNB ve SVM yöntemleri kabul edilebilir uyum (>,80) gösterirken LSTM yöntemi kabul edilebilir uyum göstermemiştir. AC1 indeksi açısından uyum değerlendirildiğinde en yüksek uyuma ,921 ile BLSTM yönteminde rastlanmış olup en düşük uyuma ise ,683 ile LSTM yönteminde rastlanmıştır. AC1 indeksi değerlerinin %20 test veri oranında BLSTM, LR, MNB ve SVM yöntemi için çok iyi (>,80), LSTM yöntemi için ise iyi uyum (<,80 Λ >,60) gösterdiği belirtilebilir. %20 test veri oranı için QWK değerleri incelendiğinde en yüksek uyuma sahip yöntemin ,917 ile SVM yöntemi olduğu en düşük uyuma sahip yöntemin ise ,684 ile LSTM yöntemi olduğu belirtilebilir. En düşük uyuma sahip ikinci yöntem ise ,868 ile MNB'dir. Görüldüğü üzere %20 test veri oranında QWK değerleri LSTM için iyi (<,80 Λ >,60); BLSTM, LR, MNB ve SVM yöntemleri için çok iyi uyuma (>,80) işaret etmektedir.

%33 test veri oranı için uyum yüzdesi en yüksek yöntem ,942 ile BLSTM yöntemidir. En düşük uyum yüzdesine sahip yöntem ise ,771 ile LSTM yöntemidir. %33 test veri oranında LSTM yöntemi dışında tüm yöntemlerde uyum yüzdesi kabul edilebilir (>,80) bulunmuştur. AC1 indeksleri incelendiğinde en yüksek uyum ,916 ile BLSTM yöntemine aittir. En düşük uyum ise ,676 ile LSTM yöntemine aittir ve bu yöntemi ,819 ile MNB yöntemi izlemektedir. AC1 indeksleri ele alındığında BLSTM, MNB, LR ve SVM yöntemleri için çok iyi uyuma (>,80), LSTM yöntemi için ise iyi uyuma (<,80 Λ >,60) ulaşıldığı görülmektedir. QWK değerleri değerlendirildiğinde en yüksek uyum ,912 ile BLSTM yönteminde, en düşük iki uyum ise ,672 ile LSTM, ,839 ile MNB yönteminde elde edilmiştir. %33 test veri oranı için elde edilen QWK değerleri BLSTM, LR, MNB ve SVM yöntemleri için çok iyi (>,80); LSTM yöntemi için iyi uyuma (<,80 Λ >,60) işaret etmektedir.

Şekil 7 incelendiğinde görülebileceği gibi tüm koşullarda LSTM yöntemine ait uyum katsayıları diğer yöntemlere ait uyum katsayılarından daha düşüktür. Tüm koşullar dikkate alındığında QWK değeri BLSTM, LR, MNB ve SVM yöntemleri için çok iyi (>,80), LSTM yöntemi için iyi uyum göstermiştir (>,60 Λ <,80). Tüm koşullarda

AC1 değerleri BLSTM, LR, MNB ve SVM yöntemleri için çok iyi (>,80), LSTM yöntemi için iyi uyum (>,60 Λ <,80) bulunduğunu göstermektedir. Tüm koşullar dikkate alındığında uyum yüzdesi BLSTM, LR, MNB ve SVM için kabul edilebilir (>,80) değerler göstermiştir. Williamson, Xi ve Breyer (2012) tarafından gerçek puanlayıcılar ile otomatik puanlama arasındaki Kappa uyum katsayısının en az ,70 olması kriterine göre %20 ve %33 test verisi oranlarında LSTM yöntemi haricinde tüm QWK değerleri kabul edilebilir bulunmuştur. Aynı kriter AC1 katsayısı için kullanıldığında LSTM yönteminde %10, %20 ve %33 test verisi oranlarında gerekli değere ulaşılmamış olup diğer tüm yöntemlerde ve test veri oranlarında kabul edilebilir değerlere ulaşılmıştır. Tablo 18 incelendiğinde görülebileceği gibi madde 8 için tüm test veri oranları ve yöntemler dikkate alındığında en yüksek uyum yüzdesi (,957), AC1 değeri (,937) ve QWK katsayısı (,937) BLSTM yönteminde %10 test veri oranında elde edilmiştir. %10 test veri oranı A_1 kitapçığını yanıtlayan 607 öğrenciden 546'sına ait verinin sistemi eğitmek amacıyla kullanıldığı anlamına gelmektedir. Elde edilen bu değerler (UY=,957, AC1=,937 ve QWK=,937) gerçek puanlayıcı grupları ile nihai puanlar arasındaki uyum yüzdesi, AC1 ve QWK değerlerine yakın bulunmuştur.

Tablo 19, B_1 kitapçığında yer alan yapılandırılmış cevap maddelerine ilişkin gerçek puanlayıcılar ile gerçek puanlayıcıların üzerinde anlaştıkları puanlar arasındaki uyumu, test veri oranının değişmesi durumunda gerçek puanlayıcıların üzerinde anlaştıkları puanlar ile otomatik puanlama yöntemleri arasındaki uyumun değişimini göstermektedir. B_1 kitapçığındaki madde incelemeleri A_1 kitapçığına benzer olarak yapılmıştır. Çapraz geçerlik kullanılarak uyum katsayıları B_1 kitapçığı için 584 veri üzerinden hesaplanmıştır. B_1 kitapçığında yer alan maddelerden örnek iki maddenin yorumuna metnin ilerleyen bölümünde yer verilmiştir. B_1 kitapçığında bulunan diğer maddelere ilişkin sonuçlar tablo 19 üzerinden incelenebilir. Tablo 19'da her bir uyum katsayısı türüne göre en yüksek uyum değerine sahip üç katsayı koyu, en düşük uyum değerine sahip üç katsayı ise italik olarak gösterilmiştir.

Tablo 19

B_1 Kitapçığındaki Yapılandırılmış Cevap Maddelerine Yönelik Gerçek Puanlayıcılar ve Otomatik Puanlama Yöntemleri Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcılar Arası Uyum			Test verisi seçim yöntemi	Otomatik Puanlama Yöntemleri ile Gerçek Puanlaycıların Üzerinde Anlaştıkları Puanlar Arası Uyum														
					SVM			LR			MNB			LSTM			BLSTM		
		UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK
Madde 3	P1-PN	,966	,952	,877	ÇG %10	,911	,879	,665	,913	,882	,667	*,906*	*,871*	,653	,913	,880	,678	**,923**	**,894**	**,719**
	P2-PN	,973	,962	,900	ÇG %20	,914	,883	,683	,911	,879	,665	*,904*	*,869*	*,642*	**,921**	**,891**	**,716**	,913	,879	,686
					ÇG %30	**,916**	**,885**	**,688**	*,906*	,872	*,644*	*,901*	*,865*	*,623*	,911	,878	,671	,911	,878	,671
Madde 5*	P1-PN	,971	,960	,972	ÇG %10	,866	,818	,884	,836	,778	,864	*,779*	*,710*	*,740*	,836	,778	,861	**,918**	**,888**	**,925**
	P2-PN	,979	,972	,979	ÇG %20	,863	,814	,882	,837	,781	,855	*,781*	*,712*	*,743*	,825	,766	,846	**,902**	**,866**	**,913**
					ÇG %30	,870	,823	,878	,844	,790	,866	,786	,720	*,744*	*,784*	*,718*	,783	**,892**	**,853**	**,904**
Madde 6*	P1-PN	,991	,988	,981	ÇG %10	,942	,915	,909	**,954**	**,933**	**,924**	,884	,833	,861	*,740*	*,628*	*,654*	**,959**	**,940**	**,939**
	P2-PN	,993	,990	,995	ÇG %20	,945	,920	,919	,947	,923	,915	,873	,819	,848	*,752*	*,649*	*,645*	,949	,925	,923
					ÇG %30	,937	,908	,916	,947	,923	,906	,846	,781	,832	*,719*	*,593*	*,682*	**,952**	**,930**	**,926**
Madde 8*	P1-PN	,950	,902	,899	ÇG %10	,827	,659	,649	,818	,645	,629	,820	,649	,632	,834	,673	,663	**,854**	**,713**	**,704**
	P2-PN	,957	,916	,913	ÇG %20	,812	,629	,618	*,800*	*,608*	*,591*	,832	,673	,656	*,805*	,618	,601	**,858**	**,719**	**,713**
					ÇG %30	,820	,646	,634	*,793*	*,593*	*,578*	,827	,662	,646	*,793*	*,590*	*,582*	**,842**	**,691**	**,679**
Madde 9*	P1-PN	,985	,971	,967	ÇG %10	,846	,711	,670	*,836*	,696	,642	*,796*	*,637*	*,538*	**,877**	**,772**	**,732**	**,885**	**,788**	**,751**
	P2-PN	,993	,987	,985	ÇG %20	,844	,706	,668	,844	,714	,658	*,796*	*,641*	*,533*	,873	,767	,722	**,882**	**,779**	**,746**
					ÇG %30	,849	,716	,679	*,837*	,698	,647	*,796*	*,643*	*,531*	,868	,760	,707	,872	,766	,716

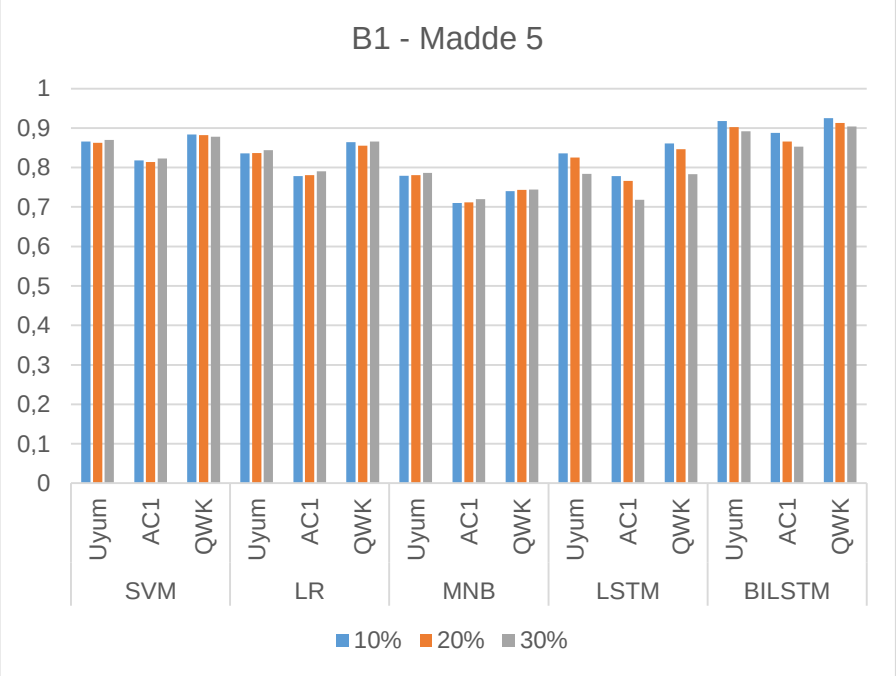
Tablo 19 (Devam)

B_1 Kitapçığındaki Yapılandırılmış Cevap Maddelerine Yönelik Gerçek Puanlayıcılar ve Otomatik Puanlama Yöntemleri Arasındaki Uyum Katsayıları

Madde Kodu	Gerçek Puanlayıcılar Arası Uyum			Test verisi seçim yöntemi	Otomatik Puanlama Yöntemleri ile Gerçek Puanlaycıların Üzerinde Anlaştıkları Puanlar Arası Uyum														
					SVM			LR			MNB			LSTM			BLSTM		
		UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK	UY	AC1	QWK
Madde 11	P1-PN	,986	,981	,987	ÇG %10	,918	,886	,912	,911	,876	,902	*,861*	*,807*	*,867*	,882	,838	,887	**,940**	**,916**	**,925**
	P2-PN	,990	,986	,989	ÇG %20	,902	,865	,893	,913	,878	,900	*,863*	*,810*	*,863*	*,878*	,833	,880	**,943**	**,920**	**,929**
					ÇG %30	,904	,867	,901	,914	,881	,899	*,861*	*,808*	*,860*	,885	,843	,894	**,930**	**,901**	**,927**
Madde 12	P1-PN	,949	,923	,932	ÇG %10	,736	,606	,667	,757	,637	,719	,707	,566	*,606*	*,654*	*,490*	,663	**,793**	**,690**	**,749**
	P2-PN	,938	,908	,937	ÇG %20	,759	,640	,718	,764	,647	**,740**	,682	,528	*,559*	*,649*	*,481*	,674	**,784**	**,677**	**,741**
					ÇG %30	,755	,634	,718	,755	,635	,719	,683	,531	*,573*	*,634*	*,467*	,654	**,774**	**,662**	,738
Madde 17*	P1-PN	,974	,963	,966	ÇG %10	,707	,580	,653	,693	,565	,631	,635	,492	,522	*,541*	*,393*	*,171*	**,743**	**,634**	**,705**
	P2-PN	,978	,968	,974	ÇG %20	**,729**	**,612**	**,675**	,678	,543	,609	,610	,456	,488	*,545*	*,391*	*,302*	**,716**	**,595**	**,671**
					ÇG %30	,680	,543	,617	,700	,575	,637	,616	,471	,478	*,575*	*,430*	*,339*	,697	,567	,644
Madde 18	P1-PN	1,000	1,000	1,000	ÇG %10	*,712*	*,425*	*,429*	,748	,497	,497	,740	,480	,485	**,784**	**,568**	**,571**	**,786**	**,572**	**,572**
	P2-PN	,995	,990	,990	ÇG %20	*,711*	*,421*	*,425*	,741	,483	,483	,726	,453	,458	,759	,517	,520	,767	,535	,534
					ÇG %30	*,719*	*,439*	*,442*	,731	,463	,462	,731	,463	,466	,755	,510	,512	**,769**	**,538**	**,538**
Madde 20	P1-PN	,969	,945	,929	ÇG %10	,818	,687	,569	,817	,685	,563	*,760*	*,562*	*,471*	**,834**	,703	**,623**	**,839**	**,717**	**,627**
	P2-PN	,969	,946	,929	ÇG %20	,815	,681	,562	,830	**,708**	,597	*,750*	*,544*	*,447*	,820	,683	,585	**,837**	**,710**	**,629**
					ÇG %30	,789	,640	,495	,810	,674	,545	*,740*	*,527*	*,421*	,793	,630	,529	,820	,691	,572

* A_1 ve B_1 kitapçıklarında yer alan ortak maddeleri göstermektedir.
Not: B_1 kitapçığındaki 5. Madde A_1 kitapçığındaki 7.Madde, B_1 kitapçığındaki 6. Madde A_1 kitapçığındaki 8. Madde, B_1 kitapçığındaki 8. Madde A_1 kitapçığındaki 10. Madde, B_1 kitapçığındaki 9. Madde A_1 kitapçığındaki 11. Madde ve B_1 kitapçığındaki 17. Madde A_1 kitapçığındaki 15. Madde ile aynıdır.

Şekil 8 otomatik puanlama yöntemleri ve test veri oranlarına göre madde 5 için elde edilen uyum değerlerini göstermektedir. Şekil 8 ve tablo 19 dikkate alınarak madde 5 için yapılan yorumlara şekil 8'in ardından yer verilmektedir.



Şekil 8. Otomatik puanlama yöntemleri ve test veri oranlarına göre B_1 kitapçığı madde 5'e ilişkin uyum değerlerini gösteren grafik

Tablo 19'da yer alan madde 5'e ait değerler incelendiğinde gerçek puanlayıcılardan oluşan ilk puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesinin ,971, AC1 indeksinin ,960, QWK değerinin ise ,972 olduğu görülmektedir. Gerçek puanlayıcılardan oluşan ikinci puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesi ,979, AC1 indeksi ,972 ve QWK değeri ise ,979 olarak bulunmuştur. Otomatik puanlama ile nihai puanlar arasındaki uyum incelendiğinde %10 test veri oranı için en yüksek uyum yüzdesi BLSTM yöntemi ile ,918 olarak elde edilmiştir. Bu uyum yüzdesi değerini ,866 ile SVM yöntemi takip etmektedir. En düşük uyum yüzdesi ise ,779 ile MNB yönteminde elde edilmiştir. Uyum yüzdeleri genel olarak incelendiğinde SVM, LR, LSTM ve BLSTM yöntemleri için kabul edilebilir değerlere (>,80) ulaşıldığı görülmektedir. AC1 indeksi

incelendiğinde en yüksek uyumun elde edildiği yöntem ,888 ile BLSTM yöntemidir. En düşük AC1 değeri ise ,710 ile MNB yöntemine ait olup bu yöntemi ,778 ile LR ve LSTM yöntemleri izlemektedir. AC1 değerlerinin BLSTM ve SVM yöntemleri için çok iyi (>,80), LR, LSTM ve MNB yöntemleri için iyi uyuma (>,60 Λ <,80) işaret ettiği bulunmuştur. %10 test veri oranında en yüksek QWK değeri BLSTM yöntemi ile ,925 olarak bulunmuş olup bu değeri ,884 ile SVM yöntemi izlemektedir. En düşük QWK değeri ,740 ile MNB yöntemine aittir. QWK değerlerinin %10 test veri oranında, AC1 indekslerinden büyük olduğu sonucuna ulaşılmıştır. QWK değeri SVM, LR, LSTM ve BLSTM yöntemleri için çok iyi (>,80), MNB yöntemi için iyi uyum (>,60 Λ <,80) bulunduğunu göstermektedir.

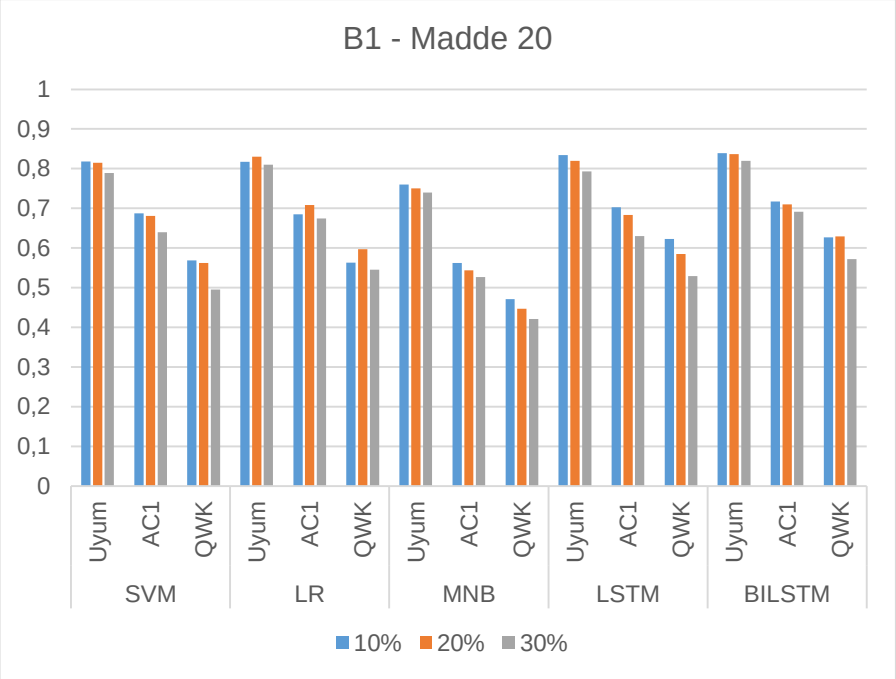
%20 test veri oranında en yüksek uyum yüzdesini ,902 ile BLSTM yöntemi göstermiş olup en düşük uyum yüzdesini ise ,781 ile MNB yöntemi göstermiştir. Uyum yüzdesine göre BLSTM, LR, LSTM ve SVM yöntemleri kabul edilebilir uyum (>,80) gösterirken MNB yöntemi kabul edilebilir uyum göstermemiştir. AC1 indeksi açısından uyum değerlendirildiğinde en yüksek uyuma ,866 ile BLSTM yönteminde rastlanmış olup en düşük uyuma ise ,712 ile MNB yönteminde rastlanmıştır. AC1 indeksi değerlerinin %20 test veri oranında BLSTM ve SVM yöntemi için çok iyi (>,80), LR, LSTM ve MNB yöntemleri için ise iyi uyum (<,80 Λ >,60) gösterdiği belirtilebilir. %20 test veri oranı için QWK değerleri incelendiğinde en yüksek uyuma sahip yöntemin ,913 ile BLSTM yöntemi olduğu, en düşük uyuma sahip yöntemin ise ,743 ile MNB yöntemi olduğu belirtilebilir. En düşük uyuma sahip ikinci yöntem ise ,846 ile LSTM'dir. Görüldüğü üzere %20 test veri oranında MNB için iyi (<,80 Λ >,60); BLSTM, LR, LSTM ve SVM yöntemleri için çok iyi uyuma (>,80) rastlanmıştır. QWK değerlerinin %20 test veri oranında AC1 indekslerinden büyük olduğu görülmektedir.

%33 test veri oranı için uyum yüzdesi en yüksek yöntem ,892 ile BLSTM yöntemidir. En düşük uyum yüzdesine sahip yöntem ise ,784 ile LSTM yöntemidir. %33 test veri oranında LSTM ve MNB yöntemi dışında tüm yöntemlerde uyum yüzdesi kabul edilebilir (>,80) bulunmuştur. AC1 indeksleri incelendiğinde en yüksek uyum ,853 ile BLSTM yöntemine aittir. En düşük uyum ise ,718 ile LSTM yöntemine aittir ve bu yöntemi ,720 ile MNB yöntemi izlemektedir. AC1 indeksleri ele alındığında BLSTM ve SVM yöntemleri için çok iyi uyuma (>,80), LR, LSTM ve MNB yöntemleri için ise iyi uyuma (<,80 Λ >,60) ulaşıldığı görülmektedir. QWK değerleri

değerlendirildiğinde en yüksek uyum ,904 ile BLSTM yönteminde en düşük iki uyum ise ,744 ile MNB, ,783 ile LSTM yönteminde elde edilmiştir. %33 test veri oranı için elde edilen QWK değerleri BLSTM, LR ve SVM yöntemleri için çok iyi (>,80); LSTM ve MNB yöntemleri için iyi uyuma (<,80 Λ >,60) işaret etmektedir. QWK değerlerinin %33 test veri oranında da AC1 indekslerinden büyük olduğu görülmektedir.

Şekil 8 incelendiğinde görülebileceği gibi tüm koşullarda MNB yöntemine ait uyum katsayıları diğer yöntemlere ait uyum katsayılarından daha düşük, BLSTM yöntemine ait uyum katsayıları ise diğer yöntemlere ait uyum katsayılarından daha yüksektir. Tüm koşullar dikkate alındığında QWK değeri BLSTM, LR ve SVM yöntemleri için tüm test veri oranlarında ve LSTM yöntemi için %10 ve %20 test veri oranında çok iyi (>,80), MNB yönteminde tüm test veri oranlarında ve LSTM yöntemi %33 test veri oranında iyi (<,80 Λ >,60) uyum göstermiştir. Tüm koşullarda AC1 değerleri BLSTM ve SVM yöntemleri için çok iyi (>,80) uyum; LR, MNB ve LSTM yöntemleri için iyi uyum (<,80 Λ >,60) bulunduğunu göstermektedir. Madde 5 için bulunan tüm AC1 katsayıları QWK katsayılarından belirgin şekilde daha düşüktür. Tüm koşullar dikkate alındığında uyum yüzdesi BLSTM, LR ve SVM için tüm test veri oranlarında; LSTM yönteminde ise %10 ve %20 test veri oranlarında kabul edilebilir (>,80) değerler göstermiştir. Williamson, Xi ve Breyer (2012) tarafından gerçek puanlayıcılar ile otomatik puanlama arasındaki Kappa uyum katsayısının en az ,70 olması kriterine göre tüm yöntemlerde ve test veri oranlarında QWK değerleri kabul edilebilir bulunmuştur. Aynı kriter AC1 katsayısı için kullanıldığında tüm yöntemlerde ve test veri oranlarında kabul edilebilir değerlere ulaşılmıştır. Tablo 19 incelendiğinde görülebileceği gibi madde 5 için tüm test veri oranları ve yöntemler dikkate alındığında en yüksek uyum yüzdesi (,918), AC1 değeri (,888) ve QWK katsayısı (,925) BLSTM yönteminde ve %10 test veri oranında elde edilmiştir. Elde edilen bu değerler (UY=,918, AC1=,888 ve QWK=,925) gerçek puanlayıcı grupları ile nihai puanlar arasındaki uyum yüzdesi, AC1 ve QWK değerlerine yakın bulunmuştur.

Şekil 9 otomatik puanlama yöntemleri ve test veri oranlarına göre madde 20 için elde edilen uyum değerlerini göstermektedir. Şekil 9 ve tablo 19 dikkate alınarak madde 20 için yapılan yorumlara şekil 9'un ardından yer verilmektedir.



Şekil 9. Otomatik puanlama yöntemleri ve test veri oranlarına göre B_1 kitapçığı madde 20'ye ilişkin uyum değerlerini gösteren grafik

Tablo 19'da yer alan madde 20'ye ait değerler incelendiğinde gerçek puanlayıcılardan oluşan ilk puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesinin ,969, AC1 indeksinin ,945, QWK değerinin ise ,929 olduğu görülmektedir. Gerçek puanlayıcılardan oluşan ikinci puanlayıcı grubu ile nihai puanlar arasındaki uyum yüzdesi ,969, AC1 indesi ,946 ve QWK değeri ise ,929 olarak bulunmuştur. Otomatik puanlama ile nihai puanlar arasındaki uyum incelendiğinde %10 test veri oranı için en yüksek uyum yüzdesi BLSTM yöntemi ile ,839 olarak elde edilmiştir. En düşük uyum yüzdesi ise ,760 ile MNB yönteminde elde edilmiştir. Uyum yüzdeleri genel olarak incelendiğinde MNB yöntemi dışındaki tüm yöntemler için kabul edilebilir değerlere (>,80) ulaşıldığı görülmektedir. AC1 indeksi incelendiğinde en yüksek uyumun elde edildiği yöntem ,717 ile BLSTM yöntemidir. En düşük AC1 değeri ise ,562 ile MNB yöntemine ait olup bu yöntemi ,685 ile LR yöntemi izlemektedir. AC1 değerlerinin BLSTM, LR, LSTM ve SVM yöntemleri için iyi uyuma (>,60 Λ <,80); MNB yöntemi için orta düzeyde uyuma (>,40 Λ <,60) işaret ettiği görülmektedir. %10 test veri oranında QWK değeri en yüksek BLSTM yöntemi ile ,627 olarak bulunmuş olup bu değeri ,623 ile LSTM yöntemi izlemektedir. En düşük QWK değeri ,471 ile MNB yöntemine aittir. QWK değerlerinin %10 test veri oranında AC1 indekslerinden küçük olduğu sonucuna ulaşılmıştır.

QWK değeri %10 test veri oranında BLSTM ve LSTM yöntemleri için iyi uyum (>,60 Λ <,80); MNB, LR ve SVM yöntemleri için orta düzeyde uyum (>,40 Λ <,60) göstermektedir.

%20 test veri oranında en yüksek uyum yüzdesini ,837 ile BLSTM yöntemi göstermiş olup en düşük uyum yüzdesini ise ,750 ile MNB yöntemi göstermiştir. Uyum yüzdesine göre MNB yöntemi dışındaki tüm yöntemler kabul edilebilir uyum (>,80) göstermiştir. AC1 indeksi açısından uyum değerlendirildiğinde en yüksek uyuma ,710 ile BLSTM yönteminde rastlanmış olup en düşük uyuma ise ,544 ile MNB yönteminde rastlanmıştır. AC1 indeksi değerlerinin %20 test veri oranında BLSTM, LR, LSTM ve SVM yöntemleri için iyi uyum (<,80 Λ >,60); MNB yöntemi için orta düzeyde uyum (>,40 Λ <,60) gösterdiği belirtilebilir. %20 test veri oranı için QWK değerleri incelendiğinde en yüksek uyuma sahip yöntemin ,629 ile BLSTM yöntemi olduğu, en düşük uyuma sahip yöntemin ise ,447 ile MNB yöntemi olduğu belirtilebilir. En düşük uyuma sahip ikinci yöntem ise ,562 ile SVM'dir. Görüldüğü üzere %20 test veri oranında BLSTM yöntemi için iyi (<,80 Λ >,60); diğer yöntemler için orta düzeyde (>,40 Λ <,60) uyuma rastlanmıştır.

%33 test veri oranı için uyum yüzdesi en yüksek yöntem ,820 ile BLSTM yöntemidir. En düşük uyum yüzdesine sahip yöntem ise ,740 ile MNB yöntemidir. %33 test veri oranında LR ve BLSTM yöntemleri için uyum yüzdesi kabul edilebilir (>,80) bulunmuştur. AC1 indeksleri incelendiğinde en yüksek uyum ,691 ile BLSTM yöntemine aittir. En düşük uyum ise ,527 ile MNB yöntemine aittir ve bu yöntemi ,630 ile LSTM yöntemi izlemektedir. AC1 değerleri ele alındığında MNB yöntemi dışındaki tüm yöntemlerin iyi uyum (<,80 Λ >,60); MNB yönteminin ise orta düzeyde uyum (>,40 Λ <,60) gösterdiği görülmektedir. QWK değerleri değerlendirildiğinde en yüksek uyum ,572 ile BLSTM yönteminde, en düşük uyum ise ,421 ile MNB yönteminde elde edilmiştir. %33 test veri oranı için elde edilen QWK değerleri tüm yöntemler için orta düzeyde uyuma (<,60 Λ >,40) işaret etmektedir.

Şekil 9 incelendiğinde görülebileceği gibi tüm koşullarda tüm yöntemlere ait uyum yüzdeleri AC1 ve QWK katsayılarından, AC1 katsayıları ise QWK katsayılarından yüksektir. Tüm koşullar dikkate alındığında QWK değeri BLSTM yönteminde %10 ve %20 test veri oranlarında, LSTM yönteminde %10 test veri oranında iyi uyum bulunduğunu göstermiştir (>,60 Λ <,80). Tüm koşullarda AC1 değerleri MNB yöntemi dışındaki tüm yöntemler için iyi uyum (>,60 Λ <,80)

bulunduğunu göstermektedir. Tüm koşullar dikkate alındığında uyum yüzdesi MNB yönteminde tüm veri oranları, LSTM ve SVM yöntemlerinde %33 test veri oranı dışında tüm yöntemler için kabul edilebilir (>,80) değerler göstermiştir. Williamson, Xi ve Breyer (2012) tarafından gerçek puanlayıcılar ile otomatik puanlama arasındaki Kappa uyum katsayısının en az ,70 olması kriteri hiçbir koşulda karşılanmamıştır. Aynı kriter AC1 katsayısı için kullanıldığında %10 ve %20 test veri oranında BLSTM yönteminde, %10 test veri oranında LSTM yönteminde, %20 test veri oranında LR yönteminde karşılanmıştır. Tablo 19 incelendiğinde görülebileceği gibi madde 20 için tüm test veri oranları ve yöntemler dikkate alındığında en yüksek uyum yüzdesi (,839) ve AC1 değeri (,717) BLSTM yönteminde %10 test veri oranında, en yüksek QWK katsayısı (,629) BLSTM yönteminde %20 test veri oranında elde edilmiştir. Dolayısıyla en yüksek uyum değerlerinin BLSTM yönteminde ve %10 test veri oranında elde edildiği belirtilebilir. Elde edilen bu değerler (UY=,839, AC1=,717, QWK=,627) gerçek puanlayıcı grupları ile nihai puanlar arasındaki uyum yüzdesi, AC1 ve QWK değerlerinden düşük olsa da kabul edilebilir değerlere ulaşılmıştır.

Otomatik puanlama yöntemleri arasında genel bir karşılaştırma yapabilmek amacıyla yöntemlerin her bir maddedeki performanslarının ortalaması alınmıştır. Tablo 20 otomatik puanlama yöntemlerinin farklı test veri oranlarındaki performanslarını ve bu performansların ortalamalarını göstermektedir. Tablo 20'de her bir test veri oranında ve ortalama performansta her bir uyum katsayısı türünde en yüksek uyumu gösteren katsayılar koyu olarak en düşük uyumu gösteren katsayılar ise italik olarak gösterilmiştir.

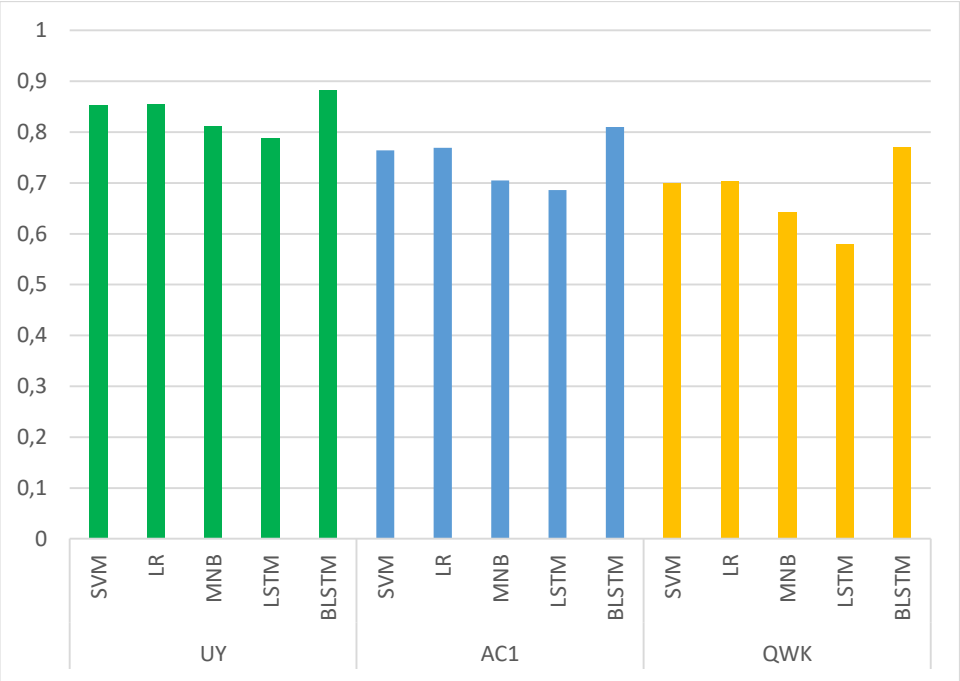
Tablo 20'de her bir test veri oranına ilişkin uyum yüzdeleri incelendiğinde değerlerin birbirine yakın olduğu görülmekle birlikte %33 test veri oranında değerlerde biraz azalma olmuştur. LSTM yöntemi dışında tüm yöntemler uyum yüzdesi açısından kabul edilebilir değerler göstermiştir. LSTM yöntemi ise kabul edilebilir uyuma yakın değerler göstermiştir. AC1 değerlerine göre değerlendirme yapıldığında değerlerin %33 test veri oranında biraz düşme eğilimi gösterdiği, bunun yanı sıra SVM, LR, MNB ve LSTM yöntemlerinin ortalama performanslarının iyi uyuma işaret ettiği görülmektedir. BLSTM yöntemi ise %10 ve %20 test veri oranlarında çok iyi, %33 test veri oranında iyi uyum göstermiştir. QWK değerleri incelendiğinde diğer uyum katsayılarına benzer şekilde %33 test veri oranında

düşme eğiliminin olduğu, bunun yanında %10, %20 ve %33 test veri oranlarında yakın değerlerin elde edildiği görülmektedir. QWK değeri açısından SVM, LR, MNB ve BLSTM yöntemleri iyi uyum bulunduğunu göstermektedir. LSTM yöntemi ise %20 test veri oranında iyi uyum, %10 ve %33 test veri oranlarında orta düzeyde uyum göstermiştir. %10, %20 ve %33 test veri oranlarının ortalamaları her bir yöntem ve uyum katsayısı açısından incelendiğinde en yüksek uyum, AC1 ve QWK değerine sahip yöntemin BLSTM olduğu görülmektedir. BLSTM yöntemi kabul edilebilir uyum yüzdesine sahip olmakla birlikte, AC1 katsayısına göre çok iyi, QWK katsayısına göre iyi uyum göstermektedir. SVM, LR ve MNB yöntemleri kabul edilebilir uyum yüzdesi, AC1 katsayısına ve QWK katsayısına göre ise iyi uyum göstermiştir. LSTM yöntemi kabul edilebilir uyum yüzdesine sahip olmayıp, AC1 indeksi açısından iyi uyum, QWK değeri açısından ise orta uyum göstermektedir. Şekil 10 yöntemlerin ortalama performanslarını göstermektedir.

Tablo 20

Otomatik Puanlama Yöntemlerinin Ortalama Performansları

		%10	%20	%33	Ortalama
	SVM	0,855	0,855	0,848	0,853
	LR	0,857	0,856	0,851	0,855
UY	MNB	0,816	0,810	0,807	0,811
	LSTM	*0,794*	*0,799*	*0,775*	*0,789*
	BLSTM	**0,889**	**0,883**	**0,874**	**0,882**
	SVM	0,768	0,767	0,756	0,764
	LR	0,773	0,771	0,762	0,769
AC1	MNB	0,712	0,704	0,700	0,705
	LSTM	*0,694*	*0,698*	*0,665*	*0,686*
	BLSTM	**0,822**	**0,810**	**0,798**	**0,810**
	SVM	0,705	0,704	0,689	0,699
	LR	0,710	0,710	0,692	0,704
QWK	MNB	0,658	0,640	0,627	0,642
	LSTM	*0,583*	*0,612*	*0,545*	*0,580*
	BLSTM	**0,782**	**0,775**	**0,755**	**0,771**



Şekil 10. Otomatik puanlama yöntemlerinin test veri oranlarına göre ortalamalarını gösteren grafik

Şekil 10'da %10, %20 ve %33 test veri oranlarında yöntemlerin gösterdiği performans ortalamaları genel olarak değerlendirildiğinde görülmektedir ki MNB ve LSTM yöntemleri diğer yöntemlerden biraz daha zayıf performans göstermektedir. En düşük performans LSTM yönteminde, en yüksek performans ise BLSTM yönteminde gözlemlenmiştir.

Tablo 20 incelenerek en iyi performansı gösteren üç koşul belirlenmiştir. Bunun yanında her bir madde için hesaplanan uyum katsayılarına göre en yüksek performansı gösteren üç koşul belirlenmiştir. İlk olarak her bir madde için ilk sırada, ikinci sırada ve üçüncü sırada en iyi performansı gösteren koşulların sayısı hesaplanmıştır. Bu şekilde birinci, ikinci ve üçüncü sırada en fazla sayıda en iyi performansı gösteren koşullar seçilmiştir. Ayrıca yapılan seçimi doğrulamak amacıyla seçilen koşulların ilk üç sıra içerisinde kaç kez yer aldığı değerlendirilmiştir. Hem madde ortalamalarının değerlendirilmesi hem de madde kapsamında yapılan değerlendirmeler sonucunda aynı koşullar seçilmiş ve BLSTM yöntemi %10 test veri oranının birinci, BLSTM yöntemi %20 test veri oranının ikinci ve BLSTM yöntemi %33 test veri oranının üçüncü sırada olduğu bulunmuştur.

2. Gerçek Puanlayıcılar ve Otomatik Puanlama Aracılığıyla Yapılan Eşitlemeye İlişkin Hatalar

Tablo 21, gerçek puanlayıcıların nihai puanları üzerinden yapılan eşitleme işlemine ait hataları bunun yanı sıra her iki test formu için otomatik puanlama işlemi sonrası elde edilen puanlarla gerçekleştirilen eşitleme işlemine ait hataları göstermektedir. Otomatik puanlamaya dayalı eşitleme işleminde BLSTM yöntemi ile %10, %20 ve %33 test veri oranlarında yapılandırılmış cevap maddeleri için elde edilen puanlar kullanılmıştır. Eşitleme işleminde KTK ve MTK'ya dayalı yöntemlerden yararlanılmış olup her iki kurama dayalı eşitleme yöntemleri için aynı kriterler kullanılarak karşılaştırma yapılmıştır. Eşitleme sonuçları değerlendirilirken SEE, BIAS ve RMSE değerlerinden yararlanılmıştır. Sistematik hatayı gösteren BIAS değeri, yüksek düzeyde negatif ve pozitif değerlerin birbirini nötrleyebilmesi (Zu ve Liu, 2010) nedeniyle yöntemlerin performanlarını karşılaştırmada kullanılmamıştır. BIAS değerinin negatif olması yeteneklerin olduğundan düşük, pozitif olması yeteneklerin olduğundan yüksek kestirildiğini (Pang, Madera, Radwan ve Zhang, 2010) gösterdiği için mutlak (absolute) BIAS değerleri üzerinde çalışılmamıştır. Seçkisiz hatayı gösteren SEE, yanlılık ve seçkisiz hatanın birleşimi olan RMSE üzerinden yöntemler karşılaştırılmıştır. En iyi yöntem seçilirken sistematik ve seçkisiz hatanın birleşimi olması nedeniyle RMSE değerlerinden yararlanılmıştır. Tablo 21'de BLSTM yöntemi aracılığıyla %10, %20, %33 test veri oranları ile elde edilen puanlar kullanılarak yapılan eşitlemeler ve gerçek puanlayıcıların puanları kullanılarak yapılan eşitlemelere ilişkin en düşük hatayı gösteren katsayılar koyu olarak, en yüksek hatayı gösteren yöntemler italik olarak gösterilmektedir.

Tablo 21

MTK ve KTK'ya Dayalı Eşitleme Yöntemleri ile Yapılan Eşitleme İşlemine Ait Hatalar

		SEE				BIAS				RMSE			
		Gerçek	BLSTM			Gerçek	BLSTM			Gerçek	BLSTM		
			%10	%20	%33		%10	%20	%33		%10	%20	%33
KTK	LC	0,211	0,213	0,209	0,215	**0,003**	**0,002**	**0,002**	**0,003**	0,211	0,213	0,209	0,215
	LT	0,198	0,201	0,197	0,202	**0,003**	**0,002**	**0,002**	**0,003**	0,198	0,201	0,197	0,202
	LT (WS=1)	0,197	0,200	0,196	0,200	**0,003**	**0,002**	**0,002**	0,004	0,197	0,200	0,196	0,200
	EC	0,351	*0,407*	0,396	*0,398*	0,061	*0,216*	*0,159*	*0,142*	0,357	*0,461*	*0,427*	*0,423*
	EF	0,330	0,336	0,347	0,336	0,062	0,032	0,052	0,071	0,336	0,337	0,351	0,344
	EF (WS=1)	0,330	0,362	0,371	0,348	0,059	0,048	0,158	0,062	0,335	0,365	0,403	0,353
	PSMEC	*0,357*	0,328	*0,405*	0,350	0,044	0,042	0,087	0,041	*0,359*	0,331	0,414	0,352
	PSMEF	0,321	0,341	0,360	0,307	0,023	0,021	0,084	0,021	0,322	0,342	0,369	0,307
	PSMEF (WS=1)	0,333	0,349	0,371	0,317	0,023	0,021	0,078	0,021	0,334	0,349	0,379	0,318
MTK	MM	0,061	0,079	0,098	0,071	-0,010	0,022	0,039	0,010	**0,062**	**0,083**	**0,106**	**0,072**
	MS	**0,050**	**0,047**	**0,006**	**0,012**	0,064	0,128	0,127	0,079	0,081	0,136	0,127	0,080
	HB	0,083	0,110	0,127	0,137	*-0,079*	-0,108	-0,087	-0,127	0,114	0,154	0,154	0,187
	SL	0,083	0,100	0,118	0,119	*-0,079*	-0,098	-0,078	-0,118	0,114	0,140	0,141	0,167

Tablo 21 incelendiğinde gerçek puanlayıcılar dikkate alındığında en düşük seçkisiz hatanın (SEE) ,050 ile MTK'ya dayalı MS yönteminde elde edildiği görülmektedir. Bu değeri ,061 ile MM yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek seçkisiz hatayı (,083) SL ve HB yöntemleri göstermiştir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük seçkisiz hatayı (,197) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,198 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) eşitleme yöntemi takip etmektedir. En yüksek seçkisiz hatanın (,357) bulunduğu yöntem ise iki değişkenli logaritmik doğrusal fonksiyon ile ön düzgünleştirme yapılan zincir eşit yüzdelikli (PSMEC) eşitleme yöntemidir. Gerçek puanlayıcılar kullanıldığı durumda genel olarak en yüksek seçkisiz hatalar eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az seçkisiz hata göstermiştir.

%10 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama işlemi sonrasında yapılan eşitleme sonuçları seçkisiz hata açısından değerlendirildiğinde en düşük seçkisiz hatanın (,047) MS yönteminde bulunduğu görülmektedir. Elde edilen bu değer (,047) gerçek puanlayıcıların kullanıldığı durumdan (,050) daha düşüktür. %10 test veri oranında BLSTM yöntemi aracılığıyla elde edilen bu değeri (,047) ,079 ile MM yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek seçkisiz hatayı (,110) HB yöntemi göstermiştir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük seçkisiz hatayı (,200) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,201 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) eşitleme yöntemi takip etmektedir. En yüksek seçkisiz hatanın (,407) bulunduğu yöntem ise zincir eşit yüzdelikli (EC) eşitleme yöntemidir. %10 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama sonrasında yapılan eşitleme işleminde genel olarak en yüksek seçkisiz hatalar eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az seçkisiz hata göstermiştir. Tüm yöntemler için hesaplanan

eşitleme hataları gerçek puanlayıcılarla gerçekleştirilen eşitleme işlemine ait hatalara yakın olmakla birlikte iki koşulda otomatik puanlama (%10 test veri oranı ile BLSTM yöntemi kullanılarak) daha az hata ile eşitleme yapmıştır.

%20 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama işlemi sonrasında yapılan eşitleme sonuçları seçkisiz hata açısından değerlendirildiğinde en düşük hatanın (,006) MS yönteminde bulunduğu görülmektedir. Elde edilen bu değer (,006) 0'a oldukça yakındır ve gerçek puanlayıcıların kullanıldığı durumda elde edilen hatadan (,050) oldukça düşüktür. %20 test veri oranında BLSTM yöntemi aracılığıyla elde edilen bu değeri (,006) ,098 ile MM yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek seçkisiz hatayı (,127) HB yöntemi göstermiştir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük seçkisiz hatayı (,196) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,197 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) eşitleme yöntemi takip etmektedir. En yüksek seçkisiz hatanın (,405) bulunduğu yöntem ise iki değişkenli logaritmik doğrusal fonksiyon ile ön düzgünleştirme yapılmış zincir eşit yüzdelikli eşitleme (PSMEC) yöntemidir. %20 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama sonrasında yapılan eşitleme işleminde genel olarak en yüksek seçkisiz hatalar eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az seçkisiz hata göstermiştir. Tüm yöntemler için hesaplanan seçkisiz eşitleme hataları gerçek puanlayıcılarla gerçekleştirilen eşitleme işlemine ait seçkisiz hatalara yakın olmakla birlikte dört koşulda otomatik puanlama (%20 test veri oranı ile BLSTM yöntemi kullanılarak) daha az seçkisiz hata ile eşitleme yapmıştır.

%33 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama işlemi sonrasında yapılan eşitleme sonuçları seçkisiz hata açısından değerlendirildiğinde en düşük hatanın (,012) MS yönteminde bulunduğu görülmektedir. Elde edilen bu değer (,012) 0'a oldukça yakındır ve gerçek puanlayıcıların kullanıldığı durumda elde edilen hatadan (,050) oldukça düşüktür. %33 test veri oranında BLSTM yöntemi aracılığıyla elde edilen bu değeri (,012) ,071 ile MM yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek

seçkisiz hatayı (,137) HB yöntemi göstermiştir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük seçkisiz hatayı (,200) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,202 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) yöntemi takip etmektedir. En yüksek seçkisiz hatanın (,398) bulunduğu yöntem ise zincir eşit yüzdelikli eşitleme (EC) yöntemidir. %33 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama sonrasında yapılan eşitleme işleminde genel olarak en yüksek seçkisiz hatalar eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az seçkisiz hata göstermiştir. Tüm yöntemler için hesaplanan seçkisiz eşitleme hataları gerçek puanlayıcılarla gerçekleştirilen eşitleme işlemine ait seçkisiz hatalara yakın olmakla birlikte dört koşulda otomatik puanlama (%33 test veri oranı ile BLSTM yöntemi kullanılarak) daha az hata ile eşitleme yapmıştır.

Tüm eşitleme işlemlerinde elde edilen seçkisiz hatalar değerlendirildiğinde hataların birbirlerine oldukça yakın olduğu görülmektedir. Otomatik puanlama aracılığıyla gerçekleştirilen eşitleme işlemlerinde bazı durumlarda gerçek puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemlerinden daha düşük seçkisiz hatalara rastlanmıştır. Gerçek puanlayıcılar kullanılsa da otomatik puanlama yapılsa da MTK'ya dayalı yöntemler daha düşük seçkisiz hatalara sahiptir. Tüm eşitleme işlemleri dikkate alındığında en düşük seçkisiz hata ,006 ile %20 test veri oranı ve BLSTM yöntemi puanları kullanılarak gerçekleştirilen eşitleme işlemiyle MS yönteminde elde edilmiştir. Tüm eşitleme işlemlerinde en yüksek seçkisiz hata ,407 ile %10 test veri oranı ve BLSTM yöntemi puanları kullanılarak gerçekleştirilen eşitleme işlemiyle zincir eşit yüzdelikli (EC) eşitleme yönteminde elde edilmiştir.

Gerçek puanlayıcılarla gerçekleştirilen eşitleme işleminde elde edilen sistematik hata (yanlılık [BIAS]) büyüklükleri ,003 ile ,079 aralığında değişmektedir. %10 test veri oranı ve BLSTM yöntemi aracılığıyla elde edilen puanlarla gerçekleştirilen eşitleme sonrasında elde edilen yanlılık değerleri büyüklükleri ,002 ile ,216 aralığında değişmektedir. %20 test veri oranı ve BLSTM yöntemi aracılığıyla elde edilen puanlarla gerçekleştirilen eşitleme sonrasında elde edilen yanlılık

değerleri büyüklükleri ,002 ile ,159 aralığında değişmektedir. %33 test veri oranı ile BLSTM yöntemi aracılığıyla elde edilen puanlarla gerçekleştirilen eşitleme sonrasında elde edilen yanlılık değerleri büyüklükleri ise ,003 ile ,142 aralığında değişmektedir.

Tablo 21 incelendiğinde gerçek puanlayıcılar dikkate alındığında en düşük hata kareleri ortalaması karekökü (RMSE) değerinin ,062 ile MTK'ya dayalı MM yönteminde elde edildiği görülmektedir. Bu değeri ,081 ile MS yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek RMSE değerini (,114) SL ve HB yöntemleri göstermiştir. Bu sonuçlar MTK'ya dayalı yöntemlerde moment yöntemlerinin (MM ve MS), karakteristik eğri yöntemlerinden (SL ve HB) daha düşük hata gösterdiği anlamına gelmektedir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük RMSE değerini (,197) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,198 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) yöntemi takip etmektedir. En yüksek RMSE değerinin (,359) bulunduğu yöntem ise iki değişkenli logaritmik doğrusal fonksiyon ile ön düzgünleştirme yapılan zincir eşit yüzdelikli (PSMEC) eşitleme yöntemidir. Gerçek puanlayıcılar kullanıldığı durumda genel olarak en yüksek RMSE değerleri eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az RMSE değerleri göstermiştir.

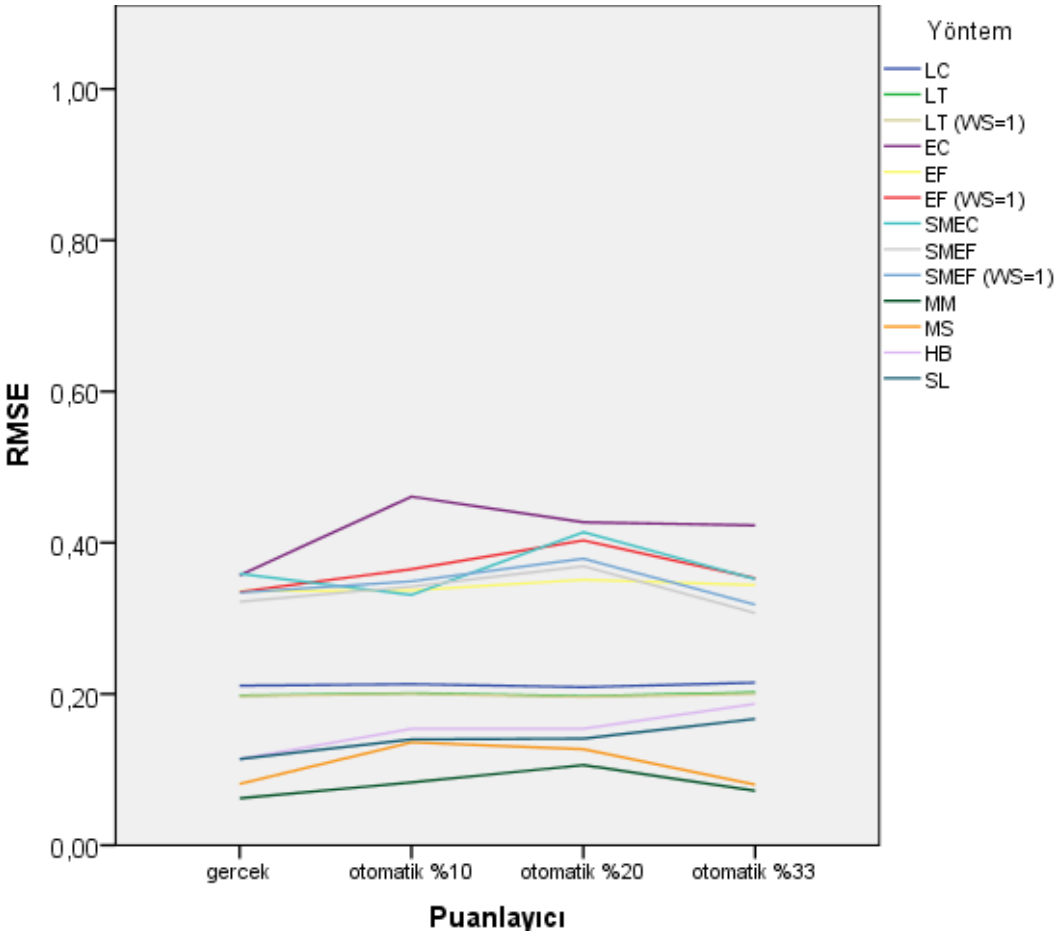
%10 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama işlemi sonrasında yapılan eşitleme sonuçları RMSE açısından değerlendirildiğinde en düşük hatanın (,083) MM yönteminde bulunduğu görülmektedir. Elde edilen bu değer (,083) gerçek puanlayıcıların kullanıldığı durumda elde edilen en düşük RMSE değerine (,062) yakındır. %10 test veri oranında BLSTM yöntemi aracılığıyla elde edilen bu değeri (,083) ,136 ile MS yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek RMSE değerini (,154) HB yöntemi göstermiştir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük RMSE değerini (,200) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,201 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem

sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) eşitleme yöntemi takip etmektedir. En yüksek RMSE değerinin (,461) bulunduğu yöntem ise zincir eşit yüzdelikli (EC) eşitleme yöntemidir. %10 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama sonrasında yapılan eşitleme işleminde genel olarak en yüksek RMSE değerleri eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az RMSE değerlerine sahiptir. Tüm yöntemler için hesaplanan eşitleme hataları gerçek puanlayıcılarla gerçekleştirilen eşitleme işlemine ait hatalara yakın olmakla birlikte bir koşulda (iki değişkenli logaritmik doğrusal fonksiyon ile ön düzgünleştirme yapılan zincir eşit yüzdelikli eşitleme yöntemi) otomatik puanlama (%10 test veri oranı ve BLSTM yöntemi kullanılarak) daha az hata ile eşitleme yapmıştır.

%20 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama işlemi sonrasında yapılan eşitleme sonuçları RMSE açısından değerlendirildiğinde en düşük hatanın (,106) MM yönteminde bulunduğu görülmektedir. Elde edilen bu değer (,106) gerçek puanlayıcıların kullanıldığı durumda elde edilen en düşük RMSE değerine (,062) yakındır. %20 test veri oranında BLSTM yöntemi aracılığıyla elde edilen bu değeri (,106) ,127 ile MS yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek RMSE değerini (,154) HB yöntemi göstermiştir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük RMSE değerini (,196) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,197 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) yöntemi takip etmektedir. En yüksek RMSE değerinin (,427) bulunduğu yöntem ise zincir eşit yüzdelikli (EC) eşitleme yöntemidir. %20 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama sonrasında yapılan eşitleme işleminde genel olarak en yüksek RMSE değerleri eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az RMSE değerlerine sahiptir. Tüm yöntemler için hesaplanan eşitleme hataları gerçek puanlayıcılarla gerçekleştirilen eşitleme işlemine ait hatalara yakın olmakla birlikte üç koşulda otomatik puanlama (%20 test veri oranı ve BLSTM yöntemi kullanılarak) daha az hata ile eşitleme yapmıştır.

%33 test veri oranı ile BLSTM yöntemi ile gerçekleştirilen otomatik puanlama işlemi sonrasında yapılan eşitleme sonuçları RMSE açısından değerlendirildiğinde en düşük hatanın (,072) MM yönteminde bulunduğu görülmektedir. Elde edilen bu değer (,072) gerçek puanlayıcıların kullanıldığı durumda elde edilen en düşük RMSE değerine (,062) oldukça yakındır. %33 test veri oranında BLSTM yöntemi aracılığıyla elde edilen bu değeri (,072) ,080 ile MS yöntemi takip etmektedir. MTK'ya dayalı yöntemler kullanıldığında en yüksek RMSE değerini (,187) HB yöntemi göstermiştir. Gerçek puanlayıcılar kullanıldığında KTK'ya dayalı eşitleme yöntemlerinde en düşük RMSE değerini (,200) gösteren yöntem sentetik evren oranının 1 olarak belirlendiği Tucker doğrusal (LT [WS=1]) eşitleme yöntemidir. Bu değeri ,202 ile seçkisiz evren oranının değiştirilmediği ve seçkisiz evren oranının örneklem sayılarına dayalı olarak belirlendiği Tucker doğrusal (LT) yöntemi takip etmektedir. En yüksek RMSE değerinin (,423) bulunduğu yöntem ise zincir eşit yüzdelikli (EC) eşitleme yöntemidir. %33 test veri oranı ve BLSTM yöntemi ile gerçekleştirilen otomatik puanlama sonrasında yapılan eşitleme işleminde genel olarak en yüksek RMSE değerleri eşit yüzdelikli eşitleme yöntemlerinde elde edilmiştir. Bu koşulda MTK'ya dayalı yöntemler genel olarak KTK'ya dayalı yöntemlerden daha az RMSE değerlerine sahiptir. Tüm yöntemler için hesaplanan eşitleme hataları gerçek puanlayıcılarla gerçekleştirilen eşitleme işlemine ait hatalara yakın olmakla birlikte dört koşulda otomatik puanlama (%33 test veri oranı ile BLSTM yöntemi kullanılarak) daha az hata ile eşitleme yapmıştır.

Şekil 11 gerçek puanlayıcılar, %10, %20 ve %33 test veri oranına göre gerçekleştirilen otomatik puanlama aracılığıyla yapılan eşitleme işlemine ait RMSE değerlerini göstermektedir. Alanyazında (Pang, Madera, Radwan ve Zhang, 2010), RMSE değerinin %1'in altında olmasının önemsiz olduğu belirtildiği için grafik 0 ile 1 aralığında çizilmiştir.



Şekil 11. Puanlayıcı türüne göre yöntemlerin RMSE değerlerini gösteren grafik

Şekil 11'de tüm eşitleme işlemlerinde elde edilen RMSE değerleri değerlendirildiğinde hataların birbirlerine yakın olduğu görülmektedir. Otomatik puanlama aracılığıyla gerçekleştirilen eşitleme işlemlerinde bazı durumlarda gerçek puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemlerinden daha düşük RMSE değerleri elde edilmiştir. Gerçek puanlayıcılar kullanılsa da otomatik puanlama yapılsa da MTK'ya dayalı yöntemler KTK'ya dayalı yöntemlerden daha düşük RMSE değerlerine sahiptir. Tüm eşitleme işlemleri dikkate alındığında en düşük RMSE değeri ,062 ile gerçek puanlayıcılar kullanılarak yapılan eşitleme işlemiyle MM yönteminde elde edilmiştir. Otomatik puanlama aracılığıyla yapılan eşitlemede ise en düşük RMSE değeri ,072 ile MM yönteminde elde edilmiştir. MTK yöntemleri her bir koşulda kendi arasında karşılaştırıldığında moment yöntemlerinin karakteristik eğri yöntemlerinden daha az hata (RMSE) gösterdiği belirtilebilir. Tüm eşitleme işlemlerinde gerçek puanlayıcılar kullanıldığında yapılan eşitlemede en yüksek RMSE değeri ,359 ile iki değişkenli logaritmik doğrusal fonksiyon ile ön düzgünleştirme yapılan zincir eşit yüzdelikli (PSMEC) eşitleme yönteminde elde

edilmiştir. Otomatik puanlamada en yüksek RMSE değeri ise ,461 ile zincir eşit yüzdelikli (EC) eşitleme yönteminde elde edilmiştir. Genel olarak eşit yüzdelikli eşitleme yöntemlerinin daha fazla hata (RMSE) ile eşitleme yaptığı belirtilebilir. Sentetik evren oranının 1 olarak değiştirilmesi doğrusal yöntemlerde RMSE değerlerini genel olarak düşürürken eşit yüzdelikli eşitleme yöntemlerinde ve eşit yüzdelikli eşitleme yöntemlerine ön düzgünleştirme uygulandığında RMSE değerlerini genel olarak yükseltmiştir. Sentetik evren oranının 1 olarak belirlenmesi katsayılarda çok büyük düşüş ya da yükselişler oluşturmamıştır. Ön düzgünleştime işlemi bazı durumlarda RMSE değerlerini düşürürken bazı durumlarda da arttırmıştır.

BLSTM yöntemiyle %10, %20 ve %33 test veri oranlarıyla otomatik puanlama yapılarak elde edilen puanlarla gerçekleştirilen eşitleme sonucunda bulunan hataların ortalamaları alınarak bu ortalamaların gerçek puanlayıcılar aracılığıyla yapılan eşitleme işlemine ait hatalardan anlamlı farklılık gösterme durumu incelenmiştir. Eşitleme yöntemleri, bu yöntemlerin sentetik evren oranlarındaki değişiklikler ve/veya öndüzgünleştirme işlemleri ile elde edilen versiyonlarının gerçek puanlayıcılar ve otomatik puanlama ortalamaları arasında farklılık gösterme durumu her bir grup için normal dağılım varsayımının karşılanmaması nedeniyle Mann Whitney U testi aracılığıyla incelenmiştir. Sonuçlar Tablo 22'de gösterilmektedir.

Tablo 22

Gerçek Puanlayıcılar ve Otomatik Puanlama Aracılığıyla Gerçekleştirilen Eşitleme İşlemleri Sonucunda Elde Edilen RMSE Değerlerine İlişkin Fark Testi

	Puanlama	N	Sıra Ortalaması	Sıra Toplamı	U	p
RMSE	Gerçek Puanlayıcı	13	12,000	156,000	65,000	,336
	Otomatik Puanlama	13	15,000	195,000		

Tablo 22 incelendiğinde gerçek puanlayıcılar aracılığıyla elde edilen 13 eşitleme yöntemine ait RMSE değerlerinin (ortanca=,211), otomatik puanlama aracılığıyla elde edilen 13 eşitleme yöntemine ait ortalama RMSE değerlerinden (ortanca=,212) anlamlı şekilde farklılaşmamadığı (U=65,000, p=,336>,05) görülmektedir. Buna göre gerçek puanlaycıların ya da otomatik puanlamanın

kullanılmasının eşitleme işlemi sonucunda elde edilen RMSE değerleri üzerinde anlamlı bir etkisinin olmadığı belirtilebilir. Etki büyüklüğü r değeri aracılığıyla incelenmiş ve ,2 bulunmuştur. Cohen (1988:79)'e göre ,1 değeri düşük düzeydeki, ,3 değeri orta düzeyde etkiye işaret etmektedir. Buna göre puanlama türünün RMSE değerleri üzerindeki etkisinin düşük düzeyin biraz üzerinde olduğu belirtilebilir. Gerçek puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemine ait hatalar (RMSE) ile otomatik puanlama aracılığıyla gerçekleştirilen eşitleme işlemine ait hata (RMSE) ortalamaları arasındaki ilişki Spearman sıra farkları korelasyonu ile değerlendirilmiş ve yüksek düzeyde ve anlamlı (r=,96, p=,00<,05) ilişkiye rastlanmıştır.

Bölüm 5
Sonuç, Tartışma ve Öneriler

Sonuç ve Tartışma

Araştırma otomatik puanlama yöntemlerini eğitim ve test veri oranı üzerinde yapılan değişikliklerle karşılaştırmaktadır. Bu amaçla %10, %20 ve %33 olarak belirlenen test veri oranlarına göre SVM, LR, MNB, LSTM ve BLSTM otomatik puanlama yöntemleri ile yapılan puanlamalar gerçek puanlayıcıların üzerinde anlaştıkları puanlarla karşılaştırılmıştır. Yapılan incelemeler göstermiştir ki en iyi otomatik puanlama BLSTM yöntemiyle gerçekleştirilmiştir. LSTM ve MNB yöntemleri SVM, LR ve BLSTM yöntemlerinden daha düşük uyum değerlerine sahiptir. Kumar ve Rama Sree (2014)'nin çeşitli sınıflandırma yöntemleri ile ilgili daha önce yaptıkları deneylerde sade Bayes yönteminin LR ve SVM yöntemlerinden daha düşük uyum yüzdelerine sahip olduğu belirlenmiştir. Bu sonuç araştırma bulgularını destekler niteliktedir. Test veri oranlarına göre yapılan karşılaştırmalar %33 test veri oranında uyum katsayılarının biraz azalma gösterdiğine işaret etse de SVM, LR, MNB ve BLSTM yöntemleri tüm koşullarda iyi ya da çok iyi uyum bulunduğunu göstermektedir. Gierl, Latifi, Lai, Boulais ve Champlain (2014) SVM yöntemi ile gerçekleştirdiği otomatik puanlama işleminde QWK değerinin çok iyi uyum gösterdiğini belirlemiştir. Bu araştırmada ise SVM yönteminin iyi uyum gösterdiği belirlenmiştir. Taghipour ve Tou Ng (2016) otomatik puanlama işleminde yinelenen sinir ağlarını karşılaştırdıkları araştırmada en yüksek QWK değerine (.746) sahip yöntemin LSTM olduğunu bulmuştur. Bu yönteme en yakın QWK değeri BLSTM yönteminde (.699) elde edilmiştir. Bu araştırmada ise BLSTM yöntemine ait QWK değeri benzer şekilde iyi uyuma işaret etmektedir. Ancak bu araştırmada LSTM yönteminin QWK değerine göre orta düzeyde uyum gösterdiği belirlenmiştir. Bu durumun nedeni LSTM yönteminde cümlenin tek yönlü, BLSTM yönteminde ise cümlelerin iki yönlü incelenmesi olabilir. Otomatik puanlama için kabul edilebilir en düşük uyum katsayısına göre karşılaştırma yapıldığında LR ve BLSTM yöntemlerinin istenilen seviyede olduğu, SVM yönteminin ise istenilen seviyeye çok yakın olduğu belirlenmiştir. Bu yönde kurulan otomatik puanlama sisteminin geniş ölçekli testlerde kullanılabileceği görülmüştür. Ayrıca bu araştırma sonucunda oluşturulan sistemin, Adesiji, Agbonifo, Adesuyi ve Olabode (2016) tarafından hazırlanan ve denetlenmeyen makine öğrenimine dayalı yöntemden daha iyi

performans gösterdiği belirtilebilir. Benzer şekilde Türk diline benzer özellik taşıyan dillerde geliştirilen otomatik puanlama sistemleri bu araştırmadaki gibi denetlenen makine öğrenimine dayalı değildir ancak bu dillerde oluşturulan sistemler de iyi performanslara sahiptir. Ishioka ve Kameda (2006) Japon dili üzerinde kurdukları otomatik puanlama sistemiyle gerçek puanlar arasında yüksek düzeyde korelasyon bulunduğu sonucuna ulaşmıştır. Jang, Kang, Noh, Kim, Sung ve Seong (2014) Kore dili üzerinde geliştirdikleri otomatik puanlama sisteminin gerçek puanlayıcılar ile yüksek düzeyde uyum gösterdiğini belirlemiştir. Kakkonen, Myller, Timonen ve Sutinen (2005) Fin dili üzerinde geliştirdiği otomatik puanlama sisteminde çoğu maddede gerçek puanlayıcılarla otomatik puanlama sistemi arasında yüksek düzeyde korelasyona rastlamıştır.

Tüm yöntemler ve test veri oranları çaprazlanarak en iyi uyum değerlerini gösteren üç otomatik puanlama koşulu ile eşitleme işlemi gerçekleştirilmiştir. Eşitleme işlemi hem gerçek puanlayıcılar hem de üç otomatik puanlama koşulu ile gerçekleştirilmiştir. Gerçek puanlayıcılar için yapılan eşitleme işleminde A_1 ve B_1 kitapçığı için gerçek puanlayıcıların nihai puanlarından, otomatik puanlama için yapılan eşitlemede her iki test formunda yer alan yapılandırılmış cevap maddelerinin otomatik puanlanması sonucunda elde edilen puanlardan yararlanılmıştır. Yapılandırılmış cevap maddeleri ve objektif puanlanan maddeler ayrı ayrı eşitleme işlemine tabi tutulmamıştır. Eşitleme yöntemi olarak KTK ve MTK'ya dayalı yöntemlerden yararlanılmıştır. Alanyazında KTK ve MTK'ya dayalı eşitleme yöntemlerini karma testlerde ve denk olmayan gruplarda ortak madde deseni kullanarak karşılaştıran benzer çalışmalar bulunmaktadır (Hagge ve Kolen, 2011; Hagge, Liu, He, Powers, Wang ve Kolen, 2011; He, 2011; Lee, Lee ve Brennan, 2012; Liu ve Kolen, 2011; Wolf, 2013). KTK'ya dayalı eşitleme yöntemlerinden Tucker doğrusal, zincir doğrusal, zincir eşit yüzdelikli, frekans eşit yüzdelikli, MTK'ya dayalı gerçek puan eşitleme yöntemlerinden ortalama-ortalama, ortalama-standart sapma, Stocking-Lord ve Haebara'dan yararlanılmıştır. Alanyazındaki araştırmaların çoğunluğu (Hagge ve Kolen, 2011; Hagge, Liu, He, Powers, Wang ve Kolen, 2011; He, 2011; Liu ve Kolen, 2011; Wolf, 2013) KTK'ya dayalı yöntemlerden zincir eşit yüzdelikli ve frekans tahmini, MTK'ya dayalı yöntemlerden gözlenen ve gerçek puan eşitlemeyi karşılaştırmaktadır. Bu araştırmalardan Hagge ve Kolen (2011) ve Hagge, Liu, He, Powers, Wang ve Kolen (2011) MTK'ya dayalı

gerçek puan eşitlemede Haebara yönteminden yaralanırken; Wolf (2013) MTK'ya dayalı gerçek puan eşitlemede eş zamanlı ölçeklemeden; He (2011), Liu ve Kolen (2011) MTK'ya dayalı gerçek puan eşitlemede Stocking-Lord yönteminden yararlanmıştır. Lee, Lee ve Brennan ise araştırmalarında Tucker, Levine gözlenen puan, Levine gerçek puan, zincir eşit yüzdelikli, frekans tahmini, Stocking-Lord ve MTK gözlenen puan eşitleme yöntemlerini karşılaştırmıştır. KTK'ya dayalı eşitleme türlerinden eşit yüzdelikli eşitleme kullanıldığı durumlarda iki değişkenli logaritmik doğrusal fonksiyonla ön düzgünleştirme uygulanmıştır. Bu araştırmadakine benzer olarak Hagge, Liu, He, Powers, Wang ve Kolen (2011), Lee, Lee ve Brennan (2012) ve Wolf (2013) logaritmik doğrusal fonksiyonla ön düzgünleştirme yapmıştır. Liu ve Kolen (2011) ise eşitleme işleminde karşılaştırma yapabilmek amacıyla evrene yönelik sonuçları elde ederken ön düzgünleştirme işleminden yararlanmıştır. Ayrıca zincir eşitleme yöntemi dışındaki eşitleme yöntemlerinin sentetik evren oranları değiştirilmiştir. Hagge ve Kolen (2011), Hagge, Liu, He, Powers, Wang ve Kolen (2011) ve Wolf (2013) bu araştırmadakine benzer şekilde sentetik evren oranını 1 olarak değiştirmiştir. Ancak bu araştırmalar sentetik evren oranının etkisini değerlendirmeyip sonuçları testi alan yeni gruba dayalı olarak göstermektedir. Sonuçta tüm yöntemler ve yöntemlerin farklı kombinasyonlarında otomatik puanlama koşullarında ve gerçek puanlayıcıların bulunduğu koşulda elde edilen hataların (RMSE) yakın olduğu bulunmuştur. Otomatik puanlama aracılığıyla gerçekleştirilen eşitleme işlemlerinde bazı durumlarda gerçek puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemlerinden daha düşük RMSE değerlerine rastlanmıştır. Ön düzgünleştirmenin bazı durumlarda RMSE değerlerini azalttığı gözlemlenirken bazı durumlarda ise arttırdığı gözlemlenmiştir. Hagge, Liu, He, Powers, Wang ve Kolen (2011) ön düzgünleştirmenin zincir eşit yüzdelikli ve frekans tahmini yöntemlerine ait standart hatayı düşürdüğünü belirlemiştir. Sentetik evren oranının değiştirilmesi ise doğrusal eşitleme türünde RMSE değerlerini düşürürken, eşit yüzdelikli eşitleme türünde RMSE değerlerini yükseltmiştir. Eşitleme işlemlerinin sonucu MTK'ya dayalı yöntemlerin KTK'ya dayalı yöntemlere göre daha düşük hata (SEE ve RMSE açısından) ile eşitleme yaptığını göstermektedir. Hagge ve Kolen (2011), Liu ve Kolen (2011) bu araştırmadakine benzer koşullarda hata kareleri ortalamasının karekköküne göre MTK'ya dayalı yöntemlerin KTK'ya dayalı yöntemlerden daha düşük hata gösterdiğini belirtmiştir. Hagge, Liu, He, Powers, Wang ve Kolen (2011) de MTK'ya dayalı yöntemlerin, KTK'ya dayalı yöntemlerden

daha düşük SEE değerlerine sahip olduğunu belirtmektedir. Her ne kadar aynı ölçüt kullanılmasa da Lee, Lee ve Brennan (2012) ve Wolf (2013) birincil düzey eşitlik korunumu dikkate alındığında MTK'ya dayalı yöntemlerin, KTK'ya dayalı yöntemlerden daha iyi performans gösterdiğini belirtmiştir. En düşük SEE ve RMSE değerini gösteren yöntem gerçek puanlayıcılar kullanıldığında da otomatik puanlama koşullarında da MTK gerçek puan eşitleme yöntemleridir. Liu ve Kolen (2011) de MTK gerçek puan eşitlemenin, frekans tahmini ve zincir eşit yüzdelikli eşitlemeden daha düşük SEE değerlerine sahip olduğunu bulmuştur. Her ne kadar aynı ölçüt dikkate alınmasa da Lee, Lee ve Brennan (2012) birincil düzey eşitlik açısından MTK gerçek puan eşitlemenin; Tucker doğrusal, zincir eşit yüzdelikli, frekans tahmini, ön düzgünleştirilmiş zincir eşit yüzdelikli ve ön düzgünleştirilmiş frekans tahmini yönteminden daha iyi bir performans sergilediğini belirlemiştir. Wolf (2013) da birincil düzey eşitlik açısından MTK gerçek puan eşitlemenin, frekans tahmini ve zincir eşit yüzdelikli eşitlemeden daha iyi performans gösterdiğini bulmuştur. Her bir koşul için MTK'ya dayalı yöntemler kendi arasında karşılaştırıldığında moment yöntemlerinin karakteristik eğri yöntemlerinden daha az hata ile eşitleme yaptığı belirtilebilir. Bu durumun ortak madde sayısı ve test uzunluğundan bunun yanında doğrusallıkla alakalı olabileceği düşünülmektedir. En yüksek RMSE ve SEE değerlerine ise eşit yüzdelikli eşitleme yöntemlerinde rastlanılmaktadır. RMSE ve SEE değerlendirildiğinde en yüksek hatalar zincir eşit yüzdelikli ve ön düzgünleştirme yapılmış zincir eşit yüzdelikli eşitlemede elde edilmiştir. Hagge ve Kolen (2011), Hagge, Liu, He, Powers, Wang ve Kolen (2011) de araştırmalarında en yüksek SEE değerine sahip yöntemin zincir eşit yüzdelikli eşitleme olduğunu belirtmektedir. Ancak He (2011) birincil düzey eşitlik ölçütüne göre zincir eşitlemenin, frekans tahmini yönteminden daha iyi performans gösterdiğini belirtmektedir. Bu araştırma ile He (2011)'nin araştırmasında karşılaşılan bu farklılığın nedeninin örneklem sayısından kaynaklı olduğu düşünülmektedir. Otomatik puanlamada her bir eşitleme yöntemi için farklı test veri oranlarının ortalama RMSE değerleri hesaplanarak bu değerlerin gerçek puanlayıcılar aracılığıyla gerçekleştirilen eşitleme işlemine ait hatalardan anlamlı farklılık gösterme durumları incelenmiştir. Sonuçta hatalar arasında anlamlı farklılık olmadığı ve hataların yüksek düzeyde uyum gösterdiği belirlenmiştir. Olgar (2015) karma testlerde otomatik puanlama kullanarak gerçekleştirdiği araştırmada bu araştırmaya benzer şekilde eşitleme işleminde benzer sonuçlara ulaşmıştır. Almond

(2014) sadece yapılandırılmış cevaplardan oluşan testlerde otomatik puanlama işleminden yararlanarak mantıklı eşitleme sonuçlarına ulaşmıştır.

Öneriler

Araştırmanın sonucuna dayalı önerilere ve ileride yapılacak araştırmalara yönelik önerilerilere aşağıda sırasıyla yer verilmektedir.

Araştırma sonuçlarına dayalı öneriler. Araştırma sonucunda elde edilen bulgulara dayalı olarak aşağıdaki öneriler listelenmiştir.

1. Otomatik puanlama yöntemlerinden BLSTM ve LR yöntemleri tercih edilebilir.
2. Otomatik puanlamada MNB ve LSTM yöntemlerinin bu araştırmada kullanılan veriye benzer özellikler taşıyan verilerde kullanılmaması önerilebilir.
3. Eşitleme yöntemleri açısından MTK'ya dayalı yöntemlerin kullanılması, eğer bu araştırmadakine benzer türde veriler üzerinde çalışılıyorsa moment yöntemlerinin kullanılması önerilebilir.

İleride yapılacak araştırmalara yönelik öneriler. Aşağıda ileride yapılacak araştırmlara yönelik öneriler listelenmiştir.

1. Bu araştırma yaklaşık olarak en az 400 veri ile yapılan otomatik puanlama sonuçlarını yansıtmaktadır. Bundan sonraki araştırmalarda daha az sayıda eğitim verisi ile otomatik puanlama yapılarak bu durumun uyum katsayılarına etkisi değerlendirilebilir. Dahası büyük örneklemlerde (>1000 ya da >3000) fazla sayıda eğitim verisi ile otomatik puanlama işlemi yapıldıktan sonra eğitim verileri seçkisiz bir biçimde kademe kademe azaltılarak bu durumun otomatik puanlamaya etkisi incelenebilir.
2. Bu araştırmada, modelin geçerliğini belirlemek üzere çapraz geçerlik ile yeniden örnekleme tekniğinden yararlanılmıştır. Daha sonraki araştırmalarda altın standart (gold standard) tekniği ile modelin geçerliği test edilebilir.
3. Bu araştırmada sentetik evren oranının değiştirilmesinin eşitleme hatalarına etkisi değerlendirilmiştir. Daha sonraki araştırmalarda eşitlenecek testlerdeki veri sayıları arasında fark bulunduğunda sentetik evren oranının ,5 olmasının etkisi incelenebilir.

4. Bu araştırma ön düzgünleştirme yönteminin etkisini ele almaktadır. İleride yapılacak araştırmalarda ön düzgünleştirme ve son düzgünleştirme yöntemleri arasında karşılaştırma yapılabilir. Ayrıca farklı ön düzgünleştirme ve son düzgünleştirme yöntemleri farklı desenler üzerinde karşılaştırılabilir.
5. Bundan sonraki araştırmalarda verideki hatalar düzeltilerek otomatik puanlama yapılarak elde edilen sonuçlar bu araştırmanın sonuçlarıyla karşılaştırılabilir.
6. Bu araştırmada veriler bilgisayar ortamına elle aktarılmıştır. Kağıt-kalem testleri üzerinde yürütülen daha sonraki araştırmalarda veri girişi optik karakter tanıma sistemleri aracılığıyla yapılarak elde edilen sonuçlar bu araştırmanın sonuçlarıyla karşılaştırılabilir.

Kaynaklar

Adesiji, K. M., Agbonifo, O. C., Adesuyi, A. T., & Olabode, O. (2016). Development of an automated descriptive text-based scoring system. *British Journal of Mathematics & Computer Science, 19*(4), 1-14. doi: 10.9734/BJMCS/2016/27558

Akar, A. (2017). *Türk dili tarihi: Dönem-eser-bibliyografya* (12. baskı). İstanbul: Ötüken.

Albano, A. (2016). equate: Observed-score linking and equating (version 2.0-5) [computer software package].

Almond, R. G. (2014). Using automated essay scores as an anchor when equating constructed response writing tests. *International Journal of Testing, 14*, 73–91. doi: 10.1080/15305058.2013.816309

Altman, D. G. (1991). *Practical statistics for medical research*. Boca Raton: CRC.

American Educational Research Association (AERA), American Pschological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, NJ: Educational Testing Service.

Araujo, J., & Born, D. G. (1985). Calculating percentage agreement correctly but writing its formula incorrectly. *The Behavior Analyst, 8*(2), 207-208.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jtla.org.

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(1), 87-101. doi: 10.1080/10705511.2014.934850

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*(2), 77-85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation, 18*(6), 1-13.

Bendermacher, N. (2010). Beyond alpha: Lower bounds for the reliability of tests. *Journal of Modern Applied Statistical Methods, 9*(1), 95-102. doi: 10.22237/jmasm/1272687000

Berg, P-C., & Gopinathan, M. (2017). *A deep learning ensemble approach to gender identification of tweet authors* (Master's thesis, Norwegian University of Science and Technology). Retrieved from https://brage.bibsys.no/xmlui/handle/11250/2458477

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted Kappa coefficients on the number of categories. Epidemiology, 7(2), 199-202.

Brookhart, S. M., & Nitko, A. J. (2015). *Educational assessment of students* (7th ed.). New Jersey: Pearson.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology, 46*(5), 423-429. doi: 10.1016/0895-4356(93)90018-V

Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal, 57*(9), 1318-1330. doi:10.1093/comjnl/bxt117

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, Y., Ben-Simon, A., & Hovav, M. (October, 2003). *The effect of specific language features on the complexity of systems for automated essay scoring.* Paper presented at the International Association of Educational Administration, Manchester.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *The Journal of Applied Psychology, 78*(1), 98-104, doi:10.1037/0021-9010.78.1.98

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Boston: Pearson.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. USA: Harcourt Brace Jovanovich College.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Deng, W., & Monfils, R. (2017). *Long-term impact of valid case criterion on capturing population-level growth under Item Response Theory equating* (Research Report 17-17). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (Research report 10-29). Princeton, NJ: Educational Testing Service.

Downing, S. M. (2009). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education*. New York, NY: Routledge.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. doi: 10.1111/bjop.12046

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Eker, S. (2017). *Çağdaş Türk dili* (11. baskı). Ankara: Grafiker.

Ercilasun, A. B. (2016). *Başlangıçtan yirminci yüzyıla Türk dili tarihi* (17. baskı). Ankara: Akçağ.

Eugenio, B. D., & Glass, M. (2004). The Kappa statistic: A second look. *Computational Linguistics, 30*(1), 95-101. https://doi.org/10.1162/089120104773633402

Frankel, J. R., Wallen, N. E., & Hyun, H. H. (2015). *How to design and evaluate research in education* (9th Ed.). New York: McGraw-Hill.

Frasco, M. (2018). *Package "Metrics". Evaluation metrics for machine learning* (version 0.1.4) [computer software package].

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2-18. Doi: 10.1037/a0024338

Gamer, M., Fellows, J., Lemon, I., & Singh, P. (2010). *Package "irr". Various coefficients of interrater reliability and agreement* (version 0.83) [computer software package].

Geisinger, K. F., & Usher-Tate, B. J. (2016). A brief history of educational testing and psychometrics. In C. S. Wells, M. Faulkner-Bond (Eds.), *Educational measurement from foundations to future* (pp. 3-20). New York: The Guilford.

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & Champlain, A. D. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education, 48*, 950-962. doi: 10.1111/medu.12517

Gonzalez, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. Switzerland: Springer.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13-34. https://doi.org/10.1207/S15327841MPEE0501_2

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. *Report of the Center for Educator Compensation Reform*. Retrieved from https://files.eric.ed.gov/fulltext/ED532068.pdf

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29-48. doi: 10.1348/000711006X126600

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement, 76*(4), 609-637. doi: 10.1177/0013164415596420

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149. https://doi.org/10.4992/psycholres1954.22.144

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups* (Doctoral dissertation). Retrieved from http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1865&context=etd

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Hagge, S. L., Liu, C., He, Y., Powers, S. J., Wang, W., & Kolen, M. J. (2011). A comparison of IRT and traditional equipercentile methods in mixed-format equating. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Essex: Pearson.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Haley, D. T. (2007). *Using a new inter-rater reliability statistic* (Report No. 2017/16). UK: The Open University.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York: Springer.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10*(1), 103-116.

He, Y. (2011). *Evaluating equating properties for mixed-format tests* (Doctoral dissertation). Retrieved from https://ir.uiowa.edu/etd/981/

Hoek, J., & Scholman, M. C. J. (2017). Evaluating discourse annotation: Some recent insights and new approaches. In H. Bunt (Ed.), *ACL Workshop on Interoperable Semantic Annotation* (pp. 1-13).

Ishioka, T., & Kameda, M. (2006). Automated Japanese essay scoring system based on articles written by experts. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, Sydney, 44, 233-240. doi: 10.3115/1220175.1220205

Jang, E-S., Kang, S-S., Noh, E-H., Kim, M-H., Sung, K-H., & Seong, T-J. (2014). KASS: Korean automatic scoring system for short-answer questions. *Proceedings of the 6th International Conference on Computer Supported Education*, Barcelona, 2, 226-230. doi: 10.5220/0004864302260230

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika, 35*(4), 401-415. https://doi.org/10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement, 34*, 111–117. https://doi.org/10.1177/001316447403400115

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005). Automatic essay grading with probabilistic latent semantic analysis. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, Ann Arbor, 5, 29-36. doi:

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different Item Response Theory scaling methods. *Educational and Psychological Measurement*, *71*(2), 362–379. https://doi.org/10.1177/0013164410375111

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking* (2nd ed.). USA: Springer.

Kubiszyn, T., & Borich, G. D. (2013). *Educational testing & measurement: Classroom application and practice* (10th ed.). Hoboken, NJ: Wiley.

Kumar, C. S., & Rama Sree, R. J. (2014). An attempt to improve classification accuracy through implementation of bootstrap aggregation with sequential minimal optimization during automated evaluation of descriptive answers. *Indian Journal of Science and Technology, 7*(9), 1369-1375. doi:

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and Best Practices in Content Analysis. *Journalism and Mass Communication Quarterly, 92*(4), 1-21. Doi: 10.1177/1077699015607338

LaFlair, G. T., Isbell, D., May, L. D. N., Arvizu, M. N. G., & Jamieson, J. (2017). Equating in small-scale language testing programs. *Language Testing, 34*(1), 127–144. https://doi.org/10.1177/0265532215620825

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lee, E., Lee, W-C., & Brennan, R. L. (2012). *Exploring equity properties in equating using AP® examinations* (Research Report 2012-4). USA: CollegeBoard.

Lilja, M. (2018). *Automatic essay scoring of Swedish essays using neural networks* (Doctoral dissertation, Uppsala University). Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1213688&dswid=9250

Liu, C., & Kolen, M. J. (2011). A comparison among IRT equating methods and traditional equating methods for mixed-format tests. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88-91. https://doi.org/10.3758/BF03192753

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement, 17*(3), 179-193. Retrieved from http://www.jstor.org/stable/1434833

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160. Retrieved from http://www.jstor.org/stable/1434012

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218. doi: 10.1207/s15326985ep3404_2

Martire, R. L. (2017). *Package 'rel'. Reliability Coefficients* (version 1.3.1) [computer software package].

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Belmont CA: Wadsworth/Thomson Learning.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and porffolio assessment* (pp. 61-73). New Jersey: Lawrence Erlbaum Associates, Inc.

Meyer, G. J. (1999). Simple procedures to estimate chance agreement and Kappa for the interrater reliability of response segments using the rorschach comprehensive system. *Journal of Personality Assessment, 72*(2), 230-255. doi: 10.1207/ S15327752JP720209

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

Milli Eğitim Bakanlığı (MEB), Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü. (2017a). *Akademik becerilerin izlenmesi ve değerlendirilmesi (ABİDE) 2016 8. sınıflar raporu*. Erişim Adresi: https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf

Milli Eğitim Bakanlığı (MEB), Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü. (2017b). *İzleme Değerlendirme Raporu 2016*. Erişim Adresi: http://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_06/23161120_2016_izleme_degYerlendirme_raporu.pdf

Mislevy, R. J., & Bock, R. D. (1997). *BILOG 3: Item analysis and test scoring with binary logistic models* [Computer software]. Mooresville, IN: Scientific Software.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (ETS Research Report 06-05). Princeton, NJ: Educational Testing Service.

Moses, T., von Davier, A. A., & Casabianca, J. (2004). *Loglinear smoothing: An alternative numerical approach using SAS* (ETS Research Report 04-27). Princeton, NJ: Educational Testing Service.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (seventh edition). Los Angeles, CA: Muthén & Muthén.

Olgar, S. (2015). *The integration of automated essay scoring systems into the equating process for mixed-format tests* (Doctoral Dissertation). Retrieved from http://diginole.lib.fsu.edu/islandora/object/fsu%3A253122

Osterlind, S. J. (2002). *Constructing test items: Multiple choice, constructed-response, performance, and other formats* (2nd ed.). USA: Kluwer Academic.

Page, E. B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan, 47*(5), 238–243. Retrieved from http://www.jstor.org/stable/20371545

Pang, X., Madera, E., Radwan, N., & Zhang, S. (2010). *A comparison of four test equating methods* (Research Report). Toronto: EQAO.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. California: Sage.

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education, 28*(2), 130-142. doi: 10.1080/08957347.2014.1002920

Preston, D., & Goodman, D. (2012). Automated Essay Scoring and The Repair of Electronics. Retrieved from https://www.semanticscholar.org/

Price, L. R. (2017). *Pschometric methods: Theory into practice*. New York: Guilford.

R Development Core Team. (2018). *R: A language and environment for statistical computing* (version 3.5.2). Vienna, Austria: R Foundation for Statistical Computing.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25-39. https://doi.org/10.1016/j.asw.2012.10.004

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: Validity, technical adequacy and implementation* (pp. 213-231). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Salkind, N. J. (2006). *Tests & measurement for people who (think they) hate tests & measurement*. California: Sage.

SAS Instutite. (2015). *Statistical analysis software* (version 9.4). Cary, NC: SAS Institute.

Senay, A., Delisle, J., Raynauld, J. P., Morin, S. N., & Fernandes, J. C. (2015). Agreement between physicians' and nurses' clinical decisions for the management of the fracture liaison service (4iFLS): the Lucky Bone™ program. *Osteoporosis International, 27*(4), 1569-1576. Doi: 10.1007/s00198-015-3413-6

Shankar, V., & Bangdiwala, S. I. (2014). Observer agreement paradoxes in 2x2 tables: Comparison of agreement measures. *BMC Medical Research Methodology*. Advance online publication. https://doi.org/10.1186/1471-2288-14-100

Shermis, M. D., & Burnstein, J. (2003). *Automated essay scoring*. Mahvah, New Jersey: Lawrence Erlbaum Associates.

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing enviroment. In V. J. Shute, B. J. Becker (Eds.), *Innovative assessment for the 21st century*. New York: Springer.

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268. https://doi.org/10.1093/ptj/85.3.257

Siriwardhana, D. D., Walters, K., Rait, G., Bazo-Alvarez, J. C., & Weerasinghe, M. C. (2018). Cross-cultural adaptation and psychometric evaluation of the Sinhala version of Lawton Instrumental Activities of Daily Living Scale. *Plos One, 13*(6), 1-20. https://doi.org/10.1371/journal.pone.0199820

Smolentzov, A. (2013). *Automated essay scoring: Scoring essays in Swedish* (Bachelor's Thesis, Stockholm University, Stockholm, Sweden). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-87266

Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* (Doctoral Dissertation, University of Florida). Retrieved from https://archive.org/details/effectofmultidim00spen

Stevens, J. (2009). *Applied multivariate statistics for the social sciences* (5th edition). New York: Taylor & Francis.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement, 7*(2), 201-210. doi: 10.1177/014662168300700208

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed). London: Pearson.

Taghipour, K., & Tou Ng, H. (2016). A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, 1882-1891. doi: 10.18653/v1/D16-1193

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*(4), 329-346. doi: 10.1111/j.1745-3984.2000.tb01090.x.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. doi: 10.1037/a0023353

Tsai, M. H. (2012). The consistency between human raters and an automated essay scoring system in grading high school students' English writing. *Action in Teacher Education, 34*(4), 328-335. doi: 10.1080/01626620.2012.717033

Vanbelle, S. (2016). A new interpretation of the weighted Kappa coefficients. *Psychometrika, 81*(2), 399-410. https://doi.org/10.1007/s11336-014-9439-4

von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*(2). Retrieved from http://www.jtla.org.

Wang, Y., Wei, Z., Zhou, Y., & Huang, X. (2018, November). Automatic essay scoring incorporating rating schema via reinforcement learning. In E. Reloff, D. Chiang, H. Julia & T. Jun'ichi (Eds.), *Empirical Methods in Natural Language Processing* (pp. 791-797).

Weiss, D. J., & Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG* (Technical Report). MN: Assessment Systems Corporation.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

Wolf, R. (2013). *Assessing the impact of characteristics of the test, common-items, and examinees on the preservation of equity properties in mixed-format test equating* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*(61), 1-9. https://doi.org/10.1186/1471-2288-13-61

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers*. Singapore: Springer.

Yoes, M. E. (1996). *User's manual for the XCALIBRE marginal maximum-likelihood estimation program* [Computer software]. St. Paul, MN: Assessment Systems Corp.

Zu, J., & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement, 47*(4), 395-412. https://doi.org/10.1111/j.1745-3984.2010.00120.x

EK-A: Milli Eğitim Bakanlığı ABİDE 2016 8. Sınıf Türkçe Testi Örnek Maddeleri ve Dereceli Puanlama Anahtarları

İSTANBUL DEĞİŞİYOR

İstanbul'da beklenmedik bir şekilde nüfusun artması; gecekonduların çoğalmasına, altyapının kurulmasında sorunlar yaşanmasına neden olmaktadır. Kentlerin dokusunda ise önemli değişmeler görülmektedir.

İstanbul'un eski semtleri olan Beyoğlu, Sirkeci, Eminönü ve Beyazıt'ta ara sokaklarda taş veya ahşap binalar, birbirini kesen dar sokaklar ve caddeler yer almaktadır. Bakırköy, Caddebostan, Etiler, Nişantaşı, Levent gibi yeni semtlerde çoğu kez doğrusal uzanış gösteren ve birbirini dik kesen cadde ve sokaklar vardır. Ataköy, Bahçeşehir gibi planlı olarak kurulan semtlerde ise daha düzenli caddeler yer almakta, çok katlı binalar yapılmaktadır.

7 - 9. soruları yukarıdaki metne göre yanıtlayınız.

7. **Nüfusun olağan dışı artması beraberinde hangi sorunları getirmektedir? Yazınız.**

8. **Metni göz önünde bulundurduğunuzda fotoğrafta görülen yer İstanbul'un hangi semti olabilir? Gerekçesiyle yazınız.**

9. **Metinde altı çizili sözcükle anlatılmak istenen aşağıdakilerden hangisidir?**

A) Yapı
B) Büyüklük
C) Kapladığı alan
D) Gelişmişlik düzeyi

"İSTANBUL DEĞİŞİYOR" Bağlamına Ait Puanlama Anahtarı

Soru No:	**5**
Soru Kodu:	**T-2016-0007**
Bağlam Adı:	**İSTANBUL DEĞİŞİYOR**
DOĞRU YANIT-(2 PUAN)Açıklama	Gecekonduların çoğalması VE altyapı problemlerinin artması sorunlarının her ikisine birden vurgu yapan YA DA bu sorunları genelleyen ifadeleri içeren yanıtlar

Örnek Yanıtlar	Çarpık kentleşme ve imar sorunları
	Gecekonduların artması ve altyapı problemleri
	Gecekonduların artması ve yapılan yolların yeterli olmaması
KISMI DOĞRU-(1 puan) Açıklama	Metinde geçen iki sorundan "gecekonduların çoğalması" YA DA "alt yapı problemlerinin artması" ifadelerinden sadece birini içeren yanıtlar
YANLIŞ YANIT-(0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar verir
Örnek Yanıtlar	Kentlerin dokusunda önemli değişmeler görülmektedir
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	**6**
Soru Kodu:	**T-2016-0008**
Bağlam Adı:	**İSTANBUL DEĞİŞİYOR**
DOĞRU YANIT-(2 PUAN)Açıklama	"Beyoğlu, Sirkeci, Eminönü, Beyazıt semtlerinden birinin, birkaçının veya hepsinin adını içeren, gerekçe olarak "Ara sokaklarda taş veya ahşap binalar bulunur." YA DA "Birbirini kesen dar sokaklar ve caddeler bulunur." ifadelerinden birini içeren yanıtlar
Örnek Yanıtlar	Beyoğlu çünkü evler ahşap.
	Sirkeci, Eminönü çünkü ara sokaklarda taş veya ahşap binalar bulunur.
KISMİ DOĞRU-(1 puan) Açıklama	Sadece semt adını içeren ancak gerekçenin yazılmadığı yanıtlar
Örnek Yanıtlar	Beyoğlu
	Eminönü, Beyazıt
	Beyoğlu, Sirkeci, Eminönü, Beyazıt
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	**7**
Soru Kodu:	**T-2016-0009**
Bağlam Adı:	**İSTANBUL DEĞİŞİYOR**
Doğru Yanıt	**A**

BASINDA OBEZİTE

10.01.2015

12 Yaş Altı Çocuklarda Mobil Cihazların Kullanımının Yasaklanması İçin Bir Sebep: Obezite

Video oyunları ve televizyon, obezitenin artması ile ilişkilidir. Odasında bu tür cihazları kullanmasına izin verilen çocuklarda obezite görülme sıklığı %30 oranında artmaktadır. Obez olan çocukların %30'unda diyabet ortaya çıkmakta, kalp krizi ve erken felç riski artmakta ve ortalama yaşam süresi kısalmaktadır.

15.12.2014
Çocukluk Döneminde Risk: Obezite

Anne ve babanın obez olması, çocuğun yeme alışkanlığı bakımından anne ve babasını örnek alması, çocukların televizyon ve bilgisayar başında çok zaman geçirmesi, stres, kaygı gibi unsurlar çocukluk döneminde obezitenin oluşmasına neden olmaktadır.

10.11.2014
Çocukları Obez Olan Ailelere Para Cezası Geliyor!

Porto Riko'da hükümet, obeziteyle mücadele amaçlı, çocukları fazla kilolu olan anne ve babalara 800 dolara kadar para cezası verilmesini planlıyor. Gelecek nesillerin daha sağlıklı olması için bu uygulamanın yararlı olacağını düşünenlerin sayısı ülkede oldukça fazla.

10 - 12. soruları yukarıdaki metne göre yanıtlayınız.

10. **Gazetelerde obeziteyle ilgili haberlere sıklıkla yer verilmesinin nedeni nedir? Bir ya da iki cümleyle yazınız.**

11. **Mobil cihazların kullanımı obeziteyi neden arttırır? Bir ya da iki cümleyle yazınız.**

12. **Gazete haberlerine göre aşağıdakilerden hangisi söylenebilir?**

A) Obezite ve diyabet birbirleriyle ilişkilidir.
B) Televizyon izlemeyen çocuklar obeziteye yakalanmıyor.
C) Porto Riko'daki para cezası birçok ülkeye örnek olmuştur.
D) Obezite yalnızca çocukluk döneminde ortaya çıkan bir sorundur.

"BASINDA OBEZİTE" Bağlamına Ait Puanlama Anahtarı

Soru No:	**10**
Soru Kodu:	**T-2016-0010**
Bağlam Adı:	**BASINDA OBEZİTE**

DOĞRU YANIT- (2 PUAN)Açıklama	Obezite ile ilgili bilinçlendirmeye vurgu yapan yanıtlar
Örnek Yanıtlar	"Obezitenin yaygınlaşmasını önlemek için."
	"Halkı bilinçlendirmek için."
	"Obezitenin bir hastalık olduğuna dikkat çekmek."
	"Halkı uyarmak için."
	"Aileleri bilinçlendirmek için."
	"Anne ve babaların önlem almasını sağlamak için." vb.
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
Örnek Yanıtlar	Para cezasını haber vermek için
BOŞ-Açıklama	- Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	**11**
Soru Kodu:	**T-2016-0011**
Bağlam Adı:	**BASINDA OBEZİTE**
DOĞRU YANIT- (1 PUAN) Açıklama	"Uzun süre hareketsiz kalma, çocukların televizyon ve bilgisayar başında çokça vakit geçirmesi" ifadelerini içeren yanıtlar
Örnek Yanıtlar	" Çocukların bilgisayar ve televizyon başında çok zaman geçirmesi."
	"Çocukların bilgisayar başında çok zaman geçirmesinden dolayı hareketsiz kalması."
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	**12**
Soru Kodu:	**T-2016-0012**
Bağlam Adı:	**BASINDA OBEZİTE**
Doğru Yanıt	**A**

EK-B: İki Puanlayıcı ve Üç Kategorili Puanlamada Durum 1'e İlişkin Matris

		Puanlayıcı 1	
		Doğru (0)	Yanlış (1 ve 2)
Puanlayıcı 2	Doğru (0)	a	b
	Yanlış (1 ve 2)	c	d

EK-C: İki Puanlayıcı ve Üç Kategorili Puanlamada Durum 2'ye İlişkin Matris

		Puanlayıcı 1	
		Doğru (1)	Yanlış (0 ve 2)
Puanlayıcı 2	Doğru (1)	a	b
	Yanlış (0 ve 2)	c	d

EK-Ç: BLSTM Yöntemi %10, %20 ve %33 Test Veri Oranlarıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin İstatistikler

	Otomatik Puanlama					
Test Veri Yüzdesi	%10		%20		%33	
Kitapçık	A_1	B_1	A_1	B_1	A_1	B_1
Madde Sayısı	18	18	18	18	18	18
Örneklem Sayısı	607	584	607	584	607	584
Ortalama	13,259	14,300	13,283	14,361	13,273	14,346
Standart Sapma	4,331	4,777	4,333	4,765	4,313	4,760
Medyan (Ortanca)	13	15	14	15	14	15
Minimum Değer	2	0	2	0	2	1
Maksimum Değer	23	23	23	23	23	23
Çarpıklık	-,208	-,520	-,218	-,538	-,209	-,518
Güvenirlik (Alfa)	,746	,784	,746	,783	,747	,786
Güvenirlik (Omega)	,857	,885	,856	,882	,858	,884

EK-D: BLSTM Yöntemi %10 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin KTK'ya Dayalı Madde İstatistikleri

%10 Test Veri Oranı ile Gerçekleştirilen Otomatik Puanlama İşlemi

Kitapçık A_1				Kitapçık B_1			
Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)	Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)
1	,774	,149	,746	1	,909	,380	,777
2	,941	,274	,741	2	,630	,344	,776
5	,550	,369	,731	3	,846	,472	,771
6	,341	,310	,736	4	,726	,468	,769
7*	,685	,510	,728	5*	,722	,509	,773
8*	,636	,452	,735	6*	,647	,563	,767
9	,717	,260	,739	7	,716	,387	,774
10	,580	,417	,728	8	,565	,360	,775
11	,643	,344	,733	9	,644	,334	,776
12	,530	,280	,738	10	,570	,302	,778
13	,746	,440	,727	11*	,509	,584	,773
14	,740	,501	,723	12*	,529	,568	,769
15*	,697	,257	,751	13	,738	,408	,773
16	,717	,319	,735	14	,260	,021	,793
17	,257	,059	,752	17*	,697	,402	,779
18*	,301	,530	,729	18	,509	,456	,769
19*	,273	,604	,723	19	,678	,370	,774
20	,542	,482	,722	20	,305	,417	,772

EK-E: BLSTM Yöntemi %20 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin KTK'ya Dayalı Madde İstatistikleri

%20 Test Veri Oranı ile Gerçekleştirilen Otomatik Puanlama İşlemi							
	Kitapçık A_1				Kitapçık B_1		
Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)	Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)
1	,774	,151	,747	1	,909	,376	,776
2	,936	,277	,742	2	,630	,348	,774
5	,550	,359	,733	3	,839	,476	,770
6	,341	,314	,736	4	,726	,471	,768
7*	,683	,516	,728	5*	,718	,503	,772
8*	,641	,453	,736	6*	,650	,582	,764
9	,717	,264	,740	7	,716	,379	,773
10	,618	,399	,730	8	,541	,348	,774
11	,633	,336	,735	9	,627	,324	,776
12	,530	,280	,739	10	,570	,309	,777
13	,746	,441	,728	11*	,518	,592	,772
14	,728	,503	,724	12*	,558	,547	,769
15*	,693	,273	,750	13	,738	,422	,771
16	,717	,313	,737	14	,260	,024	,792
17	,257	,057	,753	17*	,707	,446	,775
18*	,307	,508	,732	18	,503	,447	,768
19*	,274	,615	,723	19	,678	,377	,773
20	,542	,487	,723	20	,324	,352	,774

EK-F: BLSTM Yöntemi %33 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin KTK'ya Dayalı Madde İstatistikleri

%33 Test Veri Oranı ile Gerçekleştirilen Otomatik Puanlama İşlemi							
	Kitapçık A_1				Kitapçık B_1		
Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)	Madde No	Güçlük	Ayırt Edicilik	Alfa (Madde Çıkarıldığında)
1	,774	,145	,748	1	,909	,377	,779
2	,941	,270	,743	2	,630	,337	,778
5	,550	,363	,734	3	,851	,490	,772
6	,341	,307	,738	4	,726	,469	,771
7*	,686	,491	,732	5*	,728	,540	,772
8*	,644	,445	,737	6*	,648	,587	,767
9	,717	,263	,741	7	,716	,387	,775
10	,605	,402	,731	8	,577	,298	,780
11	,629	,372	,733	9	,675	,313	,779
12	,530	,282	,740	10	,570	,318	,779
13	,746	,451	,729	11*	,511	,583	,775
14	,750	,499	,725	12*	,556	,607	,767
15*	,676	,300	,749	13	,738	,404	,774
16	,717	,318	,737	14	,260	,032	,794
17	,257	,066	,754	17*	,680	,416	,779
18*	,303	,516	,733	18	,498	,447	,771
19*	,279	,606	,725	19	,678	,375	,776
20	,542	,479	,725	20	,272	,376	,776

EK G: Ki Kare Tablosu

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

EK H: BLSTM Yöntemi %10 Test Veri Oranıyla Gerçekleştirilen Otomatik Puanlamaya Yönelik Ön Düzgünleştirme Modelinin Belirlenmesi

		A_1 Kitapçığı				B_1 Kitapçığı			
		ki kare	sd	AIC	BIC	ki kare	sd	AIC	BIC
Model 1	x a	407,934	92	712,080	719,741	283,950	102	637,095	645,057
Model 2	x a xa	339,333	91	675,363	685,578	271,781	101	628,111	638,727
Model 3	x x^2 a a^2	291,460	90	636,921	649,691	230,749	100	585,246	598,516
Model 4	x x^2 a a^2 xa	100,428	89	431,873	447,196	117,609	99	452,665	468,588
Model 5	x x^2 x^3 a a^2 a^3	277,263	88	630,180	648,058	213,272	98	569,474	588,052
Model 6	x x^2 x^3 a a^2 a^3 xa	**89,024**	**87**	**423,683**	**444,114**	95,256	97	437,623	458,855
Model 7	x x^2 x^3 x^4 a a^2 a^3 a^4	277,247	86	633,740	656,725	195,451	96	560,901	584,787
Model 8	x x^2 a a^2 xa x^2a xa^2 x^2a^2	90,069	86	426,025	449,009	95,271	96	431,645	455,530
Model 9	x x^2 x^3 x^4 a a^2 a^3 a^4 xa	89,240	85	427,403	452,942	76,787	95	428,650	455,190
Model 10	x x^2 x^3 a a^2 a^3 xa x^2a xa^2 x^2a^2	83,774	84	425,971	454,064	**70,767**	**94**	**431,364**	**460,557**
Model 11	x x^2 x^3 x^4 a a^2 a^3 a^4 xa x^2a xa^2 x^2a^2	82,402	82	428,782	461,982	70,154	92	432,816	467,317

EK I: BLSTM Yöntemi %20 Test Veri Oranıyla Gerçekleştirilen Otomatik Puanlamaya Yönelik Ön Düzgünleştirme Modelinin Belirlenmesi

		A_1 Kitapçığı				B_1 Kitapçığı			
		ki kare	sd	AIC	BIC	ki kare	sd	AIC	BIC
Model 1	x a	427,993	93	723,758	731,451	322,613	100	666,699	674,603
Model 2	x a xa	344,817	92	678,166	688,423	313,837	99	657,653	668,192
Model 3	x x^2 a a^2	292,090	91	639,190	652,012	263,721	98	604,637	617,810
Model 4	x x^2 a a^2 xa	94,937	90	430,925	446,312	130,789	97	461,538	477,346
Model 5	x x^2 x^3 a a^2 a^3	281,204	89	635,730	653,680	235,313	96	579,617	598,060
Model 6	x x^2 x^3 a a^2 a^3 xa	**83,726**	**88**	**420,198**	**440,713**	106,520	95	441,928	463,006
Model 7	x x^2 x^3 x^4 a a^2 a^3 a^4	279,182	87	637,661	660,740	217,412	94	570,559	594,271
Model 8	x x^2 a a^2 xa x^2a xa^2 x^2a^2	86,017	87	422,949	446,029	117,948	94	438,729	462,441
Model 9	x x^2 x^3 x^4 a a^2 a^3 a^4 xa	83,029	86	423,639	449,283	**88,630**	**93**	**433,225**	**459,572**
Model 10	x x^2 x^3 a a^2 a^3 xa x^2a xa^2 x^2a^2	76,371	85	420,851	449,059	86,795	92	439,300	468,282
Model 11	x x^2 x^3 x^4 a a^2 a^3 a^4 xa x^2a xa^2 x^2a^2	74,577	83	423,478	456,814	82,709	90	437,066	471,318

EK İ: BLSTM Yöntemi %33 Test Veri Oranıyla Gerçekleştirilen Otomatik Puanlamaya Yönelik Ön Düzgünleştirme Modelinin Belirlenmesi

		A_1 Kitapçığı				B_1 Kitapçığı			
		ki kare	sd	AIC	BIC	ki kare	sd	AIC	BIC
Model 1	x a	370,380	90	688,540	696,137	327,949	101	668,561	676,494
Model 2	x a xa	295,692	89	638,498	648,628	310,809	100	656,061	666,639
Model 3	x x^2 a a^2	250,769	88	598,599	611,262	257,310	99	606,212	619,434
Model 4	x x^2 a a^2 xa	80,769	87	410,537	425,733	154,401	98	466,195	482,062
Model 5	x x^2 x^3 a a^2 a^3	239,149	86	592,492	610,220	234,586	97	590,822	609,333
Model 6	x x^2 x^3 a a^2 a^3 xa	**67,615**	**85**	**401,037**	**421,298**	124,430	96	451,638	472,793
Model 7	x x^2 x^3 x^4 a a^2 a^3 a^4	236,363	84	594,430	617,223	214,097	95	581,487	605,286
Model 8	x x^2 a a^2 xa x^2a xa^2 x^2a^2	66,583	84	401,826	424,620	128,907	95	446,406	470,205
Model 9	x x^2 x^3 x^4 a a^2 a^3 a^4 xa	65,584	83	402,880	428,206	103,067	94	442,072	468,516
Model 10	x x^2 x^3 a a^2 a^3 xa x^2a xa^2 x^2a^2	64,712	82	403,231	431,089	**82,856**	**93**	**439,162**	**468,250**
Model 11	x x^2 x^3 x^4 a a^2 a^3 a^4 xa x^2a xa^2 x^2a^2	62,756	80	405,163	438,086	82,939	91	441,622	475,999

EK-J: %10, %20 ve %33 Test Veri Oranlarıyla BLSTM Yöntemi Kullanılarak Gerçekleştirilen Otomatik Puanlama Sonucunda A_1 ve B_1 Kitapçıklarına İlişkin Test Verilerinin Faktör Analizine Uygunluğu

		A_1 Kitapçığı			B_1 Kitapçığı		
BLSTM %10	Kaiser-Meyer-Olkin		,856			,898	
	Bartlett	χ^2	sd	p	χ^2	sd	p
		1283,4	153	,000*	1445,5	153	,000*
BLSTM %20	Kaiser-Meyer-Olkin		,855			,900	
	Bartlett	χ^2	sd	p	χ^2	Sd	p
		1288,6	153	,000*	1419,5	153	,000*
BLSTM %33	Kaiser-Meyer-Olkin		,851			,898	
	Bartlett	χ^2	sd	p	χ^2	sd	p
		1299,5	153	,000*	1460,2	153	,000*

* p<,05

EK-K: %10 Test Veri Oranıyla Gerçekleştirilen Otomatik Puanlama İçin Her Bir Test Formundaki Maddelerin Faktör Yükleri

Kitapçık A_1				Kitapçık B_1			
Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1
1	0,238	12	0,420	1	0,711	10	0,428
2	0,620	13	0,667	2	0,476	11	0,602
5	0,530	14	0,777	3	0,773	12	0,564
6	0,457	15	0,240	4	0,690	13	0,610
7	0,527	16	0,473	5	0,560	14	0,030
8	0,457	17	0,094	6	0,579	17	0,406
9	0,415	18	0,523	7	0,564	18	0,650
10	0,615	19	0,659	8	0,496	19	0,547
11	0,539	20	0,691	9	0,478	20	0,641

EK-L: %20 Test Veri Oranıyla Gerçekleştirilen Otomatik Puanlama İçin Her Bir Test Formundaki Maddelerin Faktör Yükleri

Kitapçık A_1				Kitapçık B_1			
Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1
1	0,240	12	0,417	1	0,704	10	0,442
2	0,650	13	0,664	2	0,484	11	0,589
5	0,516	14	0,774	3	0,767	12	0,549
6	0,460	15	0,252	4	0,694	13	0,624
7	0,533	16	0,457	5	0,557	14	0,035
8	0,457	17	0,091	6	0,601	17	0,452
9	0,426	18	0,503	7	0,554	18	0,637
10	0,607	19	0,670	8	0,477	19	0,550
11	0,514	20	0,698	9	0,460	20	0,539

EK-M: %33 Test Veri Oranıyla Gerçekleştirilen Otomatik Puanlama İçin Her Bir Test Formundaki Maddelerin Faktör Yükleri

Kitapçık A_1				Kitapçık B_1			
Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1	Madde No	Faktör 1
1	0,226	12	0,417	1	0,700	10	0,450
2	0,627	13	0,674	2	0,469	11	0,586
5	0,515	14	0,784	3	0,812	12	0,605
6	0,449	15	0,276	4	0,695	13	0,602
7	0,505	16	0,471	5	0,604	14	0,044
8	0,451	17	0,113	6	0,607	17	0,419
9	0,416	18	0,514	7	0,561	18	0,626
10	0,610	19	0,656	8	0,413	19	0,550
11	0,567	20	0,690	9	0,448	20	0,598

EK-N: BLSTM Yöntemi %10 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin MTK Model Veri Uyumu

		A_1 Kitapçığı				B_1 Kitapçığı			
		-2LL	ki kare	sd	SH	-2LL	ki kare	sd	SH
Model 1	1PLM ve PCM	13627	1244,884	322	,431	12821	1194,921	322	,440
Model 2	1PLM ve GPCM	13596	1230,041	317	,465	12821	1230,002	317	,462
Model 3	1PLM ve GRM	13536	1111,773	317	,493	12762	1124,752	317	,492
Model 4	2PLM ve GPCM	**12800**	**452,598**	**304**	**,481**	**12082**	**475,180**	**304**	**,463**
Model 5	2PLM ve GRM	12813	453,923	304	,486	12088	466,172	304	,465

EK-O: BLSTM Yöntemi %20 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin MTK Model Veri Uyumu

		A_1 Kitapçığı				B_1 Kitapçığı			
		-2LL	ki kare	sd	SH	-2LL	ki kare	sd	SH
Model 1	1PLM ve PCM	13657	1264,760	322	,432	12866	1172,249	322	,441
Model 2	1PLM ve GPCM	13628	1237,813	317	,465	12871	1189,869	317	,462
Model 3	1PLM ve GRM	13578	1118,384	317	,493	12820	1100,679	317	,492
Model 4	2PLM ve GPCM	**12842**	**466,323**	**304**	**,482**	**12178**	**445,380**	**304**	**,466**
Model 5	2PLM ve GRM	12861	485,936	304	,487	12189	445,077	304	,469

EK-Ö: BLSTM Yöntemi %33 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin MTK Model Veri Uyumu

		A_1 Kitapçığı				B_1 Kitapçığı			
		-2LL	ki kare	sd	SH	-2LL	ki kare	sd	SH
Model 1	1PLM ve PCM	13645	1250,580	322	,432	12733	1246,761	322	,442
Model 2	1PLM ve GPCM	13612	1233,980	317	,465	12745	1281,618	317	,463
Model 3	1PLM ve GRM	13561	1096,509	317	,492	12705	1157,796	317	,493
Model 4	2PLM ve GPCM	**12822**	**437,061**	**304**	**,483**	**11997**	**424,949**	**304**	**,462**
Model 5	2PLM ve GRM	12837	446,678	304	,488	12004	420,582	304	,465

EK-P: BLSTM Yöntemi %10 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin MTK'ya Dayalı Madde Parametreleri

Kitapçık A_1				Kitapçık B_1			
Madde No	a	b_1	b_2	Madde No	a	b_1	b_2
1	0,219	-3,432		1	0,917	-2,128	
2	0,576	-3,325		2	0,522	-0,724	
5	0,535	-0,277		3	0,997	-1,542	
6	0,492	0,922		4	0,824	-0,984	
7	0,412	-0,870	-1,150	5	0,416	-0,610	-1,710
8	0,355	-1,250	-0,500	6	0,480	-0,980	-0,520
9	0,463	-1,357		7	0,627	-1,092	
10	0,690	-0,366		8	0,547	-0,338	
11	0,585	-0,733		9	0,517	-0,798	
12	0,465	-0,181		10	0,468	-0,409	
13	0,819	-1,098		11	0,405	2,100	-2,160
14	1,035	-0,955		12	0,460	-0,380	0,100
15	0,213	-2,650	-1,350	13	0,696	-1,146	
16	0,535	-1,223		14	0,273	2,323	
17	0,295	2,208		17	0,375	-2,860	-0,210
18	0,483	0,450	1,890	18	0,753	-0,030	
19	0,554	1,550	0,450	19	0,613	-0,895	
20	0,840	-0,172		20	0,777	0,867	

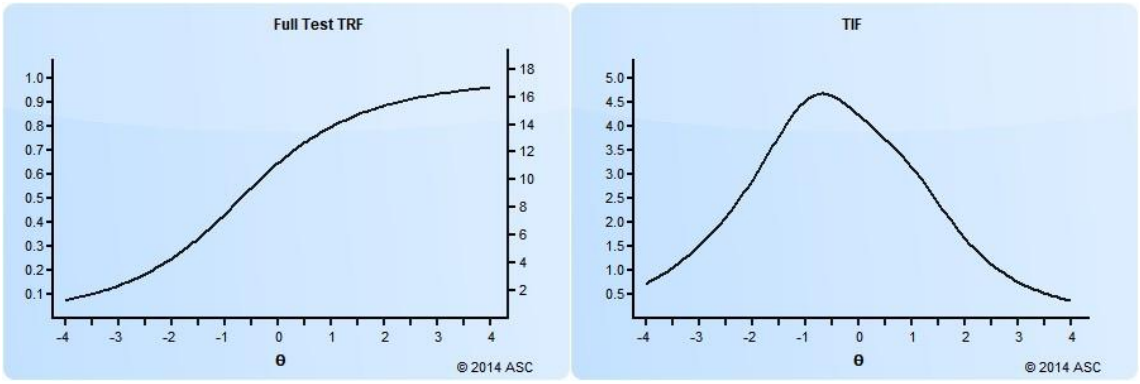
EK-R: BLSTM Yöntemi %20 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin MTK'ya Dayalı Madde Parametreleri

Kitapçık A_1				Kitapçık B_1			
Madde No	a	b_1	b_2	Madde No	a	b_1	b_2
1	0,222	-3,391		1	0,901	-2,146	
2	0,603	-3,129		2	0,531	-0,717	
5	0,519	-0,283		3	0,983	-1,507	
6	0,502	0,909		4	0,840	-0,976	
7	0,418	-0,890	-1,100	5	0,409	-0,620	-1,680
8	0,355	-1,230	-0,570	6	0,506	-0,950	-0,540
9	0,469	-1,345		7	0,618	-1,103	
10	0,675	-0,551		8	0,526	-0,219	
11	0,549	-0,709		9	0,497	-0,721	
12	0,463	-0,181		10	0,487	-0,399	
13	0,811	-1,103		11	0,404	1,850	-2,000
14	1,018	-0,909		12	0,446	-0,580	-0,030
15	0,232	-2,890	-0,930	13	0,719	-1,125	
16	0,519	-1,249		14	0,273	2,324	
17	0,295	2,208		17	0,422	-2,680	-0,270
18	0,449	0,460	1,880	18	0,727	-0,012	
19	0,577	1,450	0,520	19	0,617	-0,892	
20	0,860	-0,170		20	0,615	0,887	

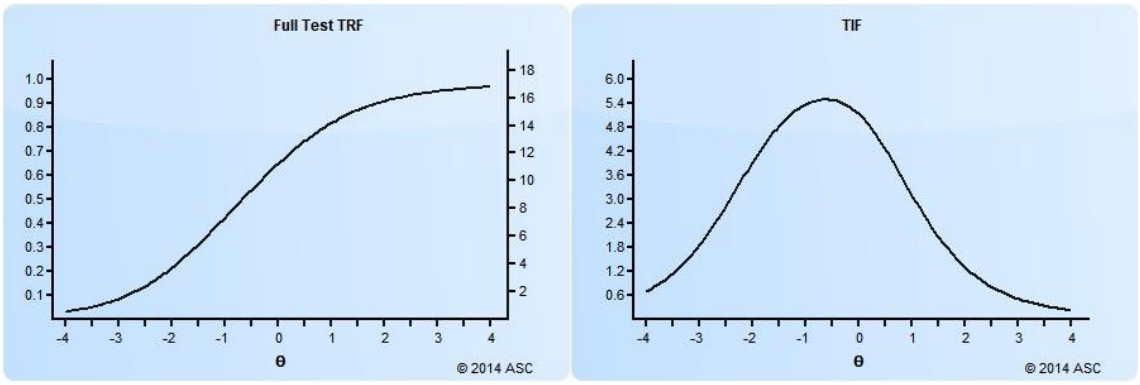
EK-S: BLSTM Yöntemi %33 Test Veri Oranıyla Otomatik Puanlama Gerçekleştirilen Testlere İlişkin MTK'ya Dayalı Madde Parametreleri

Kitapçık A_1				Kitapçık B_1			
Madde No	a	b_1	b_2	Madde No	a	b_1	b_2
1	0,210	-3,561		1	0,912	-2,131	
2	0,580	-3,309		2	0,514	-0,733	
5	0,517	-0,285		3	1,104	-1,510	
6	0,480	0,937		4	0,841	-0,975	
7	0,394	-1,020	-1,110	5	0,460	-0,610	-1,640
8	0,351	-1,330	-0,550	6	0,516	-0,960	-0,500
9	0,462	-1,361		7	0,627	-1,092	
10	0,677	-0,488		8	0,463	-0,456	
11	0,624	-0,638		9	0,486	-1,030	
12	0,462	-0,182		10	0,492	-0,396	
13	0,835	-1,089		11	0,398	1,910	-2,010
14	1,045	-1,000		12	0,519	-0,550	0,020
15	0,250	-3,130	-0,350	13	0,683	-1,160	
16	0,535	-1,225		14	0,277	2,292	
17	0,303	2,160		17	0,398	-2,960	0,110
18	0,462	0,470	1,890	18	0,723	0,011	
19	0,555	1,390	0,590	19	0,622	-0,887	
20	0,839	-0,174		20	0,705	1,083	

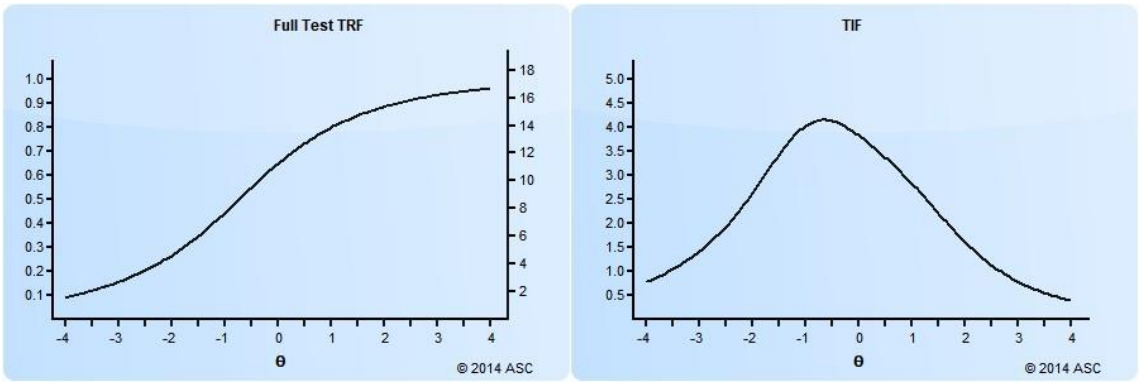
EK Ş: Gerçek Puanlayıcılar ve BLSTM Yöntemi ile %10, %20 ve %33 Test Veri Oranı ile Puanlanan A_1 ve B_1 Testlerinin Test Karakteristik Eğrileri ve Test Bilgi Fonksiyonu Grafikleri



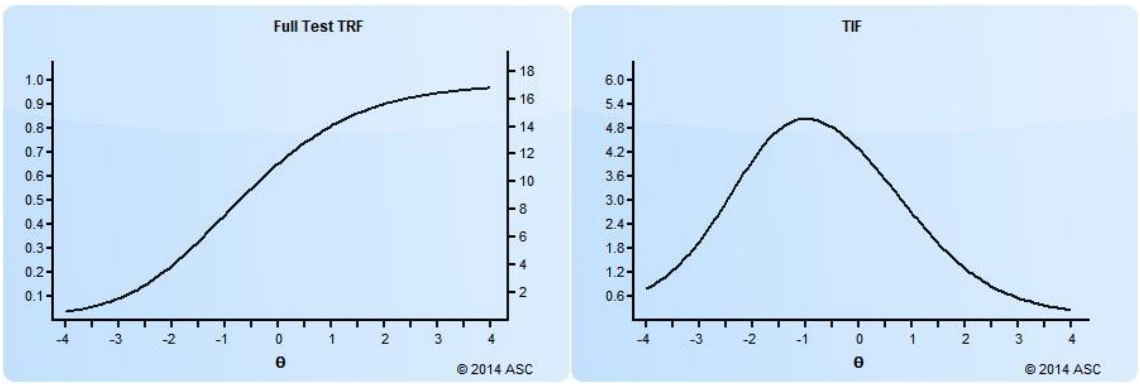
Gerçek Puanlayıcı A_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği



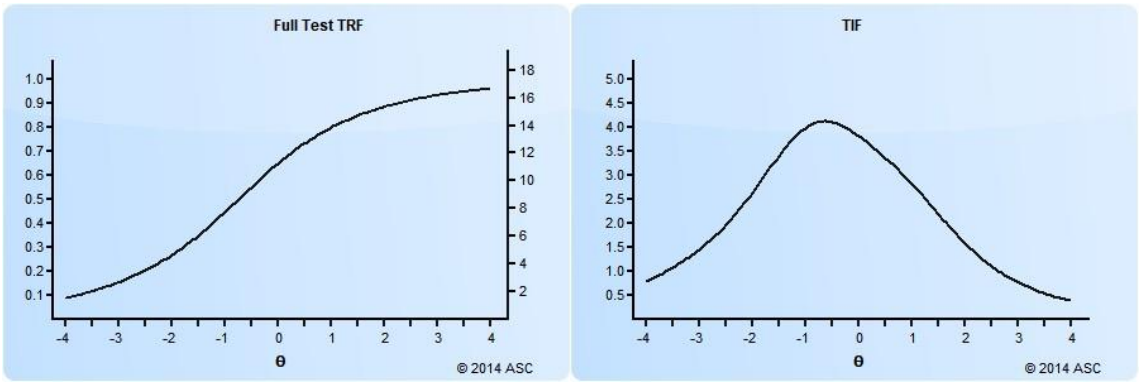
Gerçek Puanlayıcı B_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği



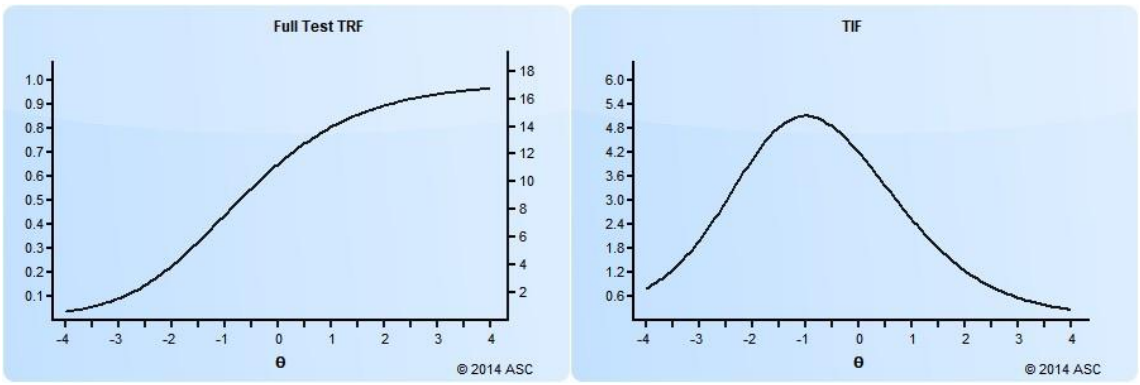
Otomatik Puanlayıcı (%10) A_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği



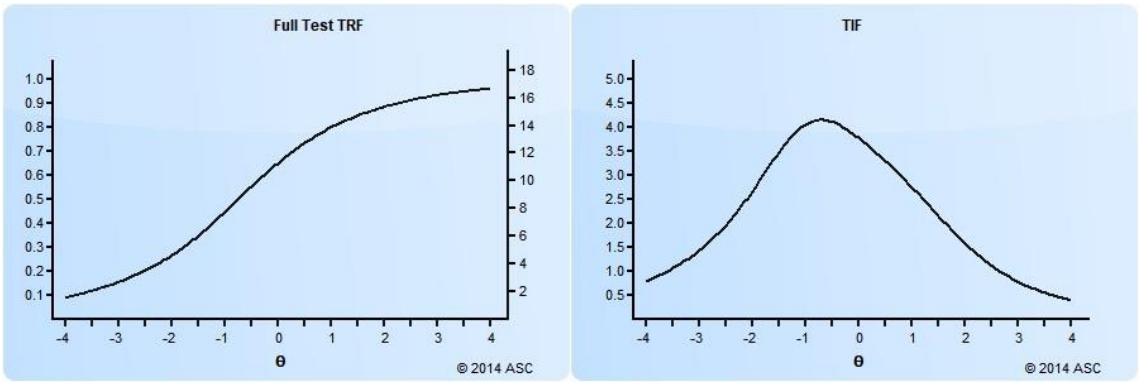
Otomatik Puanlayıcı (%10) B_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği



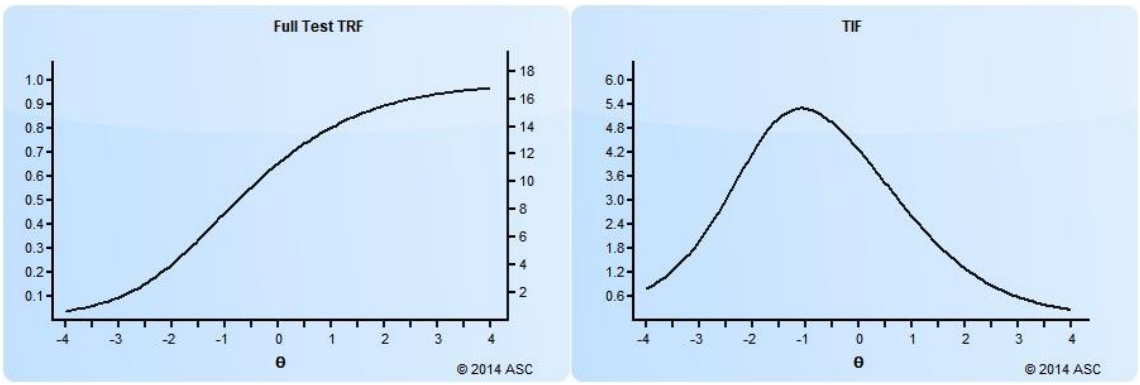
Otomatik Puanlayıcı (%20) A_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği



Otomatik Puanlayıcı (%20) B_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği



Otomatik Puanlayıcı (%33) A_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği



Otomatik Puanlayıcı (%33) B_1 Kitapçığı Sırasıyla Test Karakteristik Eğrisi ve Test Bilgi Fonksiyonu Grafiği

EK-T: Etik Komisyonu Onay Bildirimi

T.C.
HACETTEPE ÜNİVERSİTESİ
Rektörlük

30 Kasım 2017

Sayı : 35853172/ 433- 4042

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi: 14.11.2017 tarih ve 2352 sayılı yazınız.

Enstitünüz Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı doktora programı öğrencilerinden **İbrahim UYSAL'ın** **Prof. Dr. Nuri DOĞAN** danışmanlığında yürüttüğü **"Karma Formattaki Testlerde Otomatik Açık Uçlu Madde Puanlamanın Test Eşitleme Üzerindeki Etkisi"** başlıklı tez çalışması, Üniversitemiz Senatosu Etik Komisyonunun **22 Kasım 2017** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

Bilgilerinizi ve gereğini rica ederim.

Prof. Dr. Rahime M. NOHUTCU
Rektör a.
Rektör Yardımcısı

Hacettepe Üniversitesi Rektörlük 06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 3001 - 3002 • Faks: 0 (312) 311 9992
E-posta: yazimd@hacettepe.edu.tr • www.hacettepe.edu.tr

Ayrıntılı Bilgi için:
Yazı İşleri Müdürlüğü
0 (312) 305 1008

EK-U: Milli Eğitim Bakanlığı İzin Yazısı

T.C.
MİLLÎ EĞİTİM BAKANLIĞI
Ölçme, Değerlendirme ve Sınav Hizmetleri
Genel Müdürlüğü

Sayı : 57750415-480.99-E.16686081 13.10.2017
Konu : Veri Talebi

HACETTEPE ÜNİVERSİTESİ REKTÖRLÜĞÜNE
(Eğitim Bilimleri Enstitüsü Müdürlüğü)

İlgi : Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Müdürlüğünün 22.12.2016 tarihli ve 51944218/2944 sayılı yazınız.

Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) Araştırmasında kullanılan maddelerin ve puanlama anahtarlarının bir kısmı daha sonraki uygulamalarda da kullanılacağı için Genel Müdürlüğümüzün dışına çıkarılması uygun görülmemektedir. Bu nedenle çalışmanızı, kurumumuza gelerek yürütmeniz uygun görülmüştür.

Bilgilerinize rica ederim.

Mehmet Emin GÜNAYDIN
Bakan a.
Genel Müdür V.

Güvenli Elektronik İmzalı
Aslı İle Aynıdır
17 Ekim 2017

Konya Yolu Üzeri Gazi Hastanesi Karşısı 06500
Teknikokullar / ANKARA
Elektronik Ağ: www.meb.gov.tr
e-posta: sefikgol@meb.gov.tr

Ayrıntılı bilgi için: Ş. GÖL Bil. İşlet.
Tel: (0 312) 413 32 18

EK-Ü: Etik Beyanı

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı bütün bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin bütününü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

27 / 02 / 2019

İbrahim UYSAL

EK-V: Doktora Tez Çalışması Orijinallik Raporu

27 / 02 / 2019

HACETTEPE ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü
Eğitim Bilimleri Ana Bilim Dalı Başkanlığına,

Tez Başlığı: Açık Uçlu Maddelerde Otomatik Puanlamanın Güvenirliği ve Test Eşitleme Hatalarına Etkisi

Yukarıda başlığı verilen tez çalışmamın tamamı (kapak sayfası, özetler, ana bölümler, kaynakça) aşağıdaki filtreler kullanılarak **Turnitin** adlı intihal programı aracılığı ile kontrol edilmiştir. Kontrol sonucunda aşağıdaki veriler elde edilmiştir:

Rapor Tarihi	Sayfa Sayısı	Karakter Sayısı	Savunma Tarihi	Benzerlik Oranı	Gönderim Numarası
22 / 02 / 2019	103	186680	07 / 02 / 2019	%1	1081940869

Uygulanan filtreler:
1. Kaynaklar hariç
2. Alıntılar dâhil
3. 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan eder, gereğini saygılarımla arz ederim.

Ad Soyadı: İbrahim UYSAL
Öğrenci No.: N13248901
Ana Bilim Dalı: Eğitim Bilimleri
Programı: Eğitimde Ölçme ve Değerlendirme
Statüsü: ☐ Y.Lisans ☒ Doktora ☐ Bütünleşik Dr.

İmza

DANIŞMAN ONAYI

UYGUNDUR.
(Prof. Dr., Nuri DOĞAN, İmza)

EK-Y: Dissertation Originality Report

27 / 02 / 2019

HACETTEPE UNIVERSITY
Graduate School Of Educational Sciences
To The Department Of Educational Sciences

Thesis Title: The Reliability of Automated Essay Scoring and Its Effect on Test Equating Errors

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
22 / 02 / 2019	103	186680	07 / 02 / 2019	1%	1081940869

Filtering options applied:
1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Name Lastname: İbrahim UYSAL
Student No.: N13248901
Department: Educational Sciences
Program: Educational Measurement and Evaluation
Status: ☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

Signature

ADVISOR APPROVAL

APPROVED
(Prof. Dr., Nuri DOĞAN, Signature)

EK-Z: Yayımlama ve Fikri Mülkiyet Hakları Beyanı

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım haklan bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan "**Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına ilişkin Yönerge**" kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü/Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü/Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

27 / 02 / 2019

İbrahim UYSAL

"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü Üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç; imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanın önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü Üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*