HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı

Eğitimde Ölçme ve Değerlendirme Programı

İNGİLİZCE KELİME BİLGİSİ TESTİNİN (VST) BİREYSELLEŞTİRİLMİŞ BİLGİSAYARLI TEST OLARAK UYGULANABİLİRLİĞİNİN İNCELENMESİ

Mustafa GÖKCAN

Doktora Tezi

Ankara, 2023

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

İNGİLİZCE KELİME BİLGİSİ TESTİNİN (VST) BİREYSELLEŞTİRİLMİŞ BİLGİSAYARLI TEST OLARAK UYGULANABİLİRLİĞİNİN İNCELENMESİ

INVESTIGATION OF THE APPLICABILITY OF THE VOCABULARY SIZE TEST (VST) AS A COMPUTERIZED ADAPTIVE TESTING

Mustafa GÖKCAN

Doktora Tezi

Ankara, 2023

Kabul ve Onay

Eğitim Bilimleri Enstitüsü Müdürlüğüne,

Mustafa GÖKCAN'nın hazırladığı "İngilizce Kelime Bilgisi Testinin (VST) Bireyselleştirilmiş Bilgisayarlı Test Olarak Uygulanabilirliğinin İncelenmesi" 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. Nuri DOĞAN	İmza
Jüri Üyesi (Danışman)	Doç. Dr. Derya ÇOBANOĞLU AKTAN	İmza
Jüri Üyesi	Prof. Dr. Burcu ATAR	İmza
Jüri Üyesi	Prof. Dr. Dilara BAKAN KALAYCIOĞLU	İmza
Jüri Üyesi	Doç. Dr. Seher YALÇIN	İmza

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

Prof. Dr. İsmail Hakkı MİRİCİ

Eğitim Bilimleri Enstitüsü Müdürü

Öz

Bireyselleştirilmiş Bilgisayarlı Testler (BBT) ölçülmek istenen yapının daha az madde ile ölçme kesinliğinden ödün vermeden ölçülebilmesine olanak sağlar. İngilizce kelime bilgisi gibi ölçülebilmesi için çok sayıda maddeye ihtiyaç duyulan yapılar için BBT bu noktadan önem teşkil eder. Bu çalışmada, İngilizce kelime bilgisini daha kısa sürede ve daha az madde ile ölçebilmek amacıyla, alan yazında İngilizce kelime bilgisini ölçmek amacıyla en çok kullanılan testlerden biri olan Vocabulary Size Test'in (VST) BBT versiyonu (BBT-VST) geliştirilmiştir. VST 140 maddeden oluşan çoktan seçmeli bir testtir. Öncelikle VST'nin kâğıt-kalem formu ile 1622 lisans ve lisansüstü öğrenciden toplanan veriler ile VST'nin geçerlik kanıtları araştırılmıştır. VST'nin İngilizce kelime bilgisini ölçmek için geçerli bir ölçme aracı olduğunun belirlenmesinden sonra aynı verilerle post-hoc simülasyonlar yapılmıştır. Çeşitli BBT kurallarının test edildiği simülasyonlardan elde edilen bulgulara göre gerçek zamanlı BBT uygulaması için en uygun yetenek kesirim yöntemi EAP, madde seçim yöntemi MFI, sonlandırma kuralı "SH ≤ .25" ve madde kullanım sıklığı kontrol yöntemi de "randomesque = 5" olarak belirlenmiştir. En uygun BBT kurallarının tespitinden sonra Concerto platformu ile BBT-VST geliştirilmiştir. Altmış lisans ve lisansüstü öğrenciden VST'nin hem kâğıt-kalem hem de BBT versiyonu ile iki versiyondan elde edilen bulguları karşılaştırmak amacıyla veri toplanmıştır. Bulgulara göre kâğıt-kalem versiyonu ile BBT versiyonunda kestirilen yetenek puanları arasında yüksek ilişki vardır (R= .83, %95 GA = .73, .90). BBT uygulamasında ortalama test uzunluğu 12 olmuştur. Bu sonuç da göstermektedir ki BBT-VST %91 oranında daha az madde kullanarak İngilizce kelime bilgisini güvenilir bir şekilde kestirmiştir.

Anahtar sözcükler: bireyselleştirilmiş bilgisayarlı test, İngilizce kelime bilgisi testi, madde tepki kuramı, post-hoc simülasyon, yabancı dilde ölçme ve değerlendirme

Abstract

Computerized Adaptive Testing (CAT) allows the construct to be measured with fewer items without compromising measurement accuracy. For constructs such as English Vocabulary Size, which requires a large number of items to be measured, CAT is important from this point of view. In this study, in order to measure English vocabulary knowledge in a shorter period of time and with fewer items, the CAT version (CAT-VST) of the Vocabulary Size Test (VST), one of the most widely used tests in the literature to measure English vocabulary knowledge, was developed. The VST is a multiple-choice test consisting of 140 items. First, the validity evidence of the VST was investigated with the data collected from 1622 undergraduate and graduate students with the paper-and-pencil form of the VST. After determining that the VST is a valid assessment tool for measuring English vocabulary size, post-hoc simulations were conducted with the same data. According to the findings obtained from the simulations in which various CAT rules were tested, the most appropriate ability estimation method for real-time BBT application was determined as EAP, item selection method as MFI, termination rule as "SE ≤ .25" and item use frequency control method as "randomesque = 5". After the determination of the most appropriate CAT rules, the CAT-VST was developed with the Concerto platform. Data were collected from sixty undergraduate and graduate students using both the paper-and-pencil and the CAT versions of the VST in order to compare the findings obtained from the two versions. According to the findings, there was a high correlation between the scores obtained from the paper-and-pencil test and the ability scores estimated in the CAT version (R= .83, 95% CI = .73, .90). The average test length was 12 in the CAT version. This result shows that the BBT-VST reliably predicted English vocabulary size using 91% fewer items.

Keywords: computerized adaptive testing, English vocabulary size test, item response theory, language testing, post-hoc simulation

Teşekkür

Doktora eğitimim boyunca ve bu tezin yazılması sürecinde her daim bilgi ve tecrübesiyle destek olan, çalışkanlığı ile örnek olup beni gayrete getiren danışmanım Doç. Dr. Derya ÇOBANOĞLU AKTAN'a,

Tez izleme komitemde yer alan, fikir ve önerileriyle bu tezin yazılmasında önemli katkıları olan değerli Hocalarım Prof. Dr. Burcu ATAR ve Prof. Dr. Dilara BAKAN KALAYCIOĞLU'na,

Savunma jürimde yaptıkları eleştiri ve tavsiyeler ile tezimin iyileşmesine vesile olan kıymetli Hocalarım Prof. Dr. Nuri DOĞAN ve Doç. Dr. Seher YALÇIN'a,

Akademik hayatta şu anda bulunduğum yerin yegâne müsebbipleri, aldığım dersler sayesinde her birinden ayrı ayrı istifade ettiğim Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalındaki saygıdeğer Hocalarıma,

Tezin bitimine yakın, özellikle son birkaç ay kendilerine çok vakit ayıramadığım, ama bunu anlayışla karşılayıp sabreden, maddi manevi destek olan biricik eşim Merve'ye ve minik kızım Meryem'e,

Her daim sevgileriyle yanımda olan Anneme ve Babama, kardeşlerim Yasin ve Yasemin'e

… çok ama çok teşekkür ederim.

İçindekiler

Tablolar Dizini

Şekiller Dizini

Simgeler ve Kısaltmalar Dizini

1PLM: Bir Parametreli Lojistik Model

2PLM: İki Parametreli Lojistik Model

3PLM: Üç Parametreli Lojistik Model

AFA: Açımlayıcı Faktör Analizi

AWS: Amazon Web Services

BBT-VST: VST'nin Bireyselleştirilmiş Bilgisayarlı Test Versiyonu

BNC: British National Corpus

DMF: Değişen Madde Fonksiyonu

BBT: Bireyselleştirilmiş Bilgisayarlı Test

KTK: Klasik Test Kuramı

MC: Monte Carlo

MKK: Madde Karakteristik Eğrisi

MO: Maksimum Olabilirlik

MTK: Madde Tepki Kuramı

VST: Vocabulary Size Test

Bölüm 1

Giriş

Problem Durumu

Bilgisayar teknolojisinin eğitime dâhil olmasıyla, öğretim ve öğrenmede köklü değişimler meydana gelmiştir. Öğrencilerin motivasyonunu arttırması, öğretmenlerin sınavları uygulaması ve sonuçları değerlendirmesi açısından daha etkili yeni yollar sunması gibi noktalardan, bilgisayar teknolojisinin ve bu teknoloji yardımıyla geliştirilen bilgisayar temelli testlerin eğitim ve öğretime büyük katkısı olmuştur (Tseng, 2016). Örneğin bilgisayar temelli testler sayesinde bir test maddesine şekil, grafik, ses ya da video gibi multimedya araçlarını eklemek çok kolaylaşmıştır (Huang ve diğerleri, 2009). Böylece yeni madde formatları üretmek ve daha gerçekçi test ortamları oluşturmak mümkün hale gelmiştir. Ayrıca ihtiyaç anında, istenilen yer ve zamanda ölçme ve değerlendirme çalışmalarını yapmak yine bilgisayar temelli testler sayesinde kolaylaşmıştır (van der Linden & Glas, 2010). Tüm bu avantajlarının ve getirdiği kolaylıkların yanında bilgisayar teknolojisinin, öğrencilerin akademik performansını değerlendirmede en önemli katkılarından biri de bilgisayar temelli testler ve bireye uyarlanmış testlerin entegrasyonunu temsil eden, bireyselleştirilmiş bilgisayarlı testlere (BBT) imkân tanımasıdır (Chang, 2015; Tseng, 2016). BBT uygulamalarında önceden kalibre edilmiş maddelerden oluşan bir havuzdan testi alan bireyin yeteneğine uygun maddeler seçilip uygulanır. Böylece ölçme kesinliğinden taviz vermeden lineer testlere göre daha az sayıda madde ile bireyin yeteneği kestirilir (Chang, 2015; Magis ve diğerleri, 2017; Wainer, 1993; Wainer ve diğerleri, 2000).

Eğitimde ölçme ve değerlendirme alanında çalışan araştırmacılar BBT ile 1970'lerin başlarından itibaren ilgilenmektedirler. İlk BBT konferansı ABD'de 1975 yılında düzenlenmiştir ve o zamandan beri BBT uygulamalarının geliştirilmesi üzerine psikometrik ve teknolojik konular üzerine geniş çapta araştırmalar yapılmıştır (Chalhoub-Deville & Deville, 1999). BBT uygulamaları, dünya çapında genellikle yüksek riskli merkezi sınavlarda

ve lisans verme amaçlı testlerinin uygulanmasında giderek daha fazla kullanılmaktayken, sınıf içi ölçmede geleneksel lineer testler, ya da bir diğer adıyla kâğıt-kalem (K&K) testleri en yaygın ölçme ve değerlendirme yöntemi olarak yerini korumaktadır. Fakat BBT uygulamalarının sınıf içi ölçmelere de entegre edilmesi önemli bir ihtiyaçtır. Çünkü ölçme ve değerlendirme çalışmalarının her bir öğrenciye göre uyarlanması öğretmenlerin öğretim planlamasında daha isabetli kararlar almalarına olanak sağlar. BBT sayesinde her öğrencinin öğrenme süreçlerine ilişkin daha kesin ve güvenilir tespitler sağlayarak öğretmenlerin, öğrencilerin eksik oldukları alanları daha iyi belirlemelerine imkân tanır (Chang, 2015). Özellikle sınıf içi değerlendirme gibi küçük ölçekli durumlarda, düşük motivasyona sahip öğrenciler sınavların geçerliği için önemli bir sorun teşkil etmektedir. BBT uygulamaları ise geleneksel kâğıt-kalem testlerine kıyasla öğrencilerin motivasyonunu arttırmakta ve öğrencileri daha da gayretli hale getirmektedir ve böylece düşük motivasyonlu öğrenciler tarafından test geçerliğine yönelik tehditleri elimine etmekte faydalı olmaktadır (Wise, 2014). Hatta öğrencilerin yaşları arttıkça akademik olarak yorgunluğun ve bıkkınlığın neden olduğu düşük motivasyon ve derse daha az katılım gibi problemlerin olduğu durumlarda BBT uygulamalarının önemli oranda durumu pozitif yönde tersine çevirdiği bulunmuştur (Martin & Lazendic, 2018).

BBT uygulamalarının eğitimde en yaygın kullanıldığı alanlardan biri de İngilizce dil becerilerinin ölçülmesi (English Language Testing) olmasına rağmen bu alanda BBT uygulamaları üzerine çalışılmaya oldukça geç başlanmıştır. Bunun belki de en önemli sebeplerinden biri İngilizce eğitimi alanında uzunca bir süre performans temelli ölçme ve değerlendirme çalışmaları ön planda iken, BBT üzerine odaklanan araştırmacılar genellikle çoktan seçmeli testler gibi cevabın bir grup seçenek arasından seçilmesini gerektiren soru türleri ile ilgilenmeleridir (Chalhoub-Deville & Deville, 1999). Gecikmeye rağmen dil ediniminin ölçümünde BBT uygulamaları son yıllarda oldukça yaygınlaşmış ve Graduate Record Examinations (GRE) ve Graduate Management Admission Test (GMAT) gibi sınavların yanında İngilizce yeterliğin ölçüldüğü Test of English as a Foreign Language

(TOEFL)'da da BBT kullanılmaya başlanmıştır (Economides & Roupas, 2007). BBT uygulamalarının İngilizce başarısını ölçmede işe koşulması aslında çok büyük bir ihtiyaca cevap vermektedir. Malumdur ki İngilizce ya da herhangi bir dildeki yeterliği, güvenilir ve geçerli bir şekilde ölçmek için çok sayıda maddeye ihtiyaç vardır (Kaya, 2022). Tek bir oturumda öğrencileri usandırmadan bu ölçme ve değerlendirme işini yapabilmek pek mümkün değildir. Az sayıda madde ile güvenilir test sonuçları üreten BBT uygulamaları İngilizce başarısın ölçülebilmesi için çok kullanışlı bir çözüm yolu olmaktadır (Khoshsima & Toroujeni, 2017).

Dil yeterliğinin önemli unsurlarından biri de kelime bilgisidir. Bu önem Alderson (2005, s. 88) tarafından "dil yeteneği, büyük oranda kelime bilgisinin bir fonksiyonudur" şeklinde ifade edilir. Bu öneme rağmen kelime bilgisi, dil öğreniminde ihmal edilen bir boyut olmuş (Meara, 1980) ve uygulamalı dilbilim ya da İngilizce eğitimi alanlarında ancak son zamanlarda fazla sayıda araştırılır olmuştur (Nation, 2013). Kelime bilgisi bir yabancı dilin öğrenilmesinin neredeyse her boyutunda ayrı bir öneme sahiptir (Daller ve diğerleri, 2007; Milton, 2009). İngilizce kelime bilgisi ile dinleme (Li, 2019; Noreillie ve diğerleri, 2018), okuma (Zhang & Zhang, 2020), konuşma ve yazma becerileri (Milton, 2013; Miralpeix & Muñoz, 2018) arasında anlamlı pozitif korelasyonlar bulunmuş ve özellikle okuduğunu anlama ile ilgili olarak kelime bilgisi düzeyinin en önemli yordayıcı olduğu vurgulanmıştır (Stæhr, 2008). Kelime bilgisi üzerine araştırmalar son zamanlarda oldukça artsa da yeni kelime bilgisi testlerine pek rastlanmamaktadır (Mizumoto ve diğerleri, 2019). Mevcut kelime bilgisi testlerinin de geçerlik ve kullanışlılığı üzerine çok az çalışma yapılmıştır (Kremmel & Pellicer-Sánchez, 2021; Schmitt ve diğerleri, 2020). Ayrıca İngilizce kelime bilgisinin ölçülmesinde en önemli testlerden biri olan Kelime Bilgisi Testinin (Vocabulary Size Test – VST) (Nation & Beglar, 2007) şu ana kadar BBT versiyonu geliştirilmemiştir. Bu sebeple bu çalışmada VST'nin ilk BBT versiyonu geliştirilerek testin geçerlik ve kullanışlılığı araştırılmıştır.

Araştırmanın Amacı ve Önemi

İngilizcenin ikinci dil ya da yabancı dil olarak ediniminde İngilizce kelime bilgisinin önemli bir rol oynadığından bahsedilmiştir. Fakat İngilizce kelime bilgisini ölçebilmek için çok sayıda maddeye ihtiyaç duyulmaktadır. Bu sebeple İngilizce başarısında önemli bir yordayıcı olan İngilizce kelime bilgisinin, ölçme kesinliğinden taviz vermeden daha az madde ile daha kısa sürede ölçülebilmesi önem teşkil etmektedir. Bu doğrultuda bu araştırmada İngilizce kelime bilgisinin bireyselleştirilmiş bilgisayarlı test ile ölçülmesi ve elde edilen sonuçların klasik kâğıt-kalem testiyle elde edilen sonuçlarla karşılaştırılması amaçlanmıştır. Bu hedefe ulaşmak için Nation ve Beglar (2007) tarafından geliştirilen ve alan yazında kelime bilgisinin ölçülmesinde en çok atıf alan testlerden biri olan Kelime Bilgisi Testinin (Vocabulary Size Test – VST) BBT versiyonu oluşturulmuştur. Alan yazında VST'nin BBT uygulamasına rastlanmamıştır. Bu çalışmada bu eksiklik giderilmiş, hem İngilizce öğretimi alanına hem de eğitimde ölçme ve değerlendirme alanına VST gibi önemli bir ölçme aracının bireyselleştirilmiş bilgisayarlı versiyonu kazandırılmıştır.

Bireyselleştirilmiş bilgisayarlı testlerin işlevsel olabilmesi için madde havuzu, güçlük parametresinin olası her değerini kapsayacak şekilde geniş olmalıdır (Chang, 2015). Bu sebeple İngilizce kelime bilgisini ölçmek amacıyla bu çalışmada VST özellikle tercih edilmiştir. Çünkü VST İngilizce metinlerde en sık kullanılan 14000 kelime arasından seçilen kelimelerden oluşturulmuştur. Alan yazında İngilizce kelime bilgisinin BBT ile ölçüldüğü az sayıda çalışmada ise madde havuzları, en sık kullanılan 3000 (Kezer & Koç, 2014) ya da 6480 (Tseng, 2016) kelime gibi daha az sayıda kelime içinden seçilen kelimelerle oluşturulmuştur. Daha zor kelimeleri, yani güçlük parametresi olarak daha yüksek değerlere sahip maddeleri de içinde bulunduran bir testin BBT uygulamasının gerçekleştirilmesi, bu çalışmayı kelime bilgisini BBT ile ölçmeye çalışan diğer çalışmalardan ayırmaktadır.

Ayrıca, literatürde VST için detaylı bir değişen madde fonksiyonu (DMF) çalışması bulunmamaktadır. DMF, bir maddeye doğru yanıt verme olasılığı belirli bir grup üyeliğinin bir fonksiyonu olarak farklılık gösterdiğinde ortaya çıkar. *Standartlar'a* (AERA, APA &

NCME, 2014) göre, DMF, test adilliği için büyük bir tehdit oluşturmaktadır çünkü yanlı yetenek tahminlerine yol açabilmektedir. Madde yanlılığını tespit etmenin ilk adımı, DMF analizleri yaparak potansiyel olarak yanlı maddeleri tespit etmektir (Çepni & Kelecioğlu, 2021; Uysal ve diğerleri, 2019). Tarihsel olarak, çoğu DMF çalışması cinsiyet veya ırk temelli grup farklılıklarına odaklanmıştır (Kıbrıslıoğlu Uysal & Atalay Kabasakal, 2017; Zumbo, 2007). Bu çalışmada da cinsiyete bağlı DMF araştırılmış ve bu önemli testin potansiyel olarak yanlı maddeleri tespit edilerek ileride yapılacak çalışmalar ve testin kalitesinin artırılması için önerilerde bulunulmuştur.

VST'nin BBT uygulaması geliştirilmeden önce, gerçek zamanlı BBT uygulamasında kullanılacak madde seçim yöntemi, yetenek kestirim yöntemi ve sonlandırma kuralı gibi kuralların belirlenebilmesi için simülasyonlar yapılmıştır. Maliyetinin düşük olması nedeniyle Monte Carlo (MC) simülasyonlar alan yazında çok yaygın olarak kullanılmaktadır. Fakat eğer bir BBT uygulaması geliştirip ileride gerçek bireyler ile çalışılacaksa, gerçek veri kullanarak yapılan post-hoc simülasyonlar MC simülasyonlarına tercih edilmelidir (Thompson & Weiss, 2011). Çünkü MC simülasyonlarından elde edilen bulguları ve bunların yorumlamalarını gerçek test uygulamalarına genellemek her zaman mümkün olmamaktadır (Sari, 2020). Alan yazındaki bu öneriler doğrultusunda bu çalışmada post-hoc simülasyonlar tercih edilmiştir. BBT uygulamalarının gerçek bireyler ve gerçek madde havuzu kullanılarak daha etkili bir şekilde simüle edildiği (Thompson & Weiss, 2011) bu tür simülasyonlara bir örnek oluşturması açısından çalışmanın bu noktadan oldukça önemli olduğu düşünülmektedir.

Ayrıca bu zamana kadar İngilizce kelime bilgisinin VST ile ölçülmesi ve sonuçlarının analiz edilmesinde tek parametreli lojistik model ve Rasch modeli kullanılmıştır. Kullanılan bu yöntemler, VST'nin en çok eleştirilen yönü olan soruları cevaplarken şans ile doğru cevaplama olasılığının yüksek olmasını test edememektedir. Bu çalışmada şans faktörünü de hesaba katan üç parametreli lojistik model ile MTK kestirimleri yapılarak alan yazındaki

bu noktadan eksiklikler de giderilmeye çalışılmıştır. Şu ana kadar bahsedilen çalışmanın amaçları doğrultusunda araştırma problemi ve alt problemler aşağıdaki gibi sunulmuştur.

Araştırma Problemi

İngilizce Kelime Bilgisi Testinin (VST) bireyselleştirilmiş bilgisayarlı test olarak uygulanabilmesi için BBT koşulları nasıl olmalıdır ve BBT ve kâğıt-kalem testi uygulamalarının madde ve birey istatistikleri nasıldır?

Alt Problemler

1. VST'nin BBT uygulamasından önce yapılan post-hoc simülasyonlara göre en uygun BBT koşulları nelerdir?

    a. En uygun yetenek kestirim yöntemi hangisidir?

    b. En uygun madde seçim yöntemi hangisidir?

    c. En uygun sonlandırma kuralı hangisidir?

    d. En uygun madde kullanım sıklığı kontrol yöntemi hangisidir?

    e. Madde kullanım sıklığı kontrol yöntemi kullanıldığında ölçme kesinliğinde önemli bir düşüş gerçekleşmekte midir?

2. VST'nin kâğıt-kalem ve BBT versiyonları beraber uygulandığında madde ve birey istatistikleri nasıldır?

    a. BBT ve kâğıt-kalem versiyonları tarafından kestirilen yetenek parametreleri arasındaki ilişki ne düzeydedir?

    b. BBT uygulamasından elde edilen sonuçlara göre madde kullanım sıklığı istatistikleri nasıldır?

    c. BBT uygulamasında ortalama test uzunluğu ile birey yetenek parametresi arasındaki ilişkinin yönü ve düzeyi nedir?

Sınırlılıklar

Bu çalışma post-hoc simülasyonlara dahil edilen yetenek kestirim yöntemleri (ML, EAP, BM ve WL), madde seçim yöntemleri (MFI ve bOpt), test sonlandırma kuralları (standart hata; sh = 0.30, sh = 0.25 ve sh = 0.20) ve madde kullanım sıklığı kontrol yöntemi (*randomesque* = 3 ve *randomesque* = 5) ile sınırlıdır. Ayrıca bu çalışmada post-hoc simülasyonların gerçekleştirildiği catR paketi madde kullanım sıklığı kontrol yöntemi olarak sadece *randomesque* yöntemi vardır. Bu sebeple bu çalışmada madde kullanım sıklığını kontrol etmede sadece *randomesque* yönteminin etkisi test edilmiştir.

Bölüm 2

Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Bu bölümde öncelikle bireyselleştirilmiş bilgisayarlı test uygulamalarını mümkün kılan ve BBT'nin arkasındaki psikometriyi oluşturan madde tepki kuramı (MTK) anlatılmıştır. İkinci olarak BBT uygulamaları tanıtılacak, işleyişinden ve avantajlarından bahsedilmiştir. Son olarak da VST üzerine yapılmış çalışmaların yanında İngilizce kelime bilgisinin BBT uygulamaları ile ölçüldüğü ilgili araştırmalar sunulmuştur.

Madde Tepki Kuramı

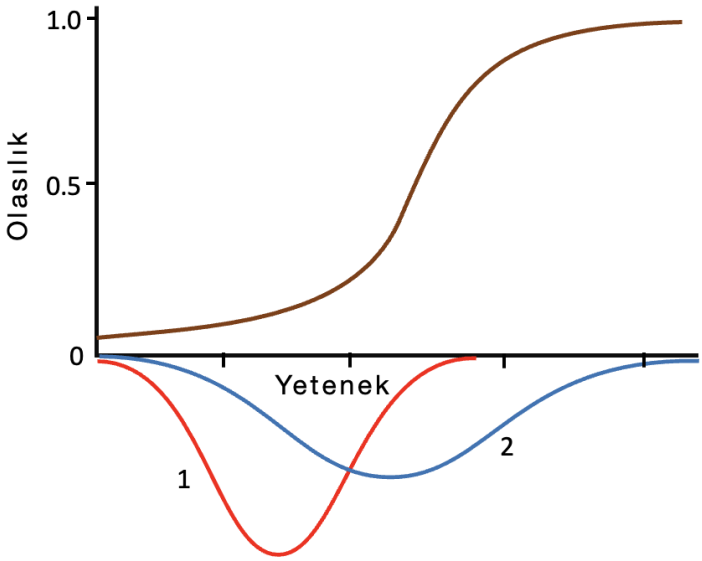
Klasik Test Kuramı (KTK) ve Modern Test Kuramı, Test Kuramının önemli iki ayrı dalıdır. İlkinde analiz birimi genellikle bireyin bir grup maddeye verdiği cevaplardan elde edilen toplam puandır (Paek & Cole, 2020). Bu kuramın sınırlılıkları, ölçme uzmanlarını alternatif ölçme modellerini aramaya sevk etmiştir. KTK'nın sınırlılıklarına örnek olarak maddeye bağımlı yetenek kestirimleri, örneklem bağımlı madde istatistikleri, belirli bir yetenek düzeyindeki bireylerin yine belli bir kısım test maddeleri karşısında nasıl bir performans göstereceklerine dair bir olasılık bilgisinin olmaması ve ölçme hatasının her bir birey için eşit olarak sınırlandırılması verilebilir (Hambleton ve diğerleri, 1991).

Modern Test Kuramlarından biri olan Madde Tepki Kuramı (MTK) ise KTK'nın bu sınırlılıklarını büyük oranda gidermiştir (Paek & Cole, 2020). Basit bir şekilde tarif edilecek olursa madde tepki kuramı modelleri, bir ölçme aracı tarafından ölçülen yetenek ya da gizil yapının (θ sembolüyle gösterilir) ölçeğe ait bir madde ile olan ilişkisini gösterir (DeMars, 2010). Yani isminden de anlaşılacağı üzere MTK bireyin test davranışını modeller fakat bunu KTK gibi test puanları düzeyinde değil de madde düzeyinde yapar. MTK eğitimde ve psikolojide oldukça yaygın olarak kullanılmaktadır ve kullanım alanı diğer sosyal bilimlere, tıp araştırmalarına ve ekonomi alanlarına doğru giderek genişlemektedir (Paek & Cole, 2020).

MTK'nın tarihçesine bakıldığında ele alınan ilk madde formatının, cevapların ya doğru ya da yanlış olarak puanlandığı ikili puanlanan format ya da iki kategorili madde formatı olduğu görülmektedir (van der Linden & Hambleton, 1997). Bu formatta maddeye verilecek doğru cevap olasılığı, testin ölçtüğü gizil yapı noktasından bireylerin sahip oldukları yeteneğin ve madde parametrelerinin bir fonksiyonu olarak modellenmesi amaçlanır (Paek & Cole, 2020). Sınavı alan bireyin madde performansı ile bu performans altında yatan gizil yapılar arasındaki ilişki madde karakteristik fonksiyonu ya da madde karakteristik eğrisi adı verilen monoton bir şekilde artan bir fonksiyon ile tanımlanabilir. Bu fonksiyona göre yetenek düzeyi arttıkça bir maddeye doğru cevap verme olasılığı artmaktadır (de Ayala, 2009; Hambleton ve diğerleri, 1991).

Şekil 1

Madde Karakteristik Eğrisi



Şekil 1'de, bir madde karakteristik eğrisi görülmektedir. X ekseninde yetenek, y ekseninde ise maddeyi doğru cevaplama olasılığı yer almaktadır. Bu gösterimde performansın altında yatan tek bir gizil yapı ve yetenek dağılımları farklı olan iki grup vardır. Görüldüğü üzere yetenek bakımından yüksek değerlere sahip bireyler ait oldukları gruplardan bağımsız olarak yeteneği düşük bireylere göre soruyu doğru cevaplama olasılıkları daha yüksektir. Yani, aynı yetenek düzeyine sahip bireyler farklı gruplarda da

olsalar, bir maddeye doğru cevap verme olasılıkları aynıdır (Embretson ve Reise, 2000; (Hambleton & Swaminathan, 1985).

KTK'nın sınırlılıklarının ölçme uzmanlarını yeni arayışlara yönlendirdiğinden bahsedilmiştir. Bu çabaların en önemlilerinden biri Lord'un 1952 yılında yayınladığı *A Theory of Test Scores* adlı kitabıdır. Bir diğeri de Birnbaum'un Lord'un bu kitabını takip eden çalışmalarıdır. Lord (1952) bireyin bir maddeyi doğru cevaplama olasılığını modelleyebilmek için eşitlik 1'de verilen normal ogive fonksiyonunu kullanmıştır.

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/z} dz. \quad \text{… eşitlik 1}$$

Birnbaum (1968) ise eşitlik 1'de verilen bu normal ogive fonksiyonunu, eşitlik 2'de verilen lojistik model ile değiştirmiştir. Aslında Birnbaum yeni bir test kuramı oluşturma motivasyonu ile ortaya çıkmamıştır. Bir istatistikçi olarak Birnbaum'un amacı Lord (1952) tarafından normal ogive model üzerine başlanan çalışmayı istatiksel olarak daha uygulanabilir yapmaktır (van der Linden & Hambleton, 1997).

$$P_i(\theta) = \frac{1}{1 + \exp\{-a_i(\theta - b_i)\}}. \quad \text{… eşitlik 2}$$

1940 ve 1950'lerde MTK uygulamalarında normal ogive tepki fonksiyonları hâkimdir. Fakat bahsedilen istatiksel ve pratiğe yönelik nedenlerden dolayı, 1950'lerin sonunda lojistik tepki fonksiyonları bu fonksiyonun yerini almıştır. Artık günümüzde, eğitimde ve psikolojide ölçme ve değerlendirme alanlarında lojistik tepki fonksiyonlarını kullanan MTK modelleri daha yaygındır (van der Linden ve Hambleton, 1997).

İkili olarak puanlanan maddeler için günümüzde en yaygın kullanılan MTK modelleri, madde özelliklerini modelleyen ve lojistik fonksiyonu oluşturan parametre sayısına göre adlandırılmaktadır. Eğer sadece bir tane madde özelliği (madde güçlüğü) modelleniyorsa MTK modeli bir parametreli lojistik model (1PLM) olarak adlandırılır. İki tane madde özelliği modelleniyorsa (madde güçlüğü ve madde ayırt ediciliği) iki parametreli lojistik model (2PLM), üç madde özelliği modelleniyorsa (madde güçlüğü, madde ayırt ediciliği ve şans

parametresi) üç parametreli lojistik model (3PLM) olarak adlandırılır. Bu üç lojistik model hakkında aşağıda kısaca bilgi verilmiştir.

Bir Parametreli Lojistik Model

Bir parametreli lojistik model en yaygın olarak kullanılan MTK modellerinden biridir. Madde karakteristik fonksiyonu (bir parametreli lojistik model için) eşitlik 3'teki gibidir.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad \text{…eşitlik 3}$$

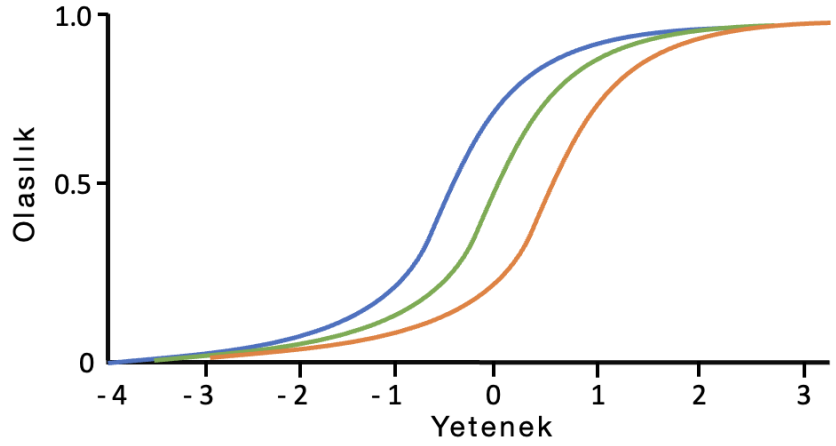
Herhangi bir madde için b parametresi, doğru cevaplama olasılığının 0.5 olduğu yetenek ölçeği üzerindeki noktadır. Bu parametre ayrıca yer (location) parametresi olarak da adlandırılır ve yetenek ölçeğine göre madde karakteristik eğrisinin pozisyonunu gösterir. b parametresinin değeri arttıkça, %50 olasılıkla maddeyi doğru cevaplayabilmek için bireyin sahip olması gereken yetenek düzeyi de artar ve sonuçta madde daha da zorlaşır. Zor maddeler yetenek ölçeğinin nispeten sağ tarafında, kolay maddeler de sol tarafında yer alır. Bir gruba ait yetenek değerleri, ortalaması 0, standart sapması 1 olacak şekilde dönüştürüldüğünde, b parametresine ait değerler genellikle -2 +2 aralığında yer alır. -2 değerine yakın b parametresine sahip maddeler çok kolay, +2 değerine yakın b parametresine sahip maddeler de zor olarak değerlendirilir (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1985). Bir parametreli modelde, birey performansını etkileyen tek madde özelliğinin madde güçlüğü olduğu kabul edilir. Klasik test kuramındaki madde ayırt edicilik indeksine denk gelen bir madde parametresi yoktur, ki bu da her maddenin eşit düzeyde ayırt ediciliğe sahip olduğunun kabul edildiğini gösterir (de Ayala, 2009; Hambleton ve diğerleri, 1991).

Şekil 2'deki madde karakteristik eğrilerine bakıldığında, düşük asimptotun 0 olduğu görülmektedir. Buna göre çok düşük yetenek düzeyindeki bireylerin soruyu doğru cevaplama olasılıkları sıfırdır. Dolayısıyla çoktan seçmeli maddelerde muhtemel bir durum olan düşük yetenek düzeyindeki bireylerin tahmin etme olasılığına bu modelde yer

verilmemektedir. Görüldüğü üzere bir parametreli modelin bazı önemli derecede kısıtlayıcı varsayımları vardır.

Şekil 2

1 PLM'ye Ait Madde Karakteristik Eğrileri



İki Parametreli Lojistik Model

Kümülatif normal dağılım (normal ogive) temelli iki parametreli madde tepki modelini ilk olarak Lord'un (1952) geliştirdiğinden bahsedilmişti. Birnbaum (1968) ise bu modelin yerine iki parametreli lojistik fonksiyonu ikame etmiştir. Normal ogive fonksiyonlarına göre, lojistik fonksiyonlar ile çalışmak daha kolaydır. Çünkü ilki integral almayı gerektirmektedir (De Mars, 2010).

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad \text{…eşitlik 4}$$

Birnbaum tarafından iki parametreli lojistik model için geliştirilen madde karakteristik eğrisi ya da fonksiyonu eşitlik 4'te verilmiştir. P(θ) ve b parametreleri ilk eşitlik 3'te tanımlandığı gibidir. İki parametreli modelin iki tane farklı öğesi olduğu eşitliğe bakıldığında görülmektedir. Bunlardan ilki olan D faktörü, ölçekleme faktörüdür ve lojistik fonksiyon ile normal ogive fonksiyonu birbirine mümkün olduğunca yaklaştırmak için kullanılır (de Ayala, 2009; Hambleton ve diğerleri, 1991). Eğer bu değer 1.7 sabitine eşitlenirse lojistik fonksiyon,

normal ogive fonksiyon ile neredeyse aynı metriğe yerleşir. θ'nın her değeri için mutlak değer olarak 0.01'den daha az fark elde edilmesini sağlar (Camilli, 1994).

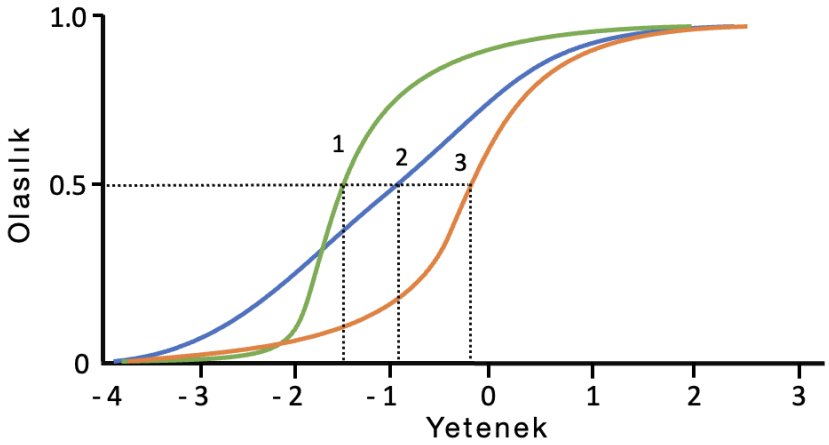
Yeni eklenen öğelerden ikincisi, madde ayırt edicilik parametresi olarak adlandırılan a parametresidir. Bu parametre, madde karakteristik eğrisinin, yetenek ölçeği üzerinde yer alan b parametresinin bulunduğu noktadaki eğimini verir. Daha dik eğime sahip maddeler bireyleri farklı yetenek düzeylerine ayırmada daha düşük eğime sahip maddelere göre daha başarılıdır (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Madde karakteristik eğrisinin eğimi, bireyin yetenek düzeyinin bir fonksiyonu olarak değişmektedir ve yetenek düzeyinin madde güçlüğüne eşit olduğu noktada bu eğim en yüksek değeri almaktadır. Madde ayırt ediciliği madde karakteristik eğrisinin genel eğimini temsil etmemektedir (Baker, 2001). Teorik olarak madde ayırt edicilik parametresi (-∞, +∞) ölçeğinde yer alır. Fakat negatif ayırt edicilik değerine sahip maddeler, madde ile alakalı bir soruna işaret ettiği için ölçekten ya da testten çıkarılır. Çünkü bu tür maddeler söz konusu olduğunda bireyin yeteneği arttıkça maddeye doğru cevap verme olasılığı düşmektedir (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Ayrıca negatif ayırt edicilik, yüksek yetenek düzeyinde bulunan bireyler arasında madde ile alakalı bir kavram yanılgısına ya da yanlış anlamaya işaret edebilir (Baker, 2001). Genellikle 2'den büyük a parametresi elde etmek pek mümkün değildir. Bu sebeple madde ayırt edicilik parametresi için genel aralık (0, 2) aralığıdır. Maddenin en başarılı şekilde ayırt edebildiği bireyler, yetenek düzeyi madde güçlüğüne yakın olan bireylerdir (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985).

Şekilde 3'te görüldüğü üzere madde karakteristik eğrileri, bir parametreli modelde olduğu gibi birbirine paralel değildir. Her MKE farklı bir eğime sahiptir ve bu da ayırt edicilik parametresi değerlerinin farklı olduğunu gösterir. Her eğrinin düşük asimptotunun sıfır olmasına dikkat edilirse yine aynı bir parametreli modelde olduğu gibi, iki parametreli modelin de tahmin parametresine yer vermediği görülmektedir. Tahmini modelden çıkaran bir varsayım en çok serbest yanıtlı maddeler için uygundur. Çoktan seçmeli testler ile makul

bir uyumun sağlanabilmesi için testin çok da zor olmayan bir test olması gerekmektedir. Mesela etkili bir eğitimden sonra yapılacak bir çoktan seçmeli başarı testi, tahminin dâhil edilmediği iki parametreli modele uyum gösterebilir (Hambleton ve diğerleri, 1991).

Şekil 3

2 PLM'ye Ait Madde Karakteristik Eğrileri



Üç Parametreli Lojistik Model

Üç parametreli lojistik modelin matematiksel ifadesi eşitlik 5'te verilmiştir. P (θ), b, a ve D parametreleri iki parametreli modelde tanımlandıkları gibidir. Bu fonksiyonda fazladan sadece c parametresi vardır ve genellikle bu parametre şans parametresi olarak adlandırılır. Bu parametre sıfırdan farklı bir düşük asimptot değeri elde edilmesine olanak sağlar ve düşük yetenek düzeyindeki bireylerin maddeyi doğru cevaplama olasılıklarını gösterir (de Ayala, 2009; Hambleton ve diğerleri, 1991).

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad \text{…eşitlik 5}$$

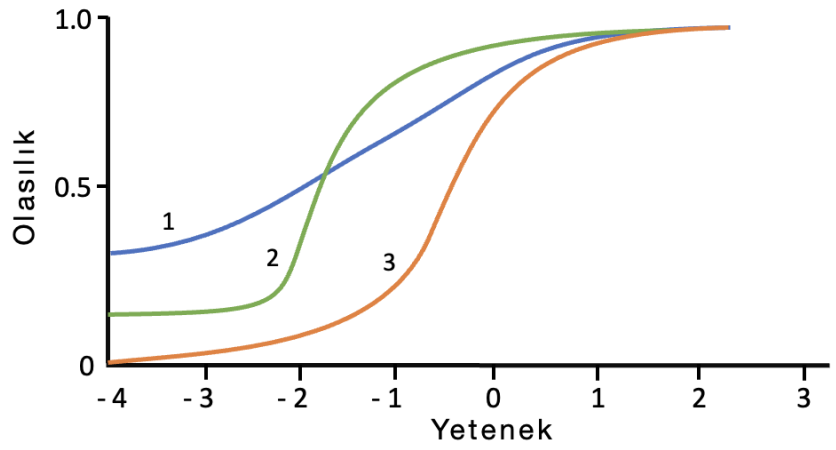
Düşük yetenek düzeyindeki bireylerin çoktan seçmeli testlerde ya da maddelerde sıfırdan farklı olan performanslarını açıklayabilmek için modele üçüncü bir parametrenin eklenmesini öneren Birnbaum'dur. Birnbaum'a göre bu sıfırdan farklı performans doğru cevabı tahmin etmekteki olasılıktan kaynaklanmaktadır. Böylelikle model eşitlik 5'teki formunu almaktadır. Eşitliğe göre sınavı alan birey doğru cevabı ya eşitlikte verilen olasılıkla

biliyordur ya da c değerine eşit olan başarı olasılığı ile yaptığı tahminle biliyordur. c parametresi tepki fonksiyonunun düşük asimptot yüksekliğidir. Her ne kadar eşitlik 5 artık bir lojistik fonksiyonu ifade etmese de, model, üç parametreli lojistik model olarak bilinmektedir (van der Linden & Hambleton, 1997).

Baker'a göre (2001) c parametresi sadece tahmin ederek maddeyi doğru olarak cevaplama olasılığıdır. Tanımı gereği, c parametresi bireyin yetenek düzeyinin bir fonksiyonu olarak değişkenlik göstermez. Bu sebeple yetenek ölçeğinde en sağda ve en solda yer alan bireylerin, yani yeteneği en yüksek ve en düşük bireylerin maddeyi şansla (tahminle) doğru cevaplandırma olasılıkları aynıdır. Lord'a (1974) göre c parametresi tahmin parametresi olarak adlandırılmamalıdır. Çünkü düşük asimptotun sıfırdan farklı olması durumu, madde geliştiricilerinin cazip ama yanlış seçenekleri yazma becerilerinden de kaynaklanabilir.

Şekil 4

3 PLM'ye Ait Madde Karakteristik Eğrileri



Şekil 4'te 3 PLM'ye ait madde karakteristik eğrileri gösterilmiştir. Görüldüğü üzere 3 nolu madde hariç diğerlerinin düşük asimptot değerleri sıfırdan farklıdır. Bu da çok düşük yetenek düzeyindeki bireylerin maddeyi hangi olasılıkla doğru cevaplayacaklarını gösterdiği gibi her bireyin maddeyi şansla doğru cevaplama olasılığını da gösterir. Yani şans

parametresinin değeri doğrudan olasılık şeklinde yorumlanabilir. Örneğin; c = .12 ise, tüm yetenek düzeylerinde maddeyi şansla doğru cevaplama olasılığı %12'dir.

Madde Tepki Kuramının Varsayımları

Her matematiksel model temelli çalışmada olduğu gibi verinin modele iyi uyum sağlayabilmesi ve doğru parametre kestirimleri elde edebilmek için MTK analizlerine başlanmadan önce sağlanması gereken varsayımlar vardır. MTK çalışmalarında test edilen en önemli iki varsayım tek boyutluluk ve yerel bağımsızlık varsayımlarıdır.

Tek Boyutluluk. MTK modellerinin en bilinen varsayımlarından biri tek boyutluluktur. Bu varsayıma göre bir testi oluşturan maddeler tarafından sadece tek bir yeteneğin ölçülmesi gerekmektedir. Fakat test performansını bazı bilişsel, kişisel ve sınav anındaki faktörler az çok etkilediği için, bu varsayımın tamamen karşılanması mümkün değildir. Bu varsayımın elimizdeki veri tarafından yeterince karşılanabilmesi için gerekli olan şey, test performansını etkileyen baskın bir yapı ya da faktörün olmasıdır. Bu baskın yapı ya da faktör de test tarafından ölçülen yetenek olarak ele alınır (de Ayala, 2009; Hambleton ve diğerleri, 1991). Baskın bir yapı ya da faktör de MTK analizlerinde tek boyutluluk varsayımının karşılanması için yeterli görülmüştür (Hambleton & Swaminathan, 1985; Henning ve diğerleri, 1985).

Yerel Bağımsızlık. Yerel bağımsızlık ile kastedilen, test performansını etkileyen yetenekler sabit tutulduğunda, bireylerin herhangi bir madde çiftine verdikleri cevapların istatistiksel olarak birbirinden bağımsız olması gerektiğidir. Yani bireylerin farklı maddelere verdikleri cevaplar arasında ilişki olmaması lazımdır (Hambleton ve diğerleri, 1991). Eğer tek boyutluluk varsayımı sağlanırsa, yerel bağımsızlık da elde edilmiş olur ve bu yönüyle her iki kavram da aynı olarak kabul edilebilir (Lord 1980; Lord & Novick, 1968). Fakat yerel bağımsızlık eldeki veri tek boyutlu olmadığında da elde edilebilir (de Ayala; 2009). Mesela bir matematik testindeki bir madde eğer yüksek düzeyde bir okuma becerisi gerektiriyorsa, düşük okuma becerisine sahip bireyler, yeterince sayısal yetenekleri olsa bile bu soruyu doğru cevaplayamayabilirler. Bu yüzden, sayısal yetenekten başka bir boyut bireyin

performansını etkilemektedir. Eğer tek boyutlu bir MTK modeli veriye uydurulursa yerel bağımsızlık elde edilemez. Diğer bir yandan eğer tüm bireyler yeterli okuma becerilerine sahipse, sadece sayısal yetenek, performansı etkiler ve tek boyutlu bir model kullanıldığında yerel bağımsızlık elde edilmiş olur (Hambleton ve diğerleri, 1991).

Bireyselleştirilmiş Bilgisayarlı Test Uygulamaları

Eğitimde ve psikolojide gerçekleştirilen ölçme uygulamalarının çoğu, maddelerin sıralamasının önceden belirlendiği, madde sayısı sabit ve lineer bir formatta klasik kâğıt-kalem ya da bilgisayar temelli testlerle yürütülür. Bu testler de genellikle klasik test kuramının ilkelerine (örn., Cronbach, 1990; Gulliksen, 1950) göre oluşturulur ve ölçme aracını oluşturan maddeler, maddeler arası iç tutarlılığı (güvenirlik) mümkün olduğu kadar yüksek tutarak seçilmeye çalışılır (Weiss, 2004). Bu maddeler de, ölçülen yapı noktasından genellikle ortalama düzeyde bir öğrenci için idealdir fakat ortalamanın altında olan öğrenciler için çok zor ve ortalamanın üstünde olan öğrenciler için de çok kolay olmaktadır (Wainer ve diğerleri, 2000; Weiss, 2004). Fakat bir bireyin yeteneği hakkında en kesin ölçmenin sağlanabilmesi için testin güçlüğünün, bireyin yetenek düzeyine eşlenmesi gerekmektedir. Bir grup bireye ya da öğrenciye uygulanan tek bir test, her bir birey için aynı ölçme kesinliği sağlamaz. En ideal ölçme, bireyin yetenek düzeyine göre uyarlanmış ya da bireyselleştirilmiş test ile sağlanabilir (Hambleton ve diğerleri, 1991)

Bilindiği üzere bireyselleştirilmiş testlerin ilk uygulaması Binet'in zekâ testidir. Ondan sonra Lord'un 1960'lı yılların sonlarındaki çalışmalarına kadar bireyselleştirilmiş test alanında herhangi bir gelişme görülmemiştir. Lord'un bireyselleştirilmiş testlere verdiği önem onun, klasik sabit uzunluktaki testleri verimsiz olarak görmesinden kaynaklanmaktadır. Lord' göre özellikle düşük ve yüksek yetenek düzeyindeki bireyler için klasik testler istenilen faydayı sağlamamaktadır. Lord, her bir bireye uygulanan test maddelerinin, bireyin yetenek düzeyi hakkında maksimum bilgi sağlayacak şekilde seçilmesi halinde testlerin ölçme kesinliğinde herhangi bir azalmaya uğramadan

kısaltılabileceğini savunmaktadır. Bu da bireyselleştirilmiş test uygulamaları ile mümkün olmaktadır. Bu tür uygulamalarda bir bireye uygulanan maddelerin sıralaması, testte önceki maddelerde bireyin göstermiş olduğu performansa bağlıdır. Bireyin önceki performansına bağlı olarak, bireyin yetenek düzeyi hakkında en çok bilgi verici maddeler seçilir ve uygulanır. Böylece ölçme kesinliğinde hiçbir kayıp yaşanmadan test uzunluğu azalmış olur. Yüksek yetenek düzeyindeki bireylere, nispeten kolay maddeler; düşük yetenek düzeyindeki bireylere ise çok zor maddeler uygulanmamış olur. Çünkü bu tür maddeler bireyin yeteneği hakkında ya çok az ya da hiç bilgi vermemektedir (Hambleton ve diğerleri, 1991).

Bireyselleştirilmiş testlerin, bireyselleştirilmiş bilgisayarlı testler olarak uygulanmaya başlanması ancak bilgisayar teknolojisinin gelişmesiyle olmuştur. Çünkü BBT uygulamalarının psikometrik temellerini oluşturan MTK, kompleks istatistiksel analizler gerektirmektedir (van der Linden & Glas, 2010). BBT uygulamalarında bireyin yeteneği ve bireyin yeteneğine göre uygulanacak maddelerin seçimi MTK hesaplamaları ile belirlenmektedir. Hatta BBT uygulamalarında kullanılacak madde havuzunu oluşturan maddelerin kalibrasyonu da MTK modelleri ile yapılmaktadır (Magis ve diğerleri, 2017). Bu sebeple ancak daha güçlü bilgisayarların üretilmesinden sonra BBT uygulamaları oldukça yaygın olarak kullanılmaya başlanmıştır. Eğitim alanında kullanımının yanında BBT uygulamaları ABD'de sağlık alanında da çok sık kullanılmaktadır. Hatta devlet tarafından fonlanan ve yetişkin ya da çocuk fark etmeksizin ülke çapında milyonlarca hastanın fiziksel, psikolojik ve sosyal sağlık durumlarının takibini yapan bir girişim (PROMIS), çalışmalarının çoğunda BBT uygulamalarından faydalanmaktadır (Scalise & Allen, 2015).

BBT uygulamalarının en önemli avantajının ölçme kesinliğinin kayba uğramadan testin kısalmasına olanak sağlaması olduğu vurgulandı. Bunun yanında BBT uygulamalarının birçok avantajı vardır. Hambleton vd. (1991) şu avantajlardan bahseder;

- artan test güvenliği
- istenildiği vakit test uygulama imkânı

- cevap kâğıdına ihtiyaç bırakmaması
- bireye göre ayarlanmış sınav hızı
- anında puanlama ve raporlama
- bazı öğrenciler için test yılgınlığını (frustration) olabildiğince azaltması
- yüksek düzeyde test standardizasyonu
- farkına varıldığında kötü çalışan maddelerin kolaylıkla madde havuzundan çıkarılması
- madde formatı seçiminde fazlaca esneklik sağlaması
- testin uygulanma süresinde azalma

BBT uygulamaları üzerine yapılan çalışmalar genellikle altı alan üzerine odaklanmıştır. Bunlar; MTK modelinin seçimi, madde havuzu, ölçme işlemi için başlama noktası, sıralı gelecek maddelerin seçimi, yetenek kestirimi ve test uygulamasının ne zaman sonlandırılacağına karar vermek için bir yöntemin seçilmesidir. Hambleton vd. (1991), Reshetar (1990) tarafından BBT'nin işleyişi hakkında verdiği bir örneği kısaca şöyle özetlemişlerdir.

BBT örneği için Reshetar 13 maddeden oluşan bir madde havuzu oluşturmuştur. Bu 13 maddeye ait parametreler Tablo 1'de verilmiştir.

Bir BBT uygulamasında sırasıyla şu işlemler gerçekleşir;

1. Örneğin üçüncü madde ilk madde olarak seçilir. Çünkü bu maddenin güçlüğü ortalama bir değerdir ve yüksek ayırt ediciliğe sahiptir. Sınavı alan bireyin bu maddeyi doğru cevaplandırdığını varsayalım. Bilindiği üzere cevaplanan tüm maddelerin doğru ya da yanlış olduğu durumlarda maksimum olabilirlik (MO) kestirim yöntemi ile yetenek kestirimi yapılamaz. Bu sebeple bu ilk adımda herhangi bir yetenek kestirimi yapılamamıştır. Bireyin en az bir soruyu doğru bir soruyu da yanlış yapması gerekir ki MO yöntemi ile kestirim yapılabilsin.

Tablo 1

BBT Örneğinin Maddeleri ve Parametreleri

Madde	Madde Parametreleri		
	b	a	c
1	0.09	1.11	0.22
2	0.47	1.21	0.24
3	- 0.55	1.78	0.22
4	1.01	1.39	0.08
5	- 1.88	1.22	0.07
6	- 0.82	1.52	0.09
7	1.77	1.49	0.02
8	1.92	0.71	0.19
9	0.69	1.41	0.13
10	- 0.28	0.98	0.01
11	1.47	1.59	0.04
12	0.23	0.72	0,02
13	1.21	0.58	0.17

2. Birey ilk maddeye doğru cevap verdiği için, bu aşamada, ilk maddeden daha zor bir soru sorulmaktadır. Bu sebeple 12. madde seçilmiştir ve yine varsayalım ki birey bu maddeyi de doğru cevaplamıştır. Bu aşamada da yetenek kestirimi yapılamamıştır.

Tablo 2

BBT Uygulamasında Aşama Aşama Kestirilen Parametreler

Aşama	Madde	Cevap	θ'	*I*(θ')	SE(θ')
1	3	1	---	---	
2	12	1	---	---	
3	7	0	1.03	0.97	1.02
4	4	1	1.46	2.35	0.65
5	11	0	1.13	3.55	0.55
6	9	1	1.24	4.61	0.47
7	2	1	1.29	5.05	0.45
8	1	1	1.31	5.27	0.44
9	8	0	1.25	5.47	0.43

3. Bu aşamada ilk iki maddeden daha zor bir madde olan 7. madde uygulanmıştır ve öğrenci bu sefer doğru cevaplayamamıştır. Bireyin madde tepki (cevap) vektörü (1, 1, 0) olarak gösterilebilir. Önceden bilinen madde parametreleri ile bireye ait yetenek değeri (θ’ = 1.03) olarak hesaplanır. Bu hesaplanan yetenek düzeyinde 3 madde için hesaplanan test bilgisi I(θ’ = 1.03) = 0.97 olarak bulunur. Şu ana kadar yapılan geçici yetenek kestirimine ait hata da SE (θ’) = 1.02 olarak hesaplanır. Tüm bu değerler Tablo 2’de görülmektedir.

4. Bu aşamada θ’ = 1.03 yetenek düzeyinde geriye kalan her bir madde tarafından sağlanan bilgi hesaplanır ve en çok bilgi sağlayan madde bir sonraki madde olarak seçilir. Tablo 3’teki değerlere bakıldığında θ’ = 1.03 yetenek düzeyinde bireyin yeteneği hakkında en çok bilgi sağlayan maddenin 4. madde olduğu görülmektedir. Bu sebeple bir sonraki madde 4. madde olmuştur ve öğrenci bu maddeyi doğru cevaplamıştır. (1, 1, 0, 1) cevap örüntüsü için yeni bir yetenek kestirimi hesaplanmıştır (θ’ = 1.46).

Tablo 3

BBT Uygulamasındaki Maddeler Tarafından Her Aşamada Sağlanan Bilgiler

Aşama	θ	Maddeler												
		1	2	3	4	5	6	7	8	9	10	11	12	13
4	1.03	0.034	0.547	---	**1.192**	0.010	0.051	---	0.143	1.008	0.251	1.101	---	0.166
5	1.46	0.179	0.319	---	---	0.004	0.017	---	0.205	0.579	0.136	**1.683**	---	0.175
6	1.13	0.292	0.494	---	---	0.008	0.039	---	0.159	**0.917**	0.219	---	---	0.170
7	1.24	0.249	**0.433**	---	---	0.006	0.029	---	0.175	---	0.187	---	---	0.173
8	1.29	**0.232**	---	---	---	0.006	0.026	---	0.182	---	0.175	---	---	0.174
9	1.31	---	---	---	---	0.005	0.024	---	**0.186**	---	0.168	---	---	0.174
10	1.25	---	---	---	---	0.006	0.028	---	---	---	0.184	---	---	0.173

5. Geriye kalan maddeler için θ’ = 1.46 değeri için madde bilgisi hesaplanır. Maddenin uygulanmasında, yetenek kestiriminde, uygulanmayan maddeler tarafından sağlanan bilginin hesaplanmasında ve uygulanacak bir sonraki maddenin belirlenmesinde

izlenen süreç, yukarıda tarif edildiği gibi devam eder. Bu sürecin devamında, önce 11. madde seçilir, sonra sırasıyla 9., 2., 1. ve son olarak da 8. madde uygulanır. Bireyin yetenek kestirimine ait standart hata önceden belirlenen bir değerin altında azalmaya başladığında BBT uygulamasına son verilir. Tablo 2'deki değerlere bakıldığında, 9. aşamada 8. madde uygulandığında elde edilen standart hata değeri 8. aşamadaki değerden sadece 0.01 azdır. Burada BBT uygulaması son bulmuştur ve bireye ait yetenek değeri ($\theta = 1.25$) olarak kestirilmiştir.

BBT Uygulamalarının Gerçekleştirilmesi

Reshetar'ın (1990) BBT'nin işleyişi hakkındaki örneğinde görüldüğü üzere bir BBT uygulaması yapılırken öncelikle uygun bir yetenek kestirim yöntemi belirlenir, sonra uygulamaya başlanacak ilk madde seçilir, sonra öğrencinin cevabına göre diğer maddeler uygulanır ve son olarak da belli bir koşula göre test sonlandırılır. Tüm bu işlemlerin uygulanması belli kurallar dâhilinde olur. Bu bölümde, bahsedilen BBT adımlarında takip edilen kurallar anlatılacaktır.

Başlama kuralı. BBT uygulamalarında bireylere uygulanacak ilk maddeyi belirleme noktasında alan yazında birkaç öneri ile karşılaşmak mümkündür. Bunlardan ilki sınavı alan bireylerin yeteneklerinin aynı olduğu kabul edilip (örn. $\theta = 0$) herkese aynı sorunun uygulanmasıdır. Fakat bu durum madde teşhir oranı ya da madde kullanım sıklığı noktasından olumsuz sonuçlar doğuracağı için pek tercih edilmemelidir. Bu soruna çözüm olabilecek diğer bir yöntem belli bir yetenek değeri ranjı belirleyip (örn., -0.5 ile 0.5 arası) madde güçlüğü belirlenen aralığa düşen maddelerden rastgele bir seçim yapmaktır. Böylelikle ilk madde olarak seçilen maddelerin kullanım sıklığı oranı kontrol edilmiş olur (Magis & Raîche, 2012; Wainer ve diğerleri, 2000). BBT uygulamasının başlama kuralı olabilecek diğer bir yöntem de, öğrenciye ait geçmiş bilgileri kullanarak yeteneğine dair bir yorum yapma ve bu değerlendirme sonucuna göre bireyin yeteneğine yakın yani bireyin yaklaşık %50 olasılıkla doğru cevaplayabileceği soruları sormaktır (Magis ve diğerleri, 2017; Thompson & Weiss, 2011).

Madde seçme kuralı. İlk maddenin uygulanmasından sonra bireyin verdiği cevaba göre bir sonraki madde uygulanır. Bu maddenin nasıl seçileceğine dair alan yazında oldukça fazla yöntem geliştirilmiş olsa da bunlardan çok azı gerçek BBT uygulamalarında kullanılmaktadır (Han, 2018). En bilinen ve belki de en eski madde seçme kurallarından biri, bireye önceden uygulanmış maddelere bağlı olarak kestirilen geçici en son yetenek kestiriminde en fazla maksimum Fisher bilgi (MFI) değerine sahip maddeyi seçmektir. Madde seçme kurallarının da önemli bir kısmı MFI kuralının bazı farklı versiyonlarıdır. Fisher bilgisine ya da asıl yaygın kullanılan ismiyle madde bilgi fonksiyonuna (Item information function – IIF) ait fonksiyon eşitlik 6'da verilmiştir.

$$I_i(\theta) = \frac{[P_i'(\theta)]}{P_i(\theta)Q_i(\theta)'} \quad \text{… eşitlik 6}$$

Bu eşitlikte $P_i(\theta)$ belli bir yetenek düzeyindeki bireyin maddeye doğru cevap verme olasılığıdır, $Q_i(\theta)$ = 1- $P_i(\theta)$, $P_i'(\theta)$, $P_i(\theta)$ 'nin birinci türevidir. Eğer test maddeleri 2 PLM ile kalibre edildiyse $P_i(\theta)$ eşitlik 7'de verildiği şekilde hesaplanabilir. Böylece eşitlik 6'da verilen Fisher bilgisi eşitlik 8'deki fonksiyona indirgenir. Eşitlik 8'deki D, ölçekleme sabitidir ve değeri 1.702'dir.

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \quad \text{… eşitlik 7}$$

$$I_i(\theta) = D^2 a^2 P_i(\theta) Q_i(\theta) \quad \text{… eşitlik 8}$$

Fisher bilgisinin işleyişini anlamak için şöyle bir örnek verilebilir. Tablo 4'te bir sonraki madde olarak seçilebilecek 4 madde ve bu maddelere ait a ve b parametreleri verilmiştir. Şekil 5'te de bu maddelere ait madde bilgi fonksiyonları gösterilmiştir. Madde bilgi fonksiyonlarına bakıldığında görülmektedir ki, eğer bireyin son kestirilmiş yetenek düzeyi 0.5 ise MFI kuralına göre bir sonraki madde için en uygun seçenek 1. maddedir. Çünkü grafikten de anlaşılacağı üzere (θ=0.5) yetenek düzeyinde en çok bilgiyi o sağlamaktadır. θ 1.5'e eşit olduğunda ise bir sonraki madde olarak 2. madde seçilmelidir.

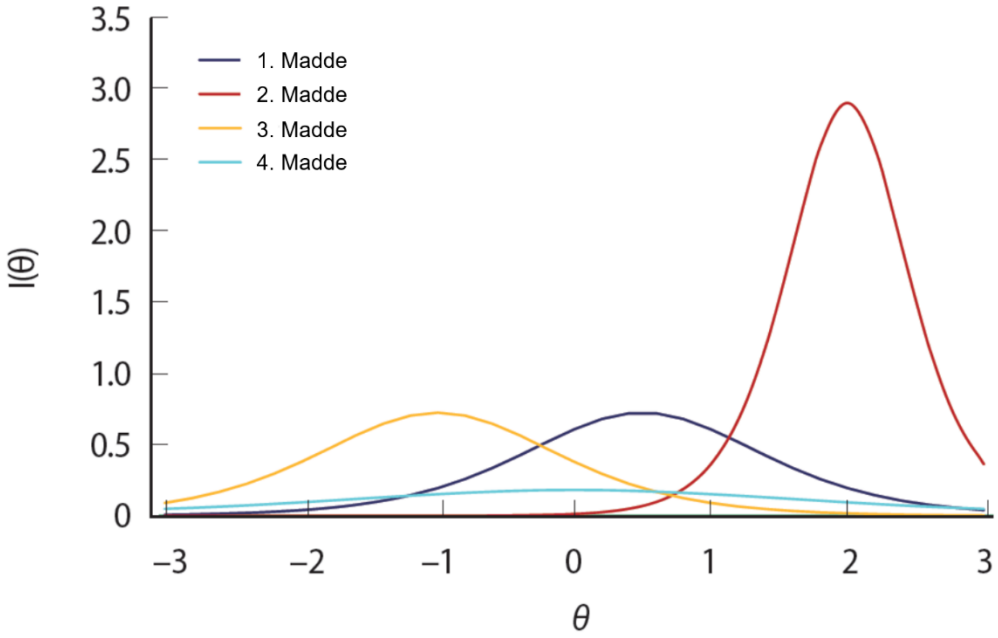
Tablo 4

Örnek IIF İşleyişini Göstermek İçin Belirlenen Seçilebilir Maddeler

Madde	a parametresi	b parametresi
1	1.0	0.5
2	2.0	2.0
3	1.0	-1.0
4	0.5	0.0

Şekil 5

Örnek Maddelere Ait Madde Bilgi Fonksiyonları



BBT madde seçim kuralları içinde MFI yöntemi anlaması ve uygulanması en kolay olanıdır ve her zaman en yüksek Fisher bilgisi değerini vermesi beklenen madde seçildiği için MFI yöntemi maksimum test bilgi fonksiyonunu sağlamasıyla bilinir. Gerçek BBT uygulamalarında MFI yöntemi nadiren uygulanır çünkü bu yöntem madde seçiminde oldukça tutucudur. Örneğin 4. Madde düşük a parametresi değerine (a=0.5) sahip olduğu için madde bilgi fonksiyonu açısından çok yüksek bir değer gösterememektedir. Hatta -3 ve +3 aralığında her θ düzeyi için daha yüksek bilgi sağlayan başka maddeler vardır. Bu sebeple MFI yöntemi söz konusu olduğunda 4. maddenin hiçbir şekilde seçilme şansı yoktur (Han, 2018).

Görüldüğü üzere MFI bireyin o anlık kestirilen yetenek düzeyine en yakın güçlüğe sahip maddeyi seçmeyi hedefler ve bu madde o düzeydeki en dik madde karakteristik eğrisine sahip madde olmalıdır. MFI gibi geleneksel madde seçim yöntemlerinin yüksek ayırt ediciliğe sahip maddeleri çok sık kullanması bazı öğrencilerin bu maddeleri ezberlemesine ve kopya çekmeye imkân tanıdığı için test güvenliğini tehlikeye düşürmektedir (Hau & Chang, 2001; Xiao & Bulut, 2022). Ayrıca bu seçicilik israfa da yol açar çünkü madde yazmak çok masraflı bir iştir ve ABD şartlarına göre bu masraf madde başına 300 doları bulmaktadır (Chang, 2015; Downing, 2006). Hatta işin içine MTK parametrelerinin hesaplanması ya da simülasyonların yapılması gibi karışık istatistiksel işlemler de dahil olursa masraf madde başına 1000 dolar ve üzeri bir rakama ulaşabilmektedir. Çok büyük madde havuzuna sahip testlerde masraf böylece milyon dolarları bulabilmektedir (Downing, 2006).

MFI yönteminin bu gibi dezavantajlarına çözüm bulabilmenin bir yolu maddelerin bu kadar sık kullanılmasının önüne geçmektir. Yani madde kullanım sıklığını kontrol etmektir. Madde kullanım sıklığı kontrol yöntemi olarak genellikle iki farklı yola başvurulmaktadır. Bunlardan ilki maksimum madde kullanım sıklığını kontrol altında tutabilmek için her madde için bir madde kullanım sıklığı parametresi belirlemektir (Xiao & Bulut, 2022). Bu yöntemde BBT algoritmasında .6 gibi bir oran tanımlanır. Bu tanımlanan orana göre bir maddenin kullanım oranı en fazla %60 olacak şekilde bireylere uygulanır. Yani 100 kişilik bir öğrenci grubunda yüksek ayırt ediciliğe sahip maddeler en fazla 60 kişiye uygulanır. İkinci yol ise en uygun (optimal) madde sayısını birden fazla olacak şekilde seçip, sonra bu maddeler arasından rastgele tekrar bir seçim yaparak bireye uygulayan “randomesque” yöntemidir (Xiao & Bulut, 2022). *Randomesque* değeri yine BBT algoritmasında tanımlanır fakat bu sefer bir tam sayı olarak girilir. Örneğin bu değer 5 olarak girildiğinde bireyin anlık kestirilen yetenek düzeyinde en çok bilgi veren 5 madde belirlenir. Bu 5 madde içinden bir tanesi tesadüfi olarak seçilir ve bireye uygulanır. Böylece yüksek ayırt ediciliğe sahip maddelerin kullanım sıklığı bir derece kontrol edilmiş olur. Madde kullanım sıklığını kontrol amacıyla

genel olarak bu iki yöntem kullanılır fakat alan yazında, öncül simülasyonlara bağlı olarak belirlenen maksimum madde kullanım sıklığı oranına göre oluşturulan, Sympson ve Hetter yöntemi (Hetter & Sympson, 1997), Davey ve Parshall yaklaşımı (Davey & Parshall, 1995; Parshall ve diğerleri, 1998) ve Stocking ve Lewis yöntemleri (Stocking & Lewis 1995, 1998) gibi daha karmaşık yöntemler de bulunmaktadır (Magis & Raîche, 2012). Bu çalışmada BBT uygulamasını geliştirirken kullanılan Concerto platformu, arka planda *catR* paketini kullanır (Aybek, 2016; Scalise & Allen, 2015). *catR* paketinde de madde kullanım sıklığı kontrol yöntemi olarak şu an için sadece *randomesque* yöntemi bulunmaktadır.

Bazı maddelerin çok sık kullanılmasına yönelik bir diğer çözüm yolu da MFI kadar seçici olmayan madde seçim yöntemlerini kullanmaktır. Alan yazında bu amaçla geliştirilen birçok madde seçim yöntemi vardır. *catR* paketin ilk sürümlerinde MFI (maximum Fisher information – maksimum Fisher bilgi) kuralı da dâhil olmak üzere MEPV (minimum expected posterior variance – minimum beklenen sonsal varyans), MLWI (maximum likelihood weighted information – maksimum olabilirlikle ağırlıklandırılmış bilgi), MPWI (maximum posterior weighted information – maksimum sonsal dağılımla ağırlıklandırılmış bilgi) ve MEI (maximum expected information – maksimum beklenen bilgi) Urry kriteri ve tamamen rassal seçim olmak üzere 7 farklı madde seçim kuralı bulunmaktadır (Magis & Raîche, 2012). *catR*'ın yeni versiyonunda theOpt kuralı, KL yöntemi (Barrada ve diğerleri; 2009), KLP yöntemi, aşamalı (progressive) yöntem ve oransal (proportional) yöntem (Segall, 2004) olmak üzere 5 tane daha madde seçim kuralı eklenmiştir (Magis & Barrada, 2017). Madde seçim yöntemleri sadece *catR* paketinde bulunan yöntemlerle de sınırlı değildir. Farklı amaçlarla geliştirilmiş birçok yöntem vardır ve ilgili kaynaklardan (Magis ve diğerleri, 2017; van der Linden & Pashley, 2010) bu yöntemler hakkında ayrıntılı bilgi elde edinilebilir.

Sonlandırma kuralı. BBT uygulamalarında testin ne zaman sonlandırılacağı çok önemlidir. Çünkü test olması gerekenden çok kısa tutulduğunda yetenek kestirimleri çok doğru bir şekilde yapılamayacaktır. Eğer test çok uzun sürerse de zaman ve kaynak

bakımından israfa gidilmiş olur ve geçersiz sonuçlar elde edilmesine neden olacak düzeyde sınavı alan bireyde yorulmalar ve performans düşüklüğü baş gösterebilir (Linacre, 2000).

BBT uygulaması şu durumlarda sonlandırılır (Linacre, 2000; Tian ve diğerleri, 2007);

1. Madde havuzundaki tüm maddeler kullanıldığında test sonlandırılır. Bu duruma genelde az sayıda madde içeren madde havuzlarının kullanıldığı çalışmalarda rastlanır.

2. Maksimum test uzunluğuna erişildiğinde BBT uygulamasına son verilir. Bu durumlarda sınavı alan öğrencilere uygulanacak soru sayısı önceden bir sayı ile sınırlandırılmıştır. Bu sayı genelde testin kâğıt-kalem formundaki madde sayısına eşit olur. Sabit uzunluklu sonlandırma kuralı, sınavı alan bireylerin aynı sayıda madde ile karşılaştıklarını bilip rahatlamaları açısından tercih edilebilir. Fakat bu durum her bir bireyin aynı ölçme kesinliğinde ölçülememesi gibi bir yan etki doğuracaktır (Chalhoub-Deville ve Deville, 1999).

3. Bireye ait yetenek yeterli kesinlikte ölçüldüğü zaman test sonlandırılır. Birey soruları cevapladıkça onun yeteneği hakkında daha fazla istatiksel bilgi elde edilir ve ölçmenin standart hatasında giderek azalma ile yetenek kestiriminin kesinliğinde artış görülür. Ölçme kesinliği yeterli seviyeye ulaştığında teste son verilir. Genelde bu standart hata değeri 0.2 logit olarak belirlenir.

4. Yetenek kestiriminin geçme-kalma kriterinden yeterince uzak olduğu durumlarda da BBT uygulamasına son verilebilir. Bu durum sınavı alan bireylerin düzeyinin önceden belirlenmiş bir geçme-kalma ölçütüne göre istatiksel olarak artık kesinleşmesine binaen ortaya çıkar. Yetenek kestiriminin ölçüt olarak belirlenen düzeyden en az iki standart sapma uzakta olduğu ya da mevcut geçme-kalma kararını değiştirecek yeterli sayıda madde kalmadığı durumlarda test sonlandırılır.

5. Sınavı alan bireyin test dışı davranışlar göstermesi de testi sonlandırmak için bir kural olarak belirlenebilir. BBT için kullanılan yazılım ya da program, sürekli aynı cevabı işaretleme ya da aynı cevap örüntüsünü işaretleme gibi anormal cevap setlerini tespit

edebilir. Çok hızlı ya da çok yavaş cevaplama gibi normal olmayan durumlarla karşılaşıldığında da test sonlandırılabilir.

Şu ana kadar bu bölümde çalışmanın kuramsal temelini oluşturan MTK'dan bahsedildi. Ayrıca arka planında MTK'yı kullanarak ölçme kesinliğinden ödün vermeden daha kısa testler geliştirilmesine olanak tanıyan BBT uygulamaları ayrıntılı bir şekilde anlatıldı. Bundan sonraki kısımda ise bu çalışmanın amacı ve yöntemi noktasından benzerlikler gösteren alan yazındaki çalışmalara değinilmiştir.

İlgili Araştırmalar

Bu bölümde öncelikle İngilizce kelime bilgisini ölçmek amacıyla BBT uygulamalarının kullanıldığı araştırmalara yer verilmiştir. Sonra çalışmada İngilizce kelime bilgisini ölçmek amacıyla kullanılan VST hakkında yapılan geçmiş çalışmalara değinilmiştir. Son olarak da madde havuzunun ikili puanlanan maddelerden oluştuğu BBT çalışmalarında madde kullanım sıklığını kontrol edebilmek için *randomesque* yönteminin kullanıldığı çalışmaların bulgularından kısaca bahsedilmiştir. Bölüm sonunda ise ilgili araştırmalar özetlenip, yapılan çalışmanın ne gibi farklılık ve yenilikler içerdiği anlatılmıştır.

İngilizce Kelime Bilgisinin BBT Uygulamaları ile Ölçüldüğü Araştırmalar

Yurtiçi ve yurtdışı beraber olmak üzere tüm alan yazına bakıldığında İngilizce kelime bilgisini BBT uygulamaları ile ölçen yayınlanmış sadece birkaç çalışma (Kezer & Koç, 2014; Mizumoto ve diğerleri, 2019; Tseng, 2016) olduğu görülmektedir. Hatta İngilizce kelime bilgisini BBT ile ölçen ilk çalışma ülkemizde gerçekleştirilmiştir. Kezer & Koç (2014) geleneksel kâğıt-kalem test yöntemi ile BBT uygulamalarını karşılaştırmak amacıyla yürüttükleri çalışmalarında bir İngilizce kelime bilgisi testi kullanmışlardır. Test, çalışma kapsamında araştırmacılar tarafından üç İngilizce uzmanı ile birlikte geliştirilmiştir ve İngilizce metinlerde en sık kullanılan ilk 3000 kelime içinden seçilen kelimelerle 100 soruluk bir test oluşturulmuştur. Madde parametreleri 2 parametreli lojistik modele göre kalibre edilmiştir. 2PLM ile uyum sağlamayan maddeler testten çıkarılıp, kalan 72 madde ile BBT

uygulaması için madde havuzu oluşturulmuştur. Testin BBT versiyonu PHP (Personal Home Page - Kişisel Ana Sayfa) dili kullanılarak geliştirilmiştir. Teste başlangıç kuralı olarak güçlük parametresi - .50 ile .50 aralığında olan bir maddenin rastgele olarak seçilmesi belirlenmiştir. Yetenek kestirim yöntemi olarak da ML (Maximum Likelihood – En çok olabilirlik) tercih edilmiştir. Sonlandırma kuralı olarak ise iki yöntem belirlenmiştir. Bunlardan ilki ardışık olarak kestirilen yetenek değerlerine ait hatalar arasındaki fark .01'den küçük ise testin sonlandırılmasıdır. İkincisi yetenek kestirimine ait standart hata değeri .50'nin altında ise testin sonlandırılmasıdır. Bu iki yöntemden biri sağlandığında test sonlandırılmıştır. Kâğıt-kalem uygulaması ve BBT uygulaması sonucunda kestirilen yetenek parametreleri arasında yüksek pozitif korelasyon ($p < .01$, $r = .86$) bulunmuştur. Ayrıca çalışma kapsamında İngilizce kelime bilgisi BBT ile ölçüldüğünde %78 daha az madde kullanıldığı bulunmuştur. Araştırmacılar ayrıca *post-hoc* simülasyonlar da gerçekleştirmişlerdir. Fakat bu simülasyonları gerçek BBT uygulamasında kullanılacak olan BBT koşullarını belirlemek için değil, yetenek parametrelerinin ve ölçme kesinliği göstergelerinin karşılaştırılması amacıyla yapmışlardır. İki başlangıç yetenek düzeyi (sıfır ve önceden kestirilen yetenek düzeyi), üç yetenek kestirim yöntemi (ML, EAP ve MAP), üç tane de sonlandırma kuralı (sabit uzunluk, standart hata < .50 ve standart hata < .30) olmak üzere toplam 18 koşul test edilmiştir. Koşulların hepsinde kâğıt-kalem testinden kestirilen yetenek kestirimleri ile BBT simülasyonları sonucu elde edilen yetenek kestirimleri arasında yüksek pozitif korelasyonlar bulunmuştur. Ölçme kesinliği için RMSD (Root Mean Squared Difference - Farklılıkların Ortalama Karekökü) değerleri karşılaştırılmıştır. Ölçme kesinliğinin yetenek kestirim yöntemine göre ve başlangıç yetenek düzeyine göre farklılaşmadığı bulunmuştur.

Tseng (2016) de Tayvan'da yaklaşık 1500 lise öğrencisi ile yürüttüğü çalışmasında, İngilizce kelime bilgisinin ölçülmesine bir seçenek olarak BBT uygulamasının kullanışlılığını ve uygulanabilirliğini araştırmıştır. Çalışmada oluşturulan madde havuzu, Tayvan'da üniversite giriş sınavı merkezi tarafından hazırlanan bir kelime listesinden faydalanarak geliştirilmiştir. Sınav merkezi bu listeyi oluştururken Britanya, Kanada, Japonya ve ABD'de

kullanılan çeşitli ders kitapları ve kelime listelerini dikkate almıştır. Liste 6480 kelimeden oluşmaktadır ve kelimeler kullanım sıklığına göre altı düzeye ayrılmıştır. Madde havuzu da her 27 kelimeye bir kelime düşecek şekilde yapılmış ve her düzey böylece 40 kelimeden oluşmuştur. Toplam 240 maddenin parametreleri Rasch modeli ile RUMM2030 programı kullanılarak kestirilmiştir. Modele uyum sağlamayan maddelerin havuzdan çıkarılması ile 180 maddelik bir nihai havuz oluşturulmuştur. BBT uygulaması, Assessment Systems tarafından geliştirilen CATPRO 1.0.0.4 kullanılarak gerçekleştirilmiştir. Yetenek kestiriminde Bayes kestirim yöntemlerinden MAP (Maximum a Posteriori) kullanılmıştır. Uygulanacak ilk madde, madde güçlüğü -0.5 ve +0.5 aralığında olan maddelerden rassal olarak seçilmiştir. Sonlandırma kuralı olarak sabit ve değişken uzunluklu olmak üzere iki farklı durum belirlenmiştir. Sabit uzunlukta koşullar 30, 60 ve 90 madde olarak, değişken uzunlukta ise standart hatanın .40, .30 ve .20 olduğu koşullar belirlenmiştir. Sonlandırma kuralı olarak sabit uzunluğun kullanıldığı durumlarda BBT uygulaması en fazla 30 dakika içinde tamamlanırken, değişken uzunluk kullanıldığında uygulama sadece 5-10 dakika aralığında sürmüştür. BBT uygulamasının ardından aynı öğrencilere 20 dakika sonra madde havuzundaki tüm sorular kâğıt-kalem formatında sorulmuş ve bu formun tamamlanması ortalama bir saati bulmuştur. 6 farklı sonlandırma koşulu altında elde edilen yetenek kestirimleri ile kâğıt-kalem formatında uygulanan ve tüm soruların sorulduğu testten elde edilen yetenek kestirimleri arasında çok yüksek ilişki katsayıları bulunmuştur ($r > 0.90$). Testi sonlandırmak için kullanılan ortalama madde sayısı standart hatanın .40 olduğu koşulda 10, .30 olduğu koşulda 18, .20 olduğu koşulda ise 43 olmuştur. Çalışmada ayrıca BBT uygulamasının, önceden belirlenen bir kesme noktasına göre İngilizce öğrenen bireylerin yeterli kelime bilgisine sahip olup olmadığını sınıflamada da başarılı bir şekilde kullanılabileceği gösterilmiştir. Sonuç olarak, Tseng'in çalışması BBT uygulamasının, İngilizce öğrenen bireylere ait İngilizce kelime bilgisinin tespit edilmesinde klasik kâğıt-kalem testleri kadar başarı gösterdiği ve bunu da daha az sayıda madde kullanarak, daha kısa sürede ve daha düşük hata değeri ile yaptığını göstermesi açısından önemlidir.

Mizumoto vd. (2019) yaptıkları çalışmada, Sasao ve Webb (2017) tarafından geliştirilen Kelime Birimleri Düzeyi Testinin (Word Part Levels Test – WPLT) BBT uygulamasını geliştirmişlerdir. Bir kelimenin birimleri ön ek (prefix), kök (root) ve son ekten (suffix) oluşmaktadır. Örneğin "transportable" kelimesi, trans (ön ek), port (kök) ve able (son ek) birimlerinden oluşur. Ön ek ve son ekler beraber ek (affix) olarak adlandırılırlar. WPLT İngilizce öğrenen bireylerin ek (affix) bilgisini ölçer. Araştırmacıların aktardıklarına göre kelime birimlerinin bilinmesi ile İngilizce kelime bilgisi arasında önemli ilişki bulunmaktadır. WPLT'nin daha kullanışlı ve bir tanı testi olarak daha ulaşılabilir olması için WPLT'nin BBT versiyonu geliştirilmiştir. WPLT'nin şekil (form), anlam (meaning) ve use (kullanım) olmak üzere üç içeriği vardır. Testin bu her bir bölümünde sırasıyla 115, 73 ve 56 madde bulunmaktadır. Bu maddelerin kalibrasyonu Sasao ve Webb'in (2017) çalışmasında toplanan veri kullanılarak yapılmıştır. Fakat ilk çalışmada Rasch modeli kullanılırken, bu çalışmada madde parametrelerini kestirmek için 2 PLM kullanılmıştır. Üç farklı bölüm olduğu için madde havuzu da üçe ayrılmıştır. BBT uygulamasında sonlandırma kuralı olarak sabit uzunluk seçeneğinde karar kılınmıştır. Araştırmacılar her ne kadar değişken uzunluklu koşullarla daha az hata ile kestirim yapılabileceğinin farkında iseler de, bazen bu kuralla çok fazla sayıda madde uygulanması gerektiren durumlar olabildiği için sabit uzunluğu tercih etmişlerdir. Her bölümde ne kadar madde uygulandıktan sonra testin sonlandırılacağına karar vermek için catR paketi kullanılarak Monte Carlo simülasyon çalışmaları yapmışlardır. Simülasyonların sonuçlarına göre şekil bölümü için 20, anlam bölümü için 15 ve kullanım bölümü için 10 madde, test sonlandırma kuralı olarak belirlenmiştir. Teste başlangıç kuralı olarak b parametresi -0.3 – 0.3 aralığında olan maddelerden rassal olarak bir maddenin seçilmesi belirlenmiştir. Madde seçim kuralı olarak da Maksimum Fisher Bilgi yöntemi, yetenek kestirimi için de Bayes modeli kullanılmıştır. BBT uygulamasındaki tüm bu işlemler R dili, HTML ve MySQL veri tabanı ile beraber çalışan Concerto platformu kullanılarak yapılmıştır. Çalışma grubu Japonya'nın batısında farklı üniversitelerde İngiliz dili eğitimi bölümünde öğrenim gören 760 öğrenciden oluşmuştur. Çalışma bulgularına göre WPLT'nin BBT versiyonu, öğrencilerin İngilizce kelime birimleri

bilgisini çok daha az madde ile yaklaşık 10 dakika gibi bir süre içerisinde kestirmiştir ve bunu WPLT'nin kâğıt-kalem formuna göre benzer ya da daha fazla ölçme kesinliğine ulaşarak yapmıştır.

VST ile İlgili Araştırmalar

Vocabulary Size Test (VST), Nation ve Beglar (2007) tarafından geliştirilen bir İngilizce kelime bilgisi testidir. Çoktan seçmeli bir formata sahip VST'de her bir soru kısa bir soru kökünden ve dört seçenekten oluşmaktadır. VST hakkında ayrıntılı bilgi "Veri toplama araçları" başlığında verilmiştir. Bu bölümde VST ile ilgili araştırmalara yer verilmiştir. VST üzerine yapılan çalışmalar bugüne kadar genellikle iki tarzda olmuştur. Bunlardan ilki testin çoktan seçmeli formatından dolayı şans başarısı ile doğru cevap verme olasılığının yüksek olması ve bunun da aşırıtahmin (overestimation) problemine sebep olması üzerine yapılan çalışmalardır. İkincisi de VST'nin iki dilli versiyonlarını geliştirmek üzerine yapılan çalışmalardır. VST'nin geçerlik ve güvenirliği üzerine yapılan araştırma sayısı ise sadece birkaç tanedir. Bunlardan biri Beglar (2010) tarafından yürütülen Rasch temelli bir çalışmadır.

Beglar (2010), bu çalışmadaki BBT uygulamasının madde havuzunu oluşturan kelime bilgisi testi olan VST'nin geçerlik araştırmasını yapmıştır. Messick'in (1989, 1995) eğitimde ve psikolojide ölçme çalışmaları için tanımladığı geçerlik kanıtları ve Wolfe ve Smith (2007a, 2007b) tarafından tarif edilen Rasch temelli, ölçme aracının geçerliğini test etme yöntemlerine göre bu çalışma yürütülmüştür. Çalışmaya Japonya'da çeşitli üniversitelerde okuyan 19'u ana dili İngilizce ve 178'i ana dili Japonca olan toplam 197 kişi katılmıştır. Öğrencilere ait İngilizce yeterlikleri, gerek o ana kadar girmiş oldukları sınavlarla ya da önceki aldıkları eğitimle az çok tahmin edilebildiği için öğrenciler İngilizce yeterliklerine göre üç gruba ayrılmıştır. Yeterliği yüksek olan gruba testteki maddelerin tamamı (k = 140) sorulmuş, orta düzeydeki öğrencilere 80 maddelik bir form verilmiş ve düşük düzeydekiler ise 40 maddelik bir formu cevaplamışlardır. WINSTEPS 3.64.2 (Linacre, 2007) kullanılarak Rasch modeline göre parametre kestirimi yapmıştır. Düşük yeteneğe sahip öğrencilere

sorulan 40 soru, diğer gruplara da sorulduğu için bu 40 maddeden 23'ü ortak (anchor) madde olarak seçilmiş ve geriye kalan maddelerin kalibrasyonu bu vesileyle yapılmıştır. Wright haritası ya da bir diğer adıyla madde-birey haritası kullanılarak ve madde tabakası (item strata) hesaplamaları yapılarak VST'nin, İngilizce kelime bilgisinin her düzeyinde başarılı bir ölçme yaptığı gösterilmiştir. Yani farklı İngilizce kelime bilgisi düzeyine sahip bireylerin bu düzeylerini belirleyebilmek için VST oldukça geçerli bir ölçme aracıdır ve düzey başına 10 kelime, yeterli ölçme kesinliği sağlamak için oldukça yeterlidir. Rasch modeline uyum sağlamayan maddeleri tespit etmek amacıyla Rasch standartlaştırılmış madde ağırlıklandırılmış ortalamanın karesi (Rasch standardized item weighted mean- square) uyum istatistikleri hesaplanmıştır. Düşük uyum ya da aşırı uyum gösteren madde sayısı 140 maddeye oranla %5'in altında kalmış ve bu bulgular da madde yazımında yeterince yüksek özen gösterildiğine delil olarak sunulmuştur. Rasch modeli VST için toplam varyansın %85.6'sını açıklamış ve artıklar arasında herhangi bir sistematik ilişki saptanmamıştır. Tek boyutluluk varsayımını karşılayan VST'de herhangi bir ikinci boyuta dair anlamlı bulgulara ulaşılamamıştır. Parametre değişmezliği için öncelikle değişen madde fonksiyonu (DMF) analizleri yapılmıştır. Ana dili İngilizce olan bireylerin çoğunluğu erkek, yüksek düzeyde yeterlilikleri olan bireylerin çoğu da kız öğrenciler olduğu için sadece düşük ve orta düzey İngilizce yeterliğine sahip bireylere sorulan 80 soru için cinsiyet açısından DMF analizi yapılabilmiştir. 80 maddeden sadece iki tanesinin DMF gösterdiği saptanmıştır, ki bu iki madde de önceden yapılan analizde modele uyum göstermeyen birkaç maddeden ikisidir. Değişmezlik ayrıca 140 maddeyi rassal olarak iki gruba ayırıp tüm formdan ve bu iki formdan elde edilen yetenek parametrelerinin arasındaki korelasyonlara bakarak da test edilip doğrulanmıştır. 70'er maddelik iki formun birbirleri arasındaki korelasyon 0.95 olarak hesaplanmış ve tüm form ile ilişki katsayıları aynı derecede yüksek ($r = 0.98$) bulunmuştur. Son olarak da Linacre (2007) tarafından önerilen daha güçlü bir yöntemle parametre değişmezliği test edilmiştir. Madde artık yükleri pozitif ya da negatif olmasına göre maddeler yine iki alt ölçeğe ayrılmış ve yetenek parametreleri bu iki grup madde ile tekrar hesaplanmıştır. Korelasyon katsayısı bu sefer 0.84 olarak hesaplanmıştır. Bu üç analizin

de sonuçları dikkate alındığında, VST maddelerinin farklı kombinasyonları yüksek derecede parametre değişmezliği göstermiş ve benzer birey yetenek parametreleri üretmiştir. Çalışmanın sonuçları toparlanacak olunursa; bu çalışmanın madde havuzunu oluşturacak VST maddelerinin büyük çoğunluğu Rasch modeline yeterli uyum göstermiş, madde artıkları analiz edildiğinde yüksek derecede psikometrik tek boyutluluk sağlamış ve testin farklı formları tarafından üretilen benzer yetenek parametrelerin delaletiyle de iyi derecede ölçme değişmezliği göstermiştir.

Stewart (2014) çoktan seçmeli testlerin doğasında yer alan bir gerçeği VST açısından dile getirmiştir. VST'nin çoktan seçmeli formatının bireylerin İngilizce kelime bilgisini kestirirken olduğundan yüksek kestirimler elde edilmesine neden olduğunu belirtmiştir. Bu durumda şans başarısının önemli bir rol oynadığını ve bu şans başarısını test etmede Rasch temelli analizlerin yeterli olmadığından bahsetmiştir. Şans başarısının VST maddelerini doğru cevaplama üzerindeki etkisinin ihmal edilebilir düzeyde olup olmadığının test edilebilmesi için 3PLM'ye göre de analizler gerçekleştirip, sonuçlarını Rasch ile yapılan analiz bulguları ile kıyaslanması gerektiğinden bahsetmiştir. Ayrıca VST maddelerindeki seçenek sayısının sadece 4 olduğunu, bunun şans ile doğru cevaplama olasılığını arttırdığını ve VST geliştiricilerin yeni seçenekler eklemek gibi şans başarısını elimine eden önlemleri alması gerektiğine dikkat çekmiştir.

Stoeckel ve Sukigara (2018) da, VST ile İngilizce kelime bilgisinin ölçülmesi durumunda karşılaşılan aşırıtahmin (overestimation) sorununa dair bir araştırma yapmışlardır. Alan yazından çeşitli kaynaklar ile destekledikleri görüşlerine göre VST'nin çoktan seçmeli formatı İngilizce kelime bilgisinin olduğundan fazla olarak kestirilmesine sebep olmaktadır. Bunun bir sebebi şans başarısının yüksek olması iken diğer bir sebebi de VST'nin ölçtüğü kelime bilgisi formatının gerçek hayatta İngilizce metinleri okurken ihtiyacımız olan kelime bilgisi formatından farklı olmasıdır. VST'de bize bir kelime verilir ve bunun anlamının seçenekler arasından doğru tahmin edilmesi beklenir. Fakat gerçekte, bir İngilizce metin okunurken karşılaştığımız kelimelerin anlamını seçeneksiz bilmemiz ya da

bağlamdan çıkarmamız gerekmektedir. İngilizce kelime bilgisinin gerçek koşullara uygun olarak ölçülmesi de ancak sözlü görüşmeler ya da çeviri görevleri gibi yöntemlere başvurmak ile mümkündür ve tabi ki bu tür yöntemler çoktan seçmeli testler ile yapılacak ölçmelerden çok daha zor ve vakit alıcıdır. Araştırmacılar hem çoktan seçmeli formatın uygulanabilirliğinden taviz vermemek hem de aşırıtahmin problemine bir çözüm üretebilmek için çalışmalarında VST'nin 80 soruluk (ilk sekiz düzey) farklı bir formatını geliştirmişlerdir. Bilgisayar ortamında geliştirdikleri bu test yine çoktan seçmelidir fakat seçeneklerin hepsi aynı anda bireye gösterilmemektedir. İlk önce soru kökü ekrana gelmekte, sonra seçeneklerden bir tanesi ekrana gelmektedir ve testi alan bireye bu seçeneğin doğru olup olmadığı sorulmaktadır. Eğer *doğru* seçeneğini seçerse diğer seçenekler gösterilmeden bir sonraki soruya geçilmektedir. Eğer *yanlış* seçeneği seçilirse, sorulan seçenek ekrandan silinmekte ve diğer seçenek bireye gösterilmektedir. Araştırmacılar her zaman son seçenek doğru olacak şekilde testi planlamışlardır. Fakat her soruya aynı sayıda seçenek koymamışlardır ve her soruya ait seçenek sayısı öğrenciler tarafından bilinmemektedir. Seçenek sayısı 1-10 aralığında değişmektedir ve ortalama seçenek sayısı 4.3'tür. Soruların çoğu 1, 2, 3 veya 4 seçeneklidir. Testi bu formata sokmak için araştırmacılar VST'nin formatında değişiklikler yapmış, orijinali dört seçenekli olan soruların bazılarında seçenekleri azaltırken bazılarında 10'a kadar çıkarmıştır. Yeni geliştirdikleri bu formatı Japonya'da iki farklı üniversitede 131 öğrenciye uygulamışlardır. Aşırıtahmin sorununu test edebilmek için aynı öğrencilere VST'nin orijinal formu (ilk 80 soru) ve ayrıca VST'nin soru köklerine benzer olarak geliştirilen çeviri görevlerinden oluşan bir test de uygulanmıştır. Yani seçenekleri aşamalı olarak gelen yeni formattaki VST, orijinal VST ve çeviri görevlerini içeren test (ölçüt) olmak üzere üç farklı ölçme aracı kullanılmıştır. Ölçüt test tüm gruba uygulanırken, orijinal format ile yeni format rastgele dağıtılmıştır. Çalışma sonunda elde edilen sonuçlara göre ölçüt puanı açısından orijinal formatı alan grubun ortalaması ((M = 30.9) ile yeni formatı alan grubun ortalaması (M = 31.7) arasında manidar fark bulunmamıştır, $t(129) = 0.705$, $p = .482$. Her iki formattan elde edilen puanlar da ölçüt puanları ile yüksek derecede korelasyona sahiptir fakat orijinal formatın uygulandığı

öğrencilerin İngilizce kelime bilgileri (M = 54.2), yeni formatın uygulandığı bireylerden (M = 38.4) oldukça yüksek bulunmuştur, t (129) = 13.80, p < .001, d = 2.41. Bu bulgulara göre araştırmacılar şans ile doğru cevaplama olasılığından dolayı VST'nin çoktan seçmeli formatının öğrencilerin İngilizce kelime bilgisini olduğundan fazla kestirdiğini belirtmişlerdir.

Zhang da (2013) VST'deki şans başarısını araştırmıştır ve "I don't know" (bilmiyorum) seçeneğinin teste eklenmesinin test sonuçları üzerinde etkisi olup olmadığını incelemiştir. Ayrıca VST uygulanan bireylerin doğru cevabı tahmin ile cevaplarken onların bu davranışlarını etkileyen faktörleri de araştırmışlardır. Çalışma Çin'de bir üniversitede birinci sınıf öğrencisi olan 111 kadın ve 39 erkekten oluşan toplam 150 kişi ile yürütülmüştür. Öğrencilere üç farklı test verilmiştir. Öğrenciler 50'şer kişi olarak rastgele üç gruba bölünmüş ve her bir gruba VST'nin farklı bir versiyonu verilmiştir. İlk gruba orijinal VST (V1), ikinci gruba da *bilmiyorum* seçeneği olan VST (V2) verilmiştir. Üçüncü gruba da yine *bilmiyorum* seçenekli VST (V3) verilmiştir fakat anlamını bilmedikleri bir kelimenin sorulduğu soru ile karşılaştıklarında *bilmiyorum* seçeneğini işaretlemek yerine cevabı yanlış verirlerse toplam puanlarından bir puan eksiltecek şekilde bir ceza ile karşılaşacakları belirtilmiştir. İlk iki versiyonda yanlış cevaplar için herhangi bir ceza uygulanmamıştır. VST'nin bu üç versiyonundan sonra tüm öğrencilere bir okuma parçası verilmiş ve bu okuma parçası ile alakalı birkaç alıştırma sorusunu cevaplamaları istenmiştir. Dikkat dağıtma amaçlı gerçekleştirilen bu kısımdan sonra yine öğrencilerin hepsine VST tekrar uygulanmıştır. Fakat bu sefer herhangi bir şık verilmemiş, açık uçlu bir formatta öğrencilerden veri toplanmıştır. Zhang, bu formatta bir uygulama yapmanın amacının kısmi bilgi ile cevaplama davranışını test etmek olduğunu belirtmiştir. Çalışma sonuçlarına göre V3'ten elde edilen puanlar çok daha düşük olmuştur. Çünkü yanlış cevaba bir puan eksiltme cezası verilmesi sorulara cevap verirken tahmin yapma sayısını azaltmıştır. Fakat bu cezanın, kısmi bilgi ile verilen doğru cevap sayısını da azalttığı bulunmuştur. Araştırmacı, kullanım amacına göre uygun VST versiyonun değişebileceğini belirtmiştir. Eğer amaç kısmi bilgiyi de kapsayacak şekilde öğrencilerin İngilizce kelime bilgisine dair bilgi elde etmek ise, VST'nin orijinal

formunun kullanılması daha uygun olacağı, fakat amaç eğer şans başarısından tamamen arınık bir şekilde öğrencin kelime bilgisini tam bir şekilde kestirmek ise ceza puanını ve *bilmiyorum* seçeneğini içeren versiyonun kullanılması daha uygun olacağı sonucuna varılmıştır.

Madde Kullanım Sıklığı Kontrol Yöntemi Olarak randomesque Yönteminin Kullanıldığı Araştırmalar

İkili puanlanan maddeler ile çalışılırken madde kullanım sıklığını kontrol etmek amacıyla randomesque yönteminin kullanıldığı çalışma oldukça sınırlıdır. Alan yazın tarandığında ulaşılan çalışmalara burada kısaca değinilmiştir.

Leroux vd. (2013), McClarty vd. (2006) tarafından değişken uzunluklu BBT uygulamalarında kullanılmak amacıyla geliştirilen bir madde kullanım sıklığı kontrol yöntemi olan PR-SE (progressive-restricted standard error) yöntemini ölçme kesinliği, madde havuzunun dengeli kullanımı gibi açılardan test etmişlerdir. Bu amaçla PR-SE yöntemini içeren koşulda elde edilen ölçme kesinliği ve madde kullanımı istatistiklerini, hiçbir madde kullanım sıklığı kontrol yöntemi kullanılmayan koşulda, randomesque yöntemi kullanılan koşulda ve Sympson–Hetter (SH) yöntemlerinin kullanıldığı koşulda hesaplanan istatistikler ile karşılaştırmıştır. İkili puanlanan, madde ve birey parametrelerinin 3PLM'ye göre kestirildiği ilki 540 maddeden oluşan ikincisi de 300 maddeden oluşan büyük ve küçük iki madde havuzu oluşturmuşlardır. Hem büyük hem de küçük madde havuzundaki maddeler kullanılarak farklı simülasyon çalışmaları yapılmıştır. PR-SE yöntemi her iki havuzda da oldukça iyi ölçme kesinliği değerleri üretmiş ve aynı zamanda madde havuzundaki neredeyse tüm maddeleri kullanmıştır. Her ne kadar önemli düzeyde bir fark olmasa da randomesque yöntemi, PR-SE ve SH yöntemlerinden daha iyi ölçme kesinliği değerleri üretmiştir. Fakat randomesque yönteminin kullanıldığı koşulda hiç kullanılmayan madde sayısının oranı büyük havuzda %52.4, küçük havuzda ise %30.1 olmuştur. Madde havuzunun dengeli kullanımı mümkün olmamış, madde kullanım sıklığı oranları çarpık bir dağılım göstermiştir.

Moyer vd. (2012) ise sabit uzunluklu BBT çalışmalarında madde seçimi aşamasında bazı kısıtlama yöntemlerinin ölçme kesinliği açısından gösterdikleri performansları araştırmıştır. Kısıtlama yöntemi olarak Kısıtlanmış BBT (KBBT; Constrained CAT: CCAT; Kingsbury & Zara, 1989), Modifiye Edilmiş KBBT (MKBBT; Modified CCAT: MCCAT; Leung ve diğerleri, 2003) yönteminden türetilmiş Esnek MKBBT (EMKBBT) ve Ağırlıklandırılmış Ceza Modeli (ACM; Weighted Penalty Model: WPM; Shin, 2017; Shin ve diğerleri, 2009) seçilmiştir. Yöntemler arası karşılaştırma yedinci sınıf öğrencilerinin başarılarını ölçen gerçek bir merkezi sınavın verileri kullanılarak gerçekleştirilen simülasyonlar ile yapılmıştır. Araştırmacılar simülasyon koşullarına test uzunluğunu ve madde kullanım sıklığı kontrol yöntemi olarak da randomesque yöntemini eklemişlerdir. randomesque değeri bu çalışmada 10 olarak belirlenmiş, diğer koşul ise randomesque değerinin 1 olduğu koşul olmuştur. Yani madde kullanım sıklığı kontrol edilmemiştir. Kısıtlama olarak içerik, madde türü, doğru cevabın olduğu şık değişkenleri tanımlanmıştır. Yöntemlerin bu kısıtlamaları ihlal etmemede ne düzeyde başarılı oldukları araştırılmıştır. KBBT tüm koşullarda hiçbir kısıtlama ihlali göstermeyerek en başarılı yöntem olmuştur. randomesque yönteminin dahil olduğu koşullarda ölçme kesinliğinde azalma gerçekleşmiştir. Madde kullanım sıklığını kontrol etmede ise randomesque yöntemi oldukça başarılı olmuştur. Madde kullanım sıklığının kontrol edilmediği koşullarda madde kullanım sıklığı düşük olan maddelerin sayısı oldukça fazla olurken, randomesque yönteminin kullanıldığı koşullarda bu sayı oldukça azalmıştır.

Cheng vd. (2017) randomesque yöntemini daha farklı bir alanda test etmişlerdir. Araştırmacılar çalışmalarında geleneksel BBT çalışmalarının belli bir test uzunluğunda maksimum bilgiyi elde etmeyi amaçlandığını yazmışlardır. Yani BBT çalışmalarında asıl amaç lineer testlere kıyasla daha az madde ile yüksek ölçme kesinliği sağlamaktır. Fakat bazı yüksek riskli sınavlarda testin uzunluğu gibi test için ayrılan zaman da önemli olmaktadır. Bazı maddeler öğrencinin yeteneği hakkında oldukça fazla bilgi verebilir fakat aynı maddeyi çözmek için çok vakit harcamak da gerekebilir. Bu sebeple testin tamamlanması için gereken zaman istenenden daha fazla olabilir. Bu sorunu ele alabilmek

için Fan vd. (2012), Birim Zamanda Maksimum Bilgi (BZMB; Maximum Information Per Time Unit Method: MIT) yöntemini önermişlerdir. Cheng ve ark. (2017) da, bu çalışmada, herhangi bir cevap süresi modeline uyum gerektirmeyen, MIT yönteminin basitleştirilmiş bir formunu (MIT-S: Simplified MIT) önermişlerdir. Ölçme kesinliği, test zamanında tasarruf ve madde havuzu kullanımı açısından geleneksel MFI yöntemi ve MIT-S yöntemlerini 1PLM, 2PLM ve 3PLM'ye göre farklı koşullarda gerçekleştirilen simülasyonlar ile karşılaştırmıştır. Tüm koşullara test uzunluğu da iki değer (uzun: 40 madde, kısa: 20 madde) alacak şekilde eklenmiştir. 1PLM ile yapılan simülasyon çalışmasında ise, ayırt edicilik parametresinin tüm maddeler için aynı olması ve birim zamanda maksimum bilgiyi elde etmeye dayalı MIT-S yönteminin test zamanında tasarruf ettirecek maddeleri öncelikli seçip çarpık dağılan bir madde kullanım sıklığına neden olabileceği düşünülerek, randomesque yöntemi ve aşamalı yöntem (PR: Progressive Method) de madde kullanım sıklığı kontrol etmek amacıyla ayrı birer koşul olarak eklenmiştir. 3PLM ve 2PLM'yi içeren koşullarda MIT-S yöntemi MFI yöntemine yakın ölçme kesinliği değerleri üretmiş ve bunu da daha kısa test zamanları ile başarmıştır. Fakat MFI yöntemi gibi madde havuzunu dengeli kullanamamış ve hiç kullanılmayan madde sayısı oldukça fazla olmuştur. 1PLM ile yapılan simülasyonlarda ise MFI ve MIT-S yöntemlerinin kullanıldığı koşulların yanında, randomesque değerinin 5 olduğu bir koşul (MIT-S-R5) ve aşamalı yöntemin eklendiği bir koşul daha (MIT-S-PR) ilave edilmiştir. randomesque yönteminin kullanıldığı koşullarda, hem kısa hem de uzun test formatında ölçme kesinliğinden ödün vermeden madde havuzunun dengeli kullanımı sağlanmış, hiç kullanılmayan ya da çok az kullanılan maddelerin sayısı azalmış ve MIT-S kadar olmasa da test süresinde azalma gerçekleşmiştir.

İlgili araştırmalar incelendiğinde görülmektedir ki, İngilizce kelime bilgisinin ölçülmesi amacıyla yapılan BBT çalışmalarında madde havuzunu oluşturan maddelerin içerdiği kelimeler güçlük parametresi olarak sınırlı kalmıştır. Daha çok İngilizce metinlerde kullanımı noktasından daha düşük frekanslara sahip kelimeler kullanılmıştır. Bu çalışmada ise madde havuzu b parametresi olarak daha geniş ranjda değerleri içerecek şekilde oluşturulmuştur. Ayrıca şans başarısının VST'yi cevaplamadaki etkisine 3PLM ile bakmaktan ziyade,

genelde seçenekler üzerinde değişiklikler yaparak bakıldığı gözlemlenmiştir. Bu çalışmada ise 3PLM'nin model uyumuna bakarak şans başarısının VST maddelerini cevaplamada etkisinin ihmal edilebilir düzeyde olup olmadığı araştırılmıştır. Madde ve birey parametrelerinin kestirimi de önceki çalışmalardan farklı olarak yine 3PLM ile yapılmıştır. BBT uygulamaları ile İngilizce kelime bilgisinin ölçülmesinin çalışıldığı araştırmalarda uygun BBT koşullarının belirlenmesi amacıyla herhangi bir simülasyon çalışması yapılmadığı da ayrıca dikkat çekicidir. Bu çalışmada en uygun BBT kuralları post-hoc simülasyonlar ile belirlenmiştir. Ayrıca madde kullanım sıklığı kontrol yöntemi olarak randomesque yönteminin etkinliğinin araştırıldığı araştırmalarda birey yetenek parametreleri gerçek bireylere ait cevap örüntülerinden kestirilmemiş, simüle edilmiştir. Yani post-hoc simülasyonlar ile, hatta gerçek zamanlı bir BBT uygulaması ile randomesque yönteminin madde kullanım sıklığını kontrol etmedeki ve madde havuzunun dengeli kullanımındaki başarısı şu ana kadar araştırılmamıştır. Bu çalışmada bu noktadan eksiklikler de giderilmiştir.

Bölüm 3

Yöntem

Bu bölümde öncelikle araştırmanın türü, çalışma grubu, veri toplama süreci, veri toplama araçları ve verilerin analizi hakkında bilgiler verilmiştir. Sonra, bulgulara geçmeden önce, araştırmaya katılan öğrenciler hakkında betimleyici istatistikler paylaşılmıştır.

Araştırmanın Türü

Bu çalışmanın nicel araştırmalar başlığı altında farklı araştırma türlerine örnek teşkil edebilecek aşamaları vardır. VST ile gerçekleştirilen öğrencilerin İngilizce kelime bilgisini ölçme ve VST maddelerinin MTK'ya göre kalibre edilme işlemleri *tarama araştırmaları* kapsamına girmektedir. Tarama araştırmalarında bir grubun belirli özellikleri belirlenmeye çalışılır. Bu çalışmada öğrencilerin İngilizce kelime bilgileri ile İngilizce başarıları arasındaki korelasyonlara bakılmış ve kâğıt-kalem formatındaki VST'den elde edilen puanlar ile BBT uygulamasından elde edilen puanlar arasındaki ilişkinin düzeyi de incelenmiştir. Bu yönüyle çalışmanın korelasyonel bir araştırma olduğu da söylenebilir (Fraenkel ve diğerleri, 2011).

Araştırmanın Çalışma Grubu

Bu çalışmada iki farklı araştırma grubundan veri toplanmıştır. Öncelikle VST'nin geçerlik analizlerini yapabilmek, VST maddelerini MTK'ya göre kalibre edebilmek ve post-hoc simülasyon çalışmasında kullanabilmek amacıyla pandemi nedeniyle online form ile 1622 gönüllü öğrenciden veri toplanmıştır. Toplanan ilk veri setinde katılımcılardan 165'i sadece ilk birkaç soruyu yanıtladıktan sonra testi tamamladığından, bu 165 katılımcıya ait veriler çıkarılmış ve geriye kalan 1457 öğrenciye ait bilgiler Tablo 5 ve Tablo 6'da sunulmuştur. Tablo 5'e göre, ankete katılanların 823'ü kadın, 634'ü ise erkek öğrencilerden oluşmaktadır. Çalışma grubunda 49 hazırlık, 690 lisans, 181 yüksek lisans ve 537 doktora öğrencisi bulunmaktadır.

Tablo 5

Cinsiyete Göre Öğrencilerin Eğitim Düzeyleri

	Hazırlık	Lisans	Y. lisans	Doktora	Toplam
Kadın	27	395	93	308	823
Erkek	22	295	88	229	634
Toplam	49	690	181	537	1457

Tablo 6

Cinsiyete Göre Puanlar

	En düşük	En yüksek	Ortalama	Standart hata	Çarpıklık	Basıklık
Kadın	2	135	68.51	26.65	- 0.08	- 0.42
Erkek	5	136	70.50	27.97	- 0.09	- 0.63
Toplam	2	136	69.38	27.24	- 0.08	- 0.52

Tablo 6'ya göre, kadınların ortalama puanı 68.51, erkeklerin ise 70.5'tir. Tüm grubun ortalama puanı ise 69.38 olarak hesaplanmıştır. Çarpıklık ve basıklık değerlerini incelediğimizde, test puanı dağılımının normal dağılımdan çok fazla uzaklaşmadığı görülmektedir. İlk çalışma grubundan elde edilen veriler kullanılarak varyans ve kopya analizleri gerçekleştirilmiştir. Bulgular bölümünün ilk kısmında bu analizlere ait bulgulara yer verilmiştir.

İkinci olarak ise VST'nin BBT versiyonu ile kâğıt-kalem sonuçlarını karşılaştırmak için veri toplanmıştır. Öncelikle Hacettepe Üniversitesinde Eğitim Fakültesinde görev yapmakta olan ve lisansüstü eğitimi devam eden, çalışmaya gönüllü olarak katılmayı kabul eden 35 araştırma görevlisinden veri toplanmıştır. Lisansüstü eğitim gören öğrencilerin İngilizce kelime bilgisinin yüksek olabileceği ihtimaline binaen bir devlet üniversitesinde uygulamalı bilimler fakültesinde birinci sınıfta öğrenim gören 25 lisans düzeyinde üniversite öğrencisinden daha veri toplanmıştır. Böylece toplam 60 öğrenciden VST'nin kâğıt-kalem ve BBT versiyonları ile veri toplanmıştır.

Gerçek BBT uygulamasında öğrencilerden sadece İngilizce kelime bilgisi verisi toplanmıştır. Dört öğrencinin BBT versiyonunu tamamlamasına rağmen kâğıt-kalem formunu tamamlamadığı ve ilk birkaç soruyu yanıtladıktan sonra diğer sorulara cevap vermediği gözlemlenmiştir. Bu sebeple, bu öğrencilere ait veriler çıkarılarak BBT uygulamasına ait sonuçların analizi 56 kişinin verileri ile yapılmıştır. Aşağıda BBT uygulamasına katılan öğrencilerin VST ve BBT-VST versiyonlarına verdikleri cevaplar kullanılarak kestirilen yetenek değerlerine ait betimleyici istatistikler ve grafikler verilmiştir.

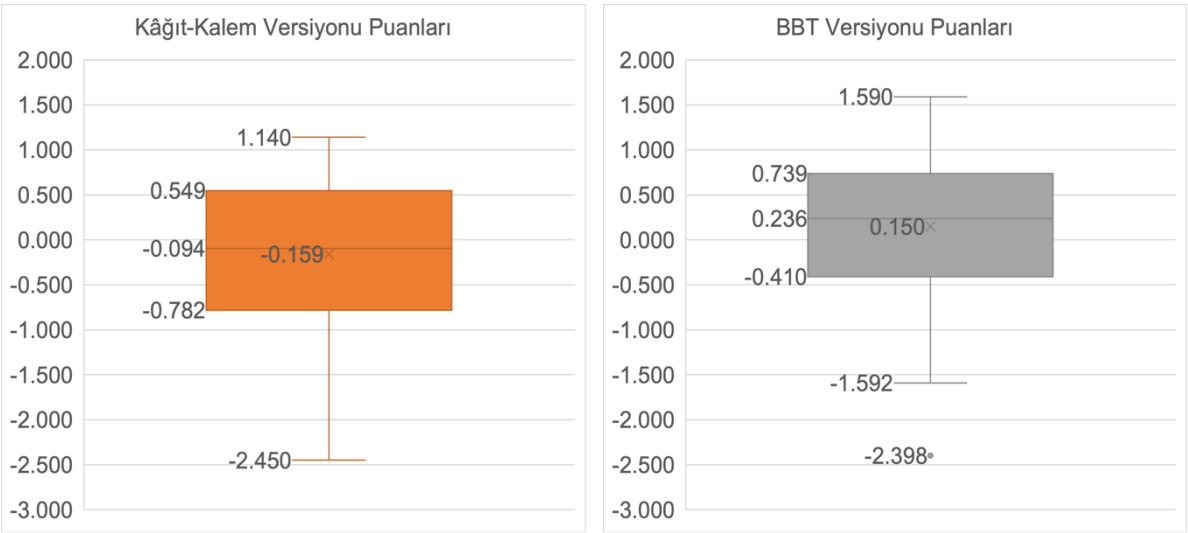
Tablo 7

VST ve BBT-VST Puanlarına Ait Betimleyici İstatistikler

	En düşük	En yüksek	Ortalama	Standart hata	Çarpıklık	Basıklık
VST	-2.45	1.14	-0.16	0.84	-0.56	2.75
BBT-VST	-2.40	1.59	0.15	0.85	- 0.59	3.13

Şekil 6

Kâğıt-kalem ve BBT Versiyonlarından Elde Edilen Puanlara Ait Kutu-Bıyık Grafikleri



Tablo 7'de BBT uygulamasına katılan öğrencilerin VST'nin hem kâğıt-kalem versiyonundan hem de BBT versiyonundan aldıkları puanlara ait en düşük ve en yüksek değerler, ortalama ve standart hata değerleri ile her iki versiyona ait çarpıklık ve basıklık değerleri verilmiştir. Her iki test için de önemli bir çarpık dağılımdan bahsedilemez. Fakat her iki versiyondan elde edilen puanların basıklık katsayısına bakıldığında puanların biraz

sivri bir dağılım gösterdiği söylenebilir. Şekil 6'da kâğıt-kalem ve BBT versiyonlarından elde edilen puanlara ait kutu-bıyık grafikleri verilmiştir. Bir kutunun üst kısmı üçüncü çeyrek dilimi, alt kısmı birinci çeyrek dilimi, x'in yanındaki sayı ortalama puanı ve çizgi ise medyanı göstermektedir. EK-A'da VST'nin kâğıt-kalem formundan elde edilen yetenek değerlerine ait histogram ve yoğunluk grafikleri verilmiştir. EK-B'de ise BBT-VST uygulamasından kestirilen yetenek puanlarına ait histogram ve yoğunluk grafikleri verilmiştir. Her iki versiyona ait puanlarının sahip olduğu sivri dağılım bu grafiklerden de gözlemlenmektedir. Kâğıt-kalem versiyonunda puanların yarıdan fazlası -0.2 ile 1.0 yetenek düzeyleri arasında yer alırken, BBT versiyonunda ise puanların yarıdan fazlası 0 ile 1.1 yetenek düzeyleri arasında bulunmaktadır.

Veri Toplama Süreci

Bu çalışma için Hacettepe Üniversitesi Etik Kurulundan gerekli izinler alındıktan sonra veriler ilk önce VST'nin 140 maddelik versiyonu aracılığıyla toplanmıştır. Bu versiyona aşağıdaki bağlantıdan ulaşılabilir: (https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size- test/Vocabulary-Size-Test-14000.pdf). VST'nin çevrimiçi versiyonu Google Forms kullanılarak oluşturulmuştur. Online VST versiyonunun linki lisans, yüksek lisans ve doktora öğrencilerine e-posta yoluyla gönderilmiştir. Çevrimiçi formda, test öncesinde VST hakkında bazı bilgiler verilmiştir. Ayrıca, bilmedikleri kelimeleri içeren soruları atlayabilecekleri ve herhangi bir zaman sınırlaması olmadığı belirtilmiştir. Veri toplama aracında ayrıca katılımcıların eğitim düzeyi, cinsiyeti ve İngilizce yeterlilik sınavı puanı (TOEFL veya Yabancı Dil Bilgisi Seviye Tespit Sınavı) gibi demografik özelliklerine ilişkin sorular da yer almıştır.

Kâğıt-kalem ve BBT versiyonlarını karşılaştırmak amacıyla gerçekleştirilen veri toplama işlemi için Cambridge Üniversitesi Psikometri Merkezi tarafından geliştirilen Concerto platformu kullanılmıştır. Concerto platformu Amazon Web Services (AWS) üzerine kurulmuş ve bu platform üzerinde VST'nin BBT versiyonu (BBT-VST) geliştirilmiştir.

Öğrenciler ya kâğıt-kalem testinden önce ya da sonra BBT-VST'yi AWS üzerinden alınan linke tıklayarak ulaşıp çözmüşlerdir. Tablet, bilgisayar ya da telefon ile testi tamamlamak mümkündür, tercih öğrencilere bırakılmıştır.

Veri Toplama Araçları

VST

Vocabulary Size Test (VST), Nation ve Beglar tarafından 2007 yılında İngilizce kelime bilgisini ölçmek amacıyla geliştirilmiştir. VST, British National Corpus'a (BNC) göre en sık kullanılan 14.000 kelimeden 140 kelime seçilerek oluşturulmuştur. İlk olarak, 14.000 kelime 14 seviyeye ayrılmış (her seviyede 1000 kelime) ve daha sonra her seviyeden 10 kelimelik bir örneklem seçilmiştir. BNC'deki kelimeler İngilizce metinlerdeki kullanım sıklığına göre sıralanmıştır. Bir kelimenin listedeki sırası arttıkça kullanım sıklığı azalmaktadır. Böylece, VST'nin 14 seviyesi arasında, ilk seviyedeki maddelerin sonraki seviyelerdekilere göre daha kolay olması öngörülmüştür. VST maddeleri çoktan seçmeli formatta sunulmaktadır. Kelime bilgisi dışında herhangi bir değişkenin sınav katılımcılarının yanıtlarını etkilememesi için soru kökleri kısa tutulmuştur. Aşağıda birinci seviyeden örnek bir madde verilmiştir. Ayrıca EK-C'de VST'nin ilk sayfası örnek olarak verilmiştir.

4. FIGURE: Is this the right **figure**?

a. answer

b. place

c. time

d. number

VST'nin Arapça, Farsça Gujarati, Japonca, Korece, Mandarin, Rusça, Tamilce, Tayca ve Vietnamca dillerinde iki dilli versiyonları da bulunmaktadır. Bu iki dilli versiyonlardan bazılarının geliştirme ve doğrulama çalışmaları da yüksek etki faktörüne sahip çeşitli dergilerde yayınlanmıştır (Elgort, 2012; Karami, 2012; Nguyen & Nation, 2011;

Zhao & Ji; 2018). Bu iki dilli versiyonlardan iki örnek madde aşağıda sunulmuştur. Maddeler incelendiğinde, iki dilli versiyonların aynı soru köklerine sahip olduğu, ancak seçeneklerin katılımcıların ana dilinde olduğu görülmektedir.

Rusça Versiyonu (Elgort, 2012)

4. FIGURE: Is this the right **figure**?

a. ответ

b. место

c. время

d. цифра

Vietnamca Versiyonu (Nguyen ve Nation, 2011)

4. FIGURE: Is this the right **figure**?

a. câu trả lời

b. địa điểm

c. thời gian

d. con số

BBT-VST

BBT-VST, VST'nin çalışma kapsamında geliştirilen bireyselleştirilmiş bilgisayarlı test versiyonudur ve Cambridge Üniversitesi Psikometri Merkezi tarafından geliştirilen Concerto platformu kullanılarak geliştirilmiştir. Concerto açık kaynaklı bir yazılımdır. Arka planda HTML, R ve MySQL'i de işe koşarak kolaylıkla BBT uygulamaları geliştirmeye olanak sağlar (Aybek, 2016).

Concerto'da oluşturulan testler, bilgisayarlardan, akıllı telefonlardan ve elektronik tabletlerden erişilebilen son derece uyarlanabilir önyüz (front-end) kullanıcı ara yüzleri aracılığıyla uygulanabilmektedir. Bu ara yüzler HTML'nin yanında JavaScript ve CSS de kullanılarak web sitelerine benzer şekilde oluşturulmaktadır. Concerto platformu, kendi kodlarını yazmak istemeyen kullanıcılar için dahili stok şablonlarıyla birlikte gelmektedir. Bu kullanıcı ara yüzü, R programlama dilini kullanarak puanlama ve BBT algoritmalarını

içerebilen arkayüz (back-end) işlevleriyle etkileşime girer (Harrison ve diğerleri, 2020). Concerto'da oluşturulan BBT uygulamalarında birçok BBT kuralı kolaylıkla tanımlanabilir. Madde seçme yöntemi, yetenek kestirim yöntemi ve sonlandırma kuralı (hataya göre ya da sabit test uzunluğu) gibi BBT koşulları herhangi bir kod yazmaya gerek kalmadan Concerto ara yüzü ile rahatlıkla belirlenebilmektedir. Ayrıca içerik dengeleme ve madde kullanım sıklığını kontrol etme işlemleri de yapılabilmektedir. Madde bazında ya da tüm test için süre tanımlanabilmekte, uygulanacak en az ya da en çok madde sayıları da birer koşul olarak teste eklenebilmektedir. Concerto ile madde havuzu sadece BBT olarak değil geleneksel lineer test olarak da uygulanabilmektedir. Hatta testin hem lineer formatı hem de BBT formatı aynı test akışına (test flow) dahil edilip, arka arkaya uygulanabilmekte ve her iki testin sonucu veri tabanına ortak kimlik bilgisi (ID) ile kaydedilebilmektedir. Böylece iki format arasında eşleştirme için herhangi bir ekstra işlem yapmaya gerek kalmamaktadır. Veri tabanına kaydedilen veriler de csv ya da pdf formatında indirilebilmektedir. Kaydedilen verilerin içinde bireyin cevapladığı test soruları, anlık olarak kestirilen yetenek ve hata değerleri ve hatta test sorularını ne kadar sürede cevapladığı da yer almaktadır. Concerto bahsedilen olanaklardan çok daha fazlasını kullanıcılarına sunmaktadır. Çalışmalarını bu platformu kullanarak yapmak isteyen araştırmacılar platformun web sitesini (https://concertoplatform.com/about) ziyaret edebilirler ya da yine platforma ait GitHub sayfasındaki (https://github.com/campsych/concerto-platform/wiki) belgelerden yararlanarak ayrıntılı bilgi elde edebilirler.

Verilerin Analizi

Çalışmanın ilk aşamasında öncelikle VST'nin geçerliğine dair kanıt toplamak için gerekli analizler gerçekleştirilmiştir. Geçerlik çalışması için "Standards for Educational and Psychological Testing"de (AERA, APA & NCME, 2014) sunulan öneriler takip edilmiştir. *Standartlar*'da belirtildiği üzere, test puanlarının belirli bir kullanım için amaçlanan yorumunun geçerliğini değerlendirmek için kullanılabilecek çeşitli geçerlik kanıtı kaynakları vardır. Farklı koşullarda, bu kaynakların çeşitli kombinasyonları kullanılabilir. Her bir

kaynağın tüm geçerlik süreçlerinde kullanılmasına dair bir gereklilik yoktur. *Standartlar* tarafından listelenen kanıt kaynakları içerik, diğer değişkenlerle ilişki (uyum geçerliği), iç yapı, yanıt süreçleri ve testin sonuçlarıdır. Bu çalışmada, testin sonuçları haricinde diğer tüm kaynaklar için kanıt toplanmıştır.

Yapı Geçerliğine İlişkin Kanıtlar

Bu çalışmada yapı geçerliği ile ilgili olarak, 3PLM temelli MTK ve DMF analizleri gerçekleştirilmiştir. Yapı geçerliğine dair kanıtlar elde etmek için MTK analizleri R yazılımı kullanılarak "mirt" paketi ile gerçekleştirilmiştir (Chalmers, 2012). Verilerin en iyi uyum sağladığı model, aynı paketteki ANOVA fonksiyonu ile belirlenmiştir. VST'nin MTK analizlerinden önce MTK'nın tek boyutluluk ve yerel bağımsızlık varsayımları test edilmiştir.

Boyutluluğun belirlenmesinde Kaiser kuralı (Kaiser, 1960), yamaç-birikinti grafiği (Cattel, 1966) ve paralel analiz (Horn, 1965) olmak üzere alan yazında oldukça yaygın olarak kullanılan üç tane yöntem vardır (Cho ve diğerleri, 2009). Weng ve Cheng (2005), ikili puanlanan maddeler ile yapılan boyutluluk analizlerinde, her ne kadar anlamsız boyutlar elde etme riski olsa da paralel analizin başarılı sonuçlar verdiğini göstermişlerdir. Fakat Tran ve Forman (2009), ikili puanlanan maddeler ile çalışıldığında ve Pearson korelasyonu kullanıldığında paralel analizin güvenirliğini oldukça düşük bulmuşlardır. Tetrakorik korelasyon kullanılsa da önemli bir iyileşme gözlemlenmemiştir. Bu sebeple VST'nin tek boyutluluk analizini yaparken paralel analiz tercih edilmemiştir. Yamaç grafiğine ve özdeğerlere bakılıp baskın bir boyut olup olmadığı gözlemlenmiştir.

Açımlayıcı faktör analizi (AFA) gerçekleştirilmiş ve kestirim yöntemi olarak WLSMV (weighted least squares means and variance adjusted - ortalama ve varyansa göre düzeltilmiş ağırlıklandırılmış en küçük kareler) yöntemi seçilmiştir. WLSMV faktör çıkarımı için tetrakorik korelasyonu kullanmaktadır. Faktör analizi sürekli değişkenlerle yapıldığında ve veriler tek değişkenli ve çok değişkenli normallik varsayımını karşıladığında, maksimum olabilirlik (ML) tahmin yöntemleri kullanılmalı, kategorik değişkenlerle yapıldığında ise en küçük kareler yöntemleri önerilmektedir (Koyuncu ve Kılıç, 2019). ML yöntemleriyle

karşılaştırıldığında, WLSMV'nin kategorik veya ikili veri içeren büyük modellerde istatistiksel performans ve analiz süresi açısından daha iyi olduğu (Muthen vd., 1997) ve hatta daha az yanlı kestirimlerde bulunduğu gösterilmiştir (Li, 2016).

Yerel bağımsızlık varsayımını test etmek için Yen'in (1993) madde çiftleri arasındaki Q3 istatistiği hesaplanmıştır. Q3 istatistiği için bir kesme değeri belirlemek amacıyla De Ayala'nın (2009) önerileri takip edilmiştir. Olası bağımlı madde çiftlerini tespit etmek için 140x140'lık bir matris incelenmiştir.

Değişen madde fonksiyonu (DMF) analizleri cinsiyet değişkenine göre R'ın "difR" paketi (Magis ve diğerleri, 2010) aracılığıyla Lojistik regresyon yöntemi, Lord'un ki-kare testi ve Mantel-Haenszel yöntemi ile gerçekleştirilmiştir. difLogistic, difLord ve difMH fonksiyonları kullanılmış ve ardından dichoDif fonksiyonu çalıştırılarak her üç yöntem tarafından da DMF olarak işaretlenen maddelerin belirlenmesi için karşılaştırmalar yapılmıştır. Önemli düzeyde DMF gösteren maddelerin DMF istatistikleri bir tabloda verilmiştir. Ayrıca, bu önemli düzeyde DMF gösteren maddelerinin madde karakteristik eğrileri (MKE) R yazılımı kullanılarak çizilmiştir.

Kapsam Geçerliğine İlişkin Kanıtlar

VST maddeleri, orijinal versiyonun geliştiricileri tarafından İngilizce kelime derlemini (corpus) mümkün olduğunca temsil edecek şekilde yazılmıştır ve bu şekilde çalışmalarında kapsamla ilgili geçerlik kanıtı sağlamışlardır. Bu çalışmada ayrıca kapsam geçerliğinin diğer kaynakları da araştırılmış ve VST'de yeterli sayıda madde olup olmadığını ve IRT modelinin yetenek ölçeğine orantılı bir şekilde yayılıp yayılmadığını kontrol etmek için R'da bir birey-madde haritası (Wright haritası) oluşturulmuştur.

Uyum Geçerliğine İlişkin Kanıtlar

Diğer değişkenlerle ilişkilere (uyum geçerliği) dair kanıtlar için, VST puanları ile Test of English As a Foreign Language (TOEFL) ve Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS) olmak üzere iki İngilizce yeterlilik sınavından alınan puanlar arasındaki korelasyonlar incelenmiştir. Yanıt süreçlerine ilişkin kanıtlar, katılımcıların soruları test geliştiricilerinin amaçladığı şekilde yanıtlayıp yanıtlamadıklarını inceler. Bu, sesli düşünme süreçleri yoluyla

kanıt toplamayı gerektirse de bu çalışmada IRT modeline tahmin etkisini dahil ederek şans eseri doğru yanıt vermenin etkisini araştırmak için dolaylı olarak kanıt toplanmıştır.

Post-Hoc Simülasyon Çalışması

VST'nin geçerlik kanıtları toplandıktan sonra, VST'nin BBT uygulamasını geliştirme aşamasına geçilmiştir. Alan yazında BBT üzerine üç tür araştırma deseni vardır (Nydick & Weiss, 2009)

1. Monte-Carlo simülasyonlar
2. Post-Hoc simülasyonlar
3. Canlı BBT uygulaması

Canlı, gerçek bir BBT uygulamasını gerçekleştirebilmek için öncelikle BBT'nin başlangıcı, devamı, sonlanması ve puanlanması için en uygun koşulların araştırmacı tarafından belirlenmesi gerekmektedir. Deneme-yanılma yoluyla yapıldığında oldukça masraflı olması, zaman kaybına ve maddi kayıplara sebep vermesi nedeniyle, bu çalışmada BBT koşulları simülasyon çalışması ile belirlenmiştir (Barnard, 2018; Nydick & Weiss, 2009).

İkili puanlanmış maddelerle çalışıldığında Monte-Carlo simülasyonları genellikle dört aşamadan oluşur (Nydick & Weiss, 2009).

1. Belli bir dağılıma göre, simüle bireylerin yetenek (θ) değerlerini üretme

2. Belli bir dağılıma göre, madde parametreleri üretme ya da önceden kalibre edilmiş bir madde havuzunun parametrelerini kullanma

3. Bir MTK modeli seçip, ona göre cevap matrisi üretme

4. Önceden belirlenmiş BBT koşullarına göre, simüle bireylerin her bir maddeye verdikleri cevapları içeren matrisi kullanarak bir BBT uygulaması gerçekleştirme

Önceden de bahsedildiği gibi eğer bir BBT uygulaması geliştirip ileride gerçek bireyler ile çalışılacaksa, MC simülasyonlarından elde edilen bulguları ve bunların yorumlamalarını gerçek test uygulamalarına genellemek her zaman mümkün olmadığı için MC simülasyonları pek tercih edilmemelidir. Onun yerine gerçek veri kullanarak yapılan

post-hoc simülasyonları tercih etmek daha uygun olmaktadır (Sari, 2020; Thompson & Weiss, 2019).

catR paketi ile hem Monte-Carlo hem de Post-Hoc BBT simülasyonlarını gerçekleştirmek mümkündür. Bu çalışmada da *catR* paketi kullanılarak 72 farklı koşul test edilmiştir. Bu koşullardan dördü yetenek kestirimine aittir. Concerto platformunda BBT uygulaması geliştirildiğinde yetenek kestirim yöntemi olarak bize dört seçenek (BM, ML, WL ve EAP) sunmaktadır. Bunların hepsi simülasyon koşullarına dahil edilmiştir.

Madde seçim yöntemi olarak MFI ve MFI kadar seçici olmayan Urry kriteri (Urry, 1970) iki farklı koşul olarak simülasyonlara eklenmiştir. Urry yöntemi MFI gibi yüksek "a" parametresine sahip maddeleri kullanmamaktadır. Bu yöntemde bireyin anlık olarak kestirilen yetenek düzeyine en yakın "b" parametresine sahip madde seçilmektedir. catR paketinin son versiyonlarında bu yönteme *bOpt* adı verilmiştir. Bundan sonra bu çalışmada da bu yöntem *bOpt* olarak adlandırılacaktır.

Testi sonlandırma kuralı olarak üç farklı koşul belirlenmiştir. Bu koşulların tamamı standart hataya (SH) bağlı koşullardır ve .20, .25 ve .30 olarak seçilmiştir. Testi sonlandırma kuralı olarak .32 değerinin altında bir SH değerinin belirlenmesi .9'un üzerinde bir güvenirlik elde edilmesini sağlar (Walter, 2009). Bilindiği üzere standart hatası daha düşük yetenek kestirimleri yapabilmek için de daha fazla sayıda maddeye ihtiyaç duyulur. Post-hoc simülasyonlar öncesi bu çalışma kapasamında yapılan öncül deneme simülasyonlarında düşük SH değerleri sonlandırma kuralı olarak belirlendiğinde makul ortalama test uzunlukları elde edildiği görülmüştür. Bu sebeple simülasyon koşullarına, güvenirliği daha yüksek kestirimler elde edebilmek için, .20, .25 ve .30 gibi nispeten düşük SH değerleri testi sonlandırma koşulları olarak eklenmiştir. Sabit bir madde sayısına göre herhangi bir koşul belirlenmemiştir. Fakat alan yazından elde edilen bilgilere göre yetenek düzeyinin her iki ucundaki bireylerin yetenek parametrelerinin kestiriminde belli bir hata değerinin altına düşmesi için bazen çok sayıda maddeye ihtiyaç duyulmaktadır. Hatta bazen havuzun tamamı uygulansa dahi standart hata olarak belirlenen kesme değerinin altına düşmemektedir. Bu sebeple BBT uygulamasının çok uzamaması için her testi sonlandırma

koşuluna, havuzdaki madde sayısının dörtte biri olan 35 madde ayrı bir kural olarak eklenmiştir. Yani eğer gerçek uygulamada 35 madde uygulandığı halde önceden belirlenen hata değerinin altına düşülmediyse test orada sonlanacaktır.

Ayrıca MFI yönteminin çok seçici olması problemine yönelik olarak Kingsbury ve Zara (1989) tarafından önerilen *randomesque* yöntemi madde kullanım sıklığı kontrol yöntemi olarak üç simülasyon koşulunu daha içerecek şekilde çalışmaya eklenmiştir. Bu yöntemde yazılım bireye anlık yetenek düzeyinde en uygun (optimal) olan maddeyi uygulamaz. O yetenek düzeyi için en uygun birkaç madde belirler ve bu maddeler arasından rastgele bir maddeyi seçer ve bireye uygular. Seçilecek en uygun madde sayısı hem simülasyonlarda hem de gerçek BBT uygulaması aşamasında araştırmacı tarafından herhangi bir tam sayı olarak belirlenebilir. Bu çalışmadaki üç koşul şöyledir.

1-) randomesque = 1

2-) randomesque = 3

3-) randomesque = 5

Birinci koşulda herhangi bir madde kullanım sıklığı kontrol yöntemi yoktur. Yani hangi madde bireyin o andaki yetenek düzeyine en uygun ise yazılım o maddeyi uygulayacaktır. İkinci koşulda ise yazılım en uygun 3 madde belirleyecek ve bunların içinden bir tanesini rastgele seçip bireye uygulayacaktır. Üçüncü koşulda ise seçilecek en uygun madde sayısı 5'tir.

Simülasyondaki tüm koşullar aşağıda özetlenmiştir.

- 4 yetenek kestirim yöntemi
  - BM
  - ML
  - WL
  - EAP
- 2 madde seçimi kuralı

- MFI
- bOpt

➢ 3 sonlandırma kuralı

- SH ≤ .20
- SH ≤ .25
- SH ≤ .30

➢ 3 madde kullanım sıklığı kontrol yöntemi

- randomesque = 1
- randomesque = 3
- randomesque = 5

Simülasyondaki koşulların her birine bir isim verilmiştir ve hangi koşulda hangi yöntemlerin dahil edildiğinin anlaşılabilmesi için belli bir örüntü takip edilmiştir. Öncelikle o koşulda tercih edilen yetenek kestirim yönteminin ilk harfi (büyük harf), sonra madde seçim yönteminin ilk harfi, sonra koşuldaki standart hata kesme değerinin noktasız yazımı ve son olarak da eğer "randomesque = 1"den farklı bir madde kullanım sıklığı kontrol yöntemi kullanıldı ise, yönteme göre "r3" ya da "r5" eklenmiştir.

1. Örnek koşul adı: Em20r5 => E - m - 20 - r5

Örneğin yukarıdaki koşulda yetenek kestirim yöntemi EAP, madde seçim yöntemi MFI, standart hata için kesme değeri .20, kullanılan madde kullanım sıklığı yöntemi ise "randomesque = 5"tir. Eğer bir koşulda "randomesque = 1" ise, yani herhangi bir madde kullanım sıklığı kontrol yöntemi uygulanmamış ve yazılım bireyin anlık olarak kestirilen yetenek düzeyi için en uygun maddeyi uygulayacaksa, koşula isim verirken madde kullanım sıklığı kontrol yöntemi kullanılmamıştır. İkinci örnekte de görüleceği üzere koşulun adı sadece yetenek kestirim yöntemi (ML), madde seçim yöntemi (bOpt) ve sonlandırma kuralı (SH ≤ .30) kullanılarak oluşturulmuştur.

2. Örnek koşul adı: Mb30 => M - b - 30

Simülasyon çalışmasında öncelikle gerçek veriye ait madde ve birey parametreleri *mirt* paketi ile kestirilmiş ve kestirilen bu parametreler iki ayrı matris olarak kaydedilmiştir. Sonra *simulateRespondents* fonksiyonu kullanılarak 72 farklı koşulda simülasyon çalışması yapılmıştır. Koşulları karşılaştırırken ortalama test uzunluğu ve ölçme kesinliğinin göstergeleri olan RMSE ve yanlılık değerleri incelenmiştir. Simülasyonda kullanılan gerçek veri, VST'nin geçerlik kanıtlarını incelemesinde kullanılan veridir. Geçerlik ve simülasyon çalışmasına dair bulgular ile beraber gerçek BBT çalışmasına dair bulgular bir sonraki bölümde verilmiştir.

Bölüm 4

Bulgular

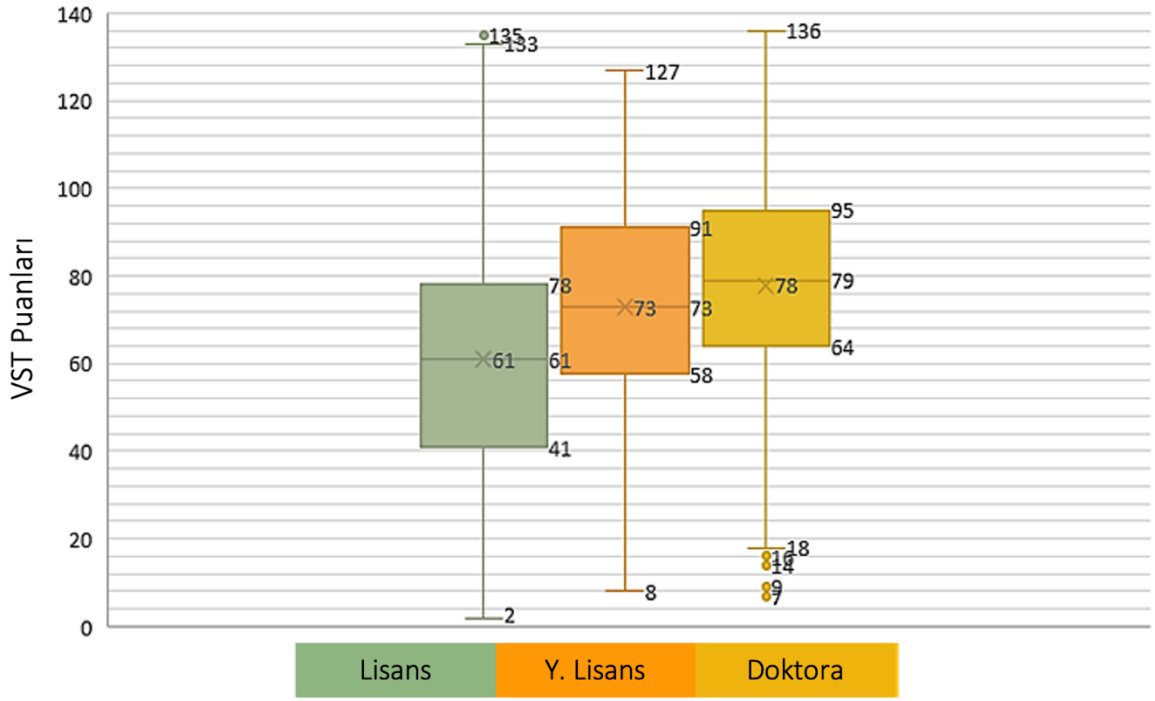
Bu bölümde öncelikle ilk çalışma grubundan elde edilen veriler üzerinden gerçekleştirilen varyans ve kopya analizlerine ait bulgulardan bahsedilmiştir. Sonrasında geçerlik, simülasyon ve gerçek zamanlı BBT uygulamasına ait bulgulara yer verilmiştir.

Varyans ve Kopya Analizlerine İlişkin Bulgular

Şekil 7'de katılımcıların VST puanlarının kutu ve bıyık grafiği verilmiştir. Bu grafikte hazırlık öğrencileri lisans grubuna dahil edilmiştir. Önceden de bahsedildiği üzere bir kutunun üst kısmı üçüncü çeyrek dilimi, alt kısmı birinci çeyrek dilimi, x'in yanındaki sayı ortalama puanı ve çizgi ise medyanı göstermektedir. Grafikten de görülebileceği gibi, katılımcıların eğitim seviyeleri arttıkça tüm puanlar (çeyrekler, ortalamalar ve medyanlar) da artmaktadır. Lisans, yüksek lisans ve doktora öğrencilerinin ortalamaları sırasıyla 61, 73 ve 78 olarak bulunmuştur.

Şekil 7

Öğrencilerin Eğitim Düzeylerine Göre Test Puanlarına Ait Kutu-Bıyık Grafikleri



Elde edilen ortalama VST puanlarının anlamlı düzeyde birbirinden farklı olup olmadıklarını test etmek için tek yönlü varyans analizi yapılmıştır. Elde edilen sonuçlara göre ortalama değerlerinden en az birinde anlamlı bir farklılık olduğu tespit edilmiştir, $F(2, 1454) = 59.90$, $p < .001$. Hangi eğitim seviyesine ait ortalamalar arasında anlamlı farkın olduğunu görebilmek için post-hoc test (Scheffe) yapılmıştır. Yüksek lisans ve doktora öğrencilerine ait ortalamalar arasında manidar fark bulunmamıştır, $p = 0.171$, %95 GA = (-9.75, 1.28). Fakat lisans ile yüksek lisans öğrencileri arasında ($p < 0.001$, %95 GA = [6.31, 16.96]) ve lisans ile doktora öğrencileri arasında ($p < 0.001$, %95 GA = [12.23, 19.51]) VST puan ortalamaları noktasında istatistiksel olarak önemli farklar bulunmuştur. Bu bulgu, VST'nin farklı eğitim seviyelerinden öğrencileri bir derece ayırt ettiğine dair bir kanıt olarak kabul edilebilir. Eğitim düzeyi, öğrencilerin İngilizce yeterliliklerini bir ölçüde yansıtmaktadır çünkü Türkiye'de araştırma görevlisi olmak ve lisansüstü programlarda öğrenim görmek için farklı düzeylerde İngilizce yeterliliği gerekmektedir. Lisansüstü programlar için talep edilen yeterlilik seviyesi lisans programlarından daha yüksektir. Bu sonuçlara dayanarak, VST'nin farklı İngilizce yeterlilik seviyelerindeki öğrencileri de ayırt edebildiği düşünülmektedir.

Geçerlik analizlerinden önce veri tarama ve temizleme işlemleri gerçekleştirilmiştir. Öncelikle kayıp veriler ve aralık dışı değerler kontrol edilmiştir. Boş bırakılan sorular yanlış cevap olarak kabul edilmiştir çünkü bir sorunun boş bırakılması, katılımcının maddede sorulan kelimenin manasını bilmediği anlamına gelmektedir. Veriler çevrimiçi formlar aracılığıyla toplandığından, herhangi bir kayıp değer ve aralık dışı değer bulunmamıştır ve buna bağlı olarak tek değişkenli uç değerler de yoktur. Çok değişkenli uç değerleri tespit etmek için her bir katılımcı için Mahalanobis uzaklığı hesaplanmıştır. Ayrıca, istatistiksel olarak anlamlı olup olmadığını görmek için her mesafe için p-değerleri hesaplanmıştır. 170 gözlemin p-değerinin .001'den küçük olduğu tespit edilmiş ve bu gözlemler çok değişkenli uç değer olarak değerlendirilmiştir. Bu gözlemler verilerden çıkarılmış ve geri kalan geçerlik ve simülasyon analizleri 1287 katılımcının verileriyle gerçekleştirilmiştir. Çok değişkenli uç değer olarak belirlenip veriden çıkarılan bireylerin özel bir grup olup olmadığı da ayrıca incelenmiştir. Bu amaçla gruplara ait frekanslar ve ortalama puanlara bakılmıştır. Grup

özelinde herhangi bir yığılma gözlenmemiştir. Çok değişkenli uç değerlerin gözlendiği bireylere ait istatistikler EK-Ç'de bir tablo ile özetlenmiştir. Toplam veride %12 oranında uç değer tespit edilmiştir. Gruplar özelinde bakıldığında, içinde en çok uç değer barından grup, %19 ile kadın lisans öğrencileridir. Erkek yüksek lisans öğrencilerine ait verilerde ise sadece %3 oranında uç değer tespit edilmiştir.

Çevrimiçi yapılan testlerde kopya gibi test hileleri ile daha sık karşılaşılır (Sanz ve diğerleri, 2020). Bu çalışmada da veriler çevrimiçi olarak toplanmıştır fakat katılımcılar kendi kelime seviyelerini öğrenmek için gönüllü olduklarından, çevrimiçi testte kopya çekmedikleri düşünülmektedir. Bu kanının ampirik olarak da desteklenip desteklenmediğini görmek için bir kopya analizi yapılmıştır. Kopya analizlerinde kopya eyleminde bulunması olası bireyleri tespit edebilmek için alanyazında genel olarak ya benzerlik indekslerine ya da birey uyum istatistiklerine bakılır (Zopluoglu, 2017). Bu çalışmada cevaplayıcılar testi bireysel olarak çözdükleri için benzerlik indekslerine değil de birey uyum istatistiklerine bakmak daha uygundur. Alan yazında birey uyumunu hesaplamak için geliştirilen oldukça fazla sayıda istatistik vardır. Hatta Karabatsos (2003) bu istatistiklerden 36 tanesini karşılaştıran kapsamlı bir araştırma yapmıştır.

Bu çalışmada birey uyum istatistiklerini elde etmek için H^T (Sijtsma, 1986; Sijtsma & Meijer, 1992) ve lz* (Snijders, 2001) değerleri R yazılımında "PerFit" paketi (Tendeiro ve diğerleri, 2016) kullanılarak hesaplanmıştır. Kopya analizlerinde birey uyum istatistikleri hesaplanarak normal olmayan cevap örüntülerine sahip bireylerin tespit edilmesi amaçlanır. Bu çalışmada tercih edilen H^T ve lz* istatistiklerinin normal olmayan örüntüleri tespitte başarılı oldukları görülmüştür (H^T için bkz. Karabatsos, 2003; lz* için bkz. de la Torre & Deng, 2008; Magis ve diğerleri, 2012). Kopya analizi yapılırken, bir önceki analizde çok değişkenli uç değer olarak değerlendirilen verilerin de dahil olduğu 1457 kişiden oluşan veri kullanılmıştır. H^T istatistiği için kesme değer olarak Karabatsos'un (2003) önerdiği .22 değeri kullanılmıştır. Bu değerin altında bir değere sahip bireyler normal olmayan cevap örüntüsüne sahip bireyler olarak işaretlenmiştir. lz* istatistiği için kesme değer ise "PerFit" paketindeki *cutoff* fonksiyonu ile – 3.49 olarak hesaplanmıştır. Bu değerin altındaki

bireylerin de cevap örüntüleri anormal olarak işaretlenmiştir. Kopya analizi sonucu elde edilen bulgular Tablo 8 ve şekil 8'de özetlenmiştir.

Tablo 8'de iki birey uyum istatistiğine göre belirlenen kesme değerlerin altında değerlere sahip olan bireylerin frekansları ve bulunduğu gruptaki toplam öğrenci sayısına göre oranları verilmiştir. Çok değişkenli uç değer olarak belirlenen bireylere ait uyum istatistikleri "Uç Değer" satırında, uç değer olarak belirlenmeyen bireylere ait uyum istatistikleri ise "Normal" satırında verilmiştir. En alt satırda ise toplam 1457 bireye ait istatistikler paylaşılmıştır.

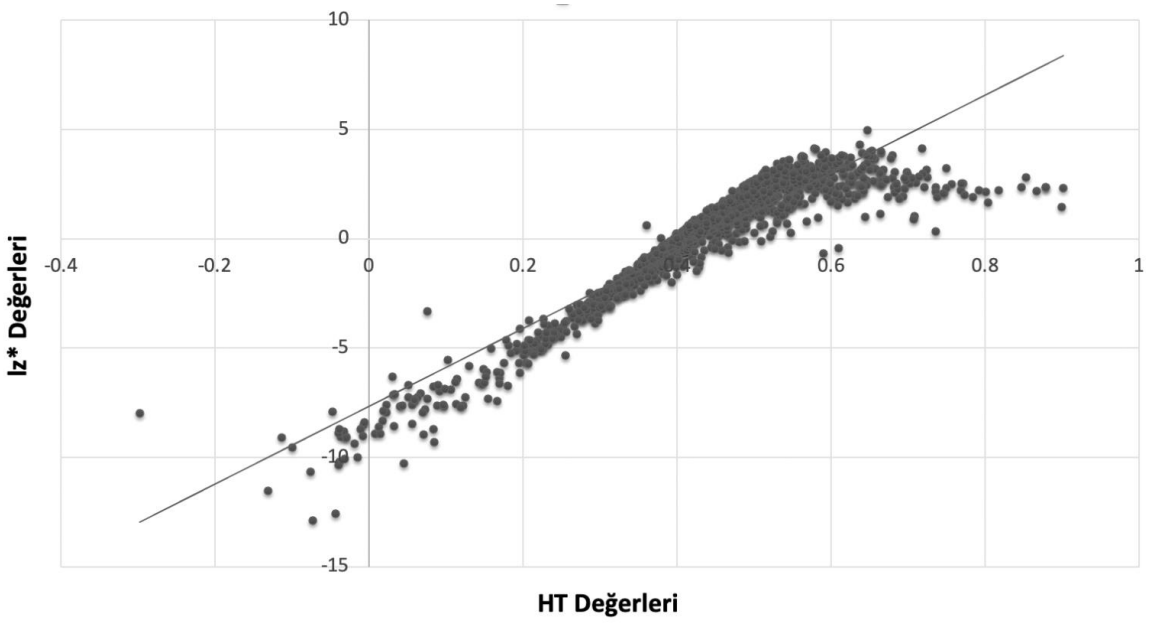
Tablo 8

Kopya Analizi Bulguları

	H^T Frekans	lz* Frekans	H^T Oran	lz* Oran
Uç Değer (n = 170)	92	111	54.11	65.29
Normal (n = 1287)l	17	41	1.32	3.18
Toplam (n = 1457)	109	152	7.48	10.43

Şekil 8

H^T ve lz Değerleri Arasındaki İlişki*



Tablodaki değerlere bakıldığında tüm öğrencilerde normal olmayan (anormal) cevap örüntüsüne sahip bireylerin toplam sayısı H^T istatistiğine göre 109, lz* istatistiğine göre ise 152'dir. Anormal cevap örüntüsüne sahip bireylerin toplam öğrenci sayısına oranı ise HT istatistiğine göre %7.48 ve lz* istatistiğine göre ise %10.43'tür. Fakat tablo ayrıntılı incelendiğinde görülmektedir ki, anormal cevap örüntüsüne sahip bireylerin çoğu uç değer olarak işaretlenip veri setinden çıkarılan bireylere aittir.

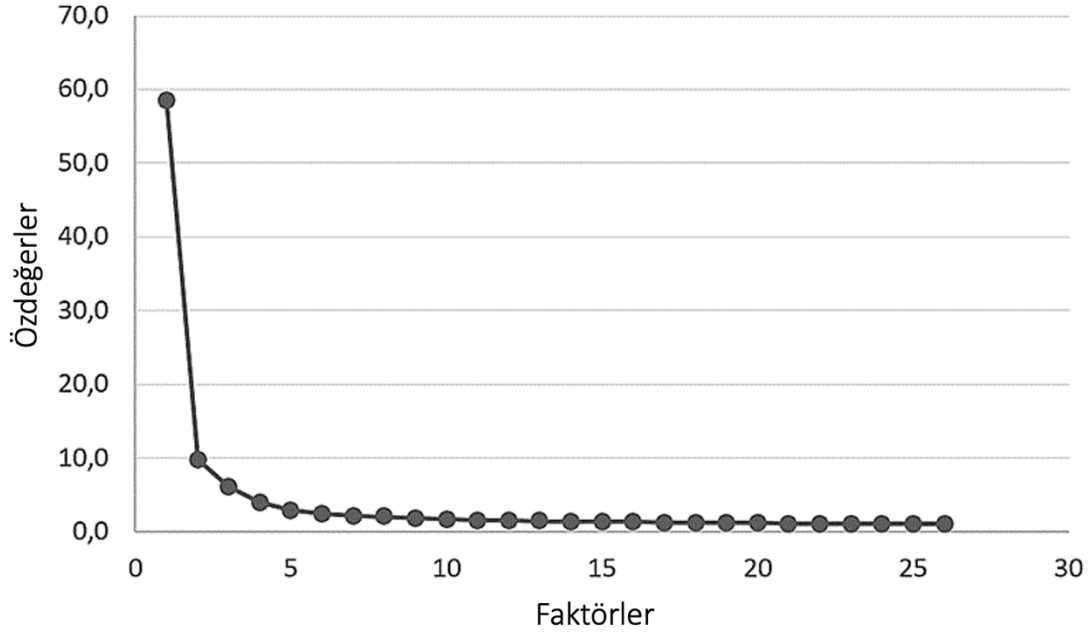
Bu çalışmada geçerlik ve simülasyon analizlerinin yapıldığı 1287 kişinin verilerinde HT istatistiğine göre sadece %1.32, lz* istatistiğine göre de sadece %3.18 oranında anormal cevap örüntüsü bulunmaktadır. Kopya çekme, anormal cevap örüntülerine neden olabilecek sebeplerden sadece bir tanesi olduğu ve bu istenmeyen örüntülere kaygı, motivasyon eksikliği, tahmin ile ya da rastgele cevaplama, dikkatsiz cevaplama ve yaratıcı cevaplama gibi sebeplerin de neden olabileceği (Gaertner & McBride, 2017; Zopluoglu, 2017) düşünüldüğünde bu çalışmada gerçekleştirilen çevrimiçi testte ciddi oranda kopya çekme gibi test hilesine başvurulmadığı söylenebilir. Şekil 8'de ise birey uyum değerlerini elde etmek için hesaplanan HT ve lz* değerleri arasındaki ilişki saçılım grafiği ile gösterilmiştir. İki istatistik arasındaki korelasyon katsayısı .927 (%95 GA = .919, .934) olarak hesaplanmıştır. Yüksek ilişkiden anlaşıldığı üzere, her iki istatistik normal olmayan cevap örüntüsüne sahip bireyleri işaretlemede birbirine yakın sonuçlar vermiştir.

Yapı Geçerliğine İlişkin Bulunan Kanıtlar - MTK

VST'nin MTK analizlerini yapmadan önce, MTK'nın iki temel varsayımı olan tek boyutluluk ve yerel bağımsızlık test edilmiştir. VST'nin boyutluluğunu araştırmak için açımlayıcı faktör analizi yapıldıktan sonra baskın bir boyut olduğu görülmüştür. Baskın bir boyut, MTK analizlerinde tek boyutluluk varsayımının karşılanması için yeterli görülmüştür (Hambleton & Swaminathan, 1985; Henning ve diğerleri, 1985). Şekil 9'da faktör analizinin özdeğerlerini gösteren yamaç-birikinti grafiği gösterilmektedir. Birinci boyutun özdeğerinin (58.602) ikinci boyutun özdeğerinden (9.784) yaklaşık altı kat daha büyük olduğu görülmektedir.

Şekil 9

Özdeğer ve Faktörlere ait Yamaç-birikinti Grafiği



Ayrıca, Tablo 9'daki model uyum indeksleri tek boyutlu modelin verilere iyi uyum sağladığını göstermektedir. İki, üç ve dört boyutlu modeller tek boyutlu modelden daha iyi uyum indekslerine sahip olsa da VST tek boyutlu olarak kabul edilmiştir. Tek boyutlu modelde maddeler toplam varyansın %42'sini açıklamaktadır. Maddelerin faktör yükleri 0.234 - 0.978 aralığındadır ve tüm maddelere ait faktör yükleri EK-D'de verilmiştir.

Tablo 9

Açımlayıcı Faktör Analizi Model Uyum İstatistikleri

Model	X^2	sd	X^2/sd	RMSEA	CFI	TLI	SRMR
1 Faktörlü	19167.1	9590	1.99	0.028	0.934	0.933	0.101
2 Faktörlü	14514.7	9451	1.53	0.020	0.965	0.964	0.074
3 Faktörlü	11842.6	9313	1.27	0.015	0.983	0.982	0.060
4 Faktörlü	11118.9	9176	1.21	0.013	0.987	0.986	0.055

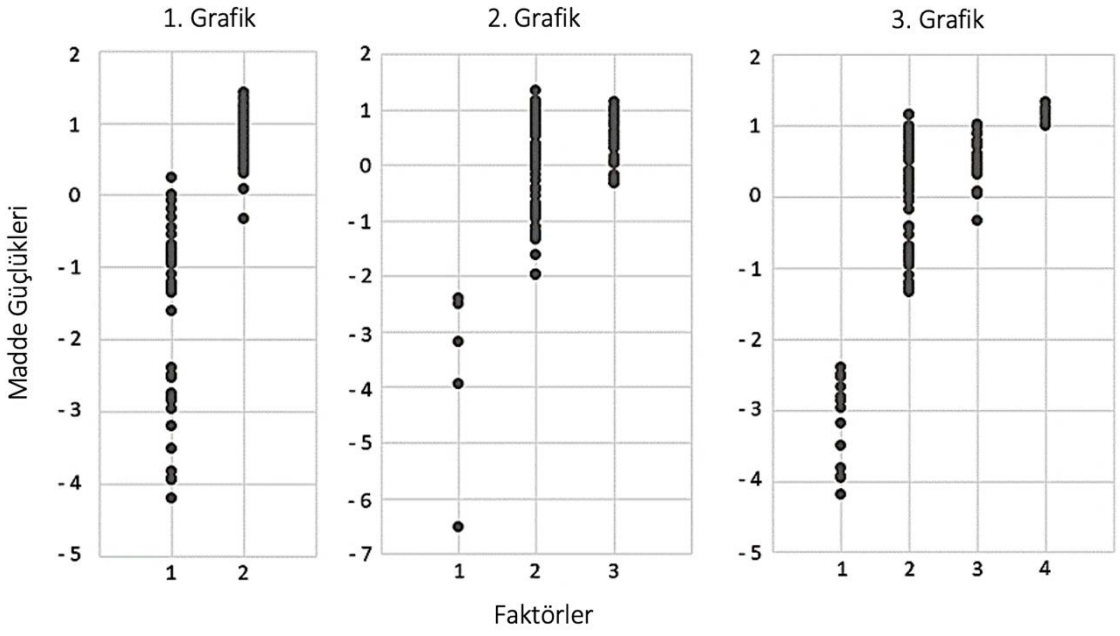
Çok boyutlu bulguların arkasındaki nedenin güçlük faktörleri olduğu düşünülmüştür (difficulty factors). "Güçlük faktörleri" sorununun geçmişi neredeyse bir asır öncesine dayanmaktadır (Hertzman, 1936; Spearman, 1927) ve ikili puanlanan maddelerin faktör analizlerinde sıklıkla karşılaşılmaktadır (daha fazla ayrıntı için bkz. Hattie, 1985). Bir testin

maddeleri güçlük parametresinde büyük ölçüde değişkenlik gösterdiğinde, madde içeriğinden bağımsız olarak madde güçlüğüne göre "sahte" faktörlerin çıkarıldığı bilinmektedir (McDonald & Ahlawat, 1974; Yang & Xia, 2015). Bu sorun genellikle "sahte/yapay faktörler" veya "güçlük faktörleri" olarak adlandırılır ve bazen kelime bilgisi gibi basit yapıların çok boyutlu görünmesine neden olur (Reckase ve diğerleri, 1988).

Şekil 10'da iki, üç ve dört boyutlu modellerin her biri ile ilgili boyutlardaki maddelerin güçlük parametreleri arasındaki ilişkileri gösteren saçılım grafikleri yer almaktadır. Birinci grafikte, birinci boyuttaki maddelerin güçlük parametrelerinin çoğunlukla -4 ile 0 arasında, ikinci boyuttakilerin ise 0 ile 1,5 arasında olduğu görülmektedir. İki boyutlu modelde, birinci boyut için madde güçlüklerinin ortalaması -1.57, ikinci boyut için ise 0.76'dır. İkinci ve 3. grafik incelendiğinde, birinci boyuta ait güçlük parametrelerinin aralıklarının diğer boyutlara ait olanlardan daha düşük değerlere sahip olduğu ve diğer boyutlar için aralıkların değerlerinin sırasıyla arttığı görülmektedir.

Şekil 10

İki, Üç ve Dört Boyutlu Modellerde Madde Güçlükleri ve Faktörler Arasındaki İlişkiyi Gösteren Saçılım Grafikleri



Üç boyutlu modelde, birinci boyut için madde güçlüklerinin ortalaması -3.70 olarak hesaplanmıştır. İkinci boyut için 0 ve üçüncü boyut için 0.45'tir. Dört boyutlu modelde ise birinci, ikinci, üçüncü ve dördüncü boyutlar için madde güçlüklerinin ortalamaları sırasıyla -3.18, 0, 0.54 ve 1.17 olarak bulunmuştur.

Tablo 10

Üç ve Dört Boyutlu Modele ait Post-hoc Analiz Bulguları

Model	Boyutlar	Fark	GA Alt Sınır	GA Üst Sınır	p değeri
Üç Boyutlu	1-2	3.70	2.78	4.62	< 0.001
	1-3	4.15	3.21	5.09	< 0.001
	2-3	0.45	0.05	0.86	0.026*
Dört Boyutlu	1-2	3.18	2.62	3.75	< 0.001
	1-3	3.72	3.10	4.34	< 0.001
	1-4	4.35	3.53	5.16	< 0.001
	2-3	0.54	0.10	0.98	0.010*
	2-4	1.17	0.47	1.85	< 0.001
	3-4	0.63	0.12	1.07	0.013*

* $p < 0.05$

Boyutlardaki madde güçlüklerine ait ortalamaların arasındaki farkların istatistiksel olarak manidarlığını test etmek için varyans analizi yapılmıştır. İki boyutlu modelde iki boyuta ait madde güçlükleri ortalamaları anlamlı düzeyde farklı bulunmuştur, $F(1, 96) = 188.9$, $p < 0.001$. Üç boyutlu modelde ise ilk yapılan analizde en az bir ortalamanın farklı olduğu bulunmuştur, $F(2, 103) = 59.89$, $p < 0.001$. Sonrasında yapılan post-hoc analizde (Scheffe), üç boyutun ortalamaları da birbirinden anlamlı derecede farklı bulunmuştur. Analiz sonuçları Tablo 10'da verilmiştir. Dört boyutlu modelde de öncelikle en az bir ortalamanın farklı olduğu bulunmuş, sonrasındaki post-hoc analizde (Scheffe) de yine dört boyuttaki maddelere ait güçlük parametrelerinin ortalamalarının manidar düzeyde farklılık gösterdiği bulunmuştur. Dört boyutlu modele ait bulgular da Tablo 10'da verilmiştir. Tablodan da görüleceği üzere güven aralığı (GA) değerlerine ait alt ve üst sınırları sıfırı içermemektedir. Elde edilen p değerlerinin hepsi en az %95 olasılıkla manidardır. Baskın bir boyut bulunması ve bu boyutun toplam varyansın %42'sini açıklaması gibi boyutsallığa

ilişkin bulgular ile madde içeriklerini incelendiğinde, tek boyutlu bir yapının söz konusu olduğu görülmektedir. Çok boyutlu görünümün bir nedeni, varyans analizi bulgularına göre güçlük faktörleri olarak açıklanabilir.

Madde artıkları (item residuals) arasındaki ilişkileri gösteren Q3 istatistiği, yerel bağımsızlık varsayımını araştırmak için hesaplanmıştır. Q3 korelasyonel bir istatistik olduğundan, değeri -1 ile +1 arasında değişir ve Q3'ün yüksek mutlak değeri yerel bağımsızlığın önemli ölçüde ihlal edildiğini gösterir (Paek & Cole, 2020). Q3 istatistiği için bir kesme değeri olarak de Ayala (2009) |Q3|≥ √0,5=.2236 değerini önermiştir. VST'de 140 madde bulunduğundan, olası bağımlı madde çiftlerini tespit etmek için 140x140'lık bir matris (19600 hücre) incelenmiştir. Q3 istatistiği .2236'nın üzerinde olan madde çiftleri yerel olarak bağımlı maddeler olarak işaretlenmiştir. Elde edilen 19600 hücreden 74 tanesinin Q3 değerinin kesme değerinden yüksek olduğu ve bu 74 hücrenin 30 farklı maddeye ait olduğu görülmüştür.

Tablo 11

Olası Yerel Bağımlı Madde Çiftleri

Madde	Maddeler	Madde	Maddeler	Madde	Maddeler	Madde	Maddeler
1	2,5,6,8	17	19,21	38	35		26,30,70,74,
2	1,5,6,8,19,21	19	2,17,21	42	34,43	103	94,105,107
5	1,2,6,8	21	2,6,17,19,22	43	34,42		70,94,103,
6	1,2,5,8,21	22	21	62	65	105	107,117
7	10	26	103	65	62	107	70,103,105
8	1,2,5,6	30	103	70	103,105,107	116	140
10	7,11	34	42,43	74	103	117	105
11	10	35	38	94	103,105	140	116

Tablo 11'de bu 30 madde, "Madde" sütunlarında ve bu 30 madde ile yüksek Q3 değerine sahip olan maddeler ise "Maddeler" sütunlarında gösterilmiştir. Olası bağımlı olan madde çiftleri de ayrıca incelenmiştir ancak herhangi bir bağımlılık nedeni görülememiştir. Bir VST maddesine verilen yanlış veya doğru yanıtın, başka bir VST maddesine verilen yanlış veya doğru yanıtlarla ilişkili olması mümkün değildir. Bunun nedeni, VST maddelerinin her biri farklı bir kelimenin anlamını sormaktadır. Soru kökleri çok kısadır ve

aynı soru köküne sahip herhangi bir madde çifti yoktur. Sorulan kelimenin ortak kökünden kaynaklı ya da benzer anlama sahip kelimelerden kaynaklı bir yerel bağımsızlık ihlali söz konusu olup olmadığı da ayrıca incelenmiş, fakat o noktadan da herhangi bir bulguya rastlanmamıştır. Olası yerel bağımlı madde çiftlerinden en yüksek 3 Q3 değerine sahip madde çiftleri EK-E'de Q3 değerleriyle birlikte verilmiştir. Şu noktaya da değinmek gerektir ki, bağımlılığın tek kaynağı ortak bir metin ya da soru köküne sahip olmak değildir. Ackerman'a (1987) göre, madde parametreleri (örn. ayırt edicilik ve güçlük) ve maddelerin sıralaması (örn. kolaydan zora ya da zordan kolaya) da yerel bağımlılığa yol açabilir. Bu çalışmada, VST maddeleri kolaydan zora doğru sıralanmıştır ve Tablo 11'den yerel olarak bağımlı madde çiftlerinin çoğunlukla komşu maddeler olduğu görülebilir. Yerel bağımsızlığı ihlal eden madde çiftlerinin içerikleri incelenmek istenildiğinde "Veri Toplama Araçları" bölümündeki bağlantıya tıklayarak VST'ye ulaşmak mümkündür.

Tablo 12

1PLM ile 2PLM'nin Karşılaştırılması

	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	sd	p
1PLM	155622.6	155657.6	155902.3	155895.7	156350.2	-77670.29	-	-	-
2PLM	152677.1	152833.5	153232.5	153219.5	154121.9	-76058.56	3223.472	139	0

Tablo 13

2PLM ile 3PLM'nin Karşılaştırılması

	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	sd	p
2PLM	152677.1	152833.5	153232.5	153219.5	154121.9	-76058.56	-	-	-
3PLM	152254.2	152662.5	153087.3	153067.7	154421.4	-75707.08	702.946	140	0

Tek boyutluluk ve yerel bağımsızlık varsayımlarını test ettikten sonra, hangi MTK modelinin verilere en iyi uyduğunu belirlemek için MTK analizleri yapılmıştır. İlk olarak, kestirimler bir ve iki parametreli lojistik modellerle yapılmıştır. Daha sonra olabilirlik oran testi yapılıp model uyum indeksleri incelenerek iki model karşılaştırılmıştır. Tablo 12, olabilirlik oranı testinin sonuçlarını göstermektedir. İki parametreli modelde AIC, SABIC ve

BIC değerlerinde düşüşler görülmektedir. Ayrıca, daha büyük bir logLik değeri hesaplanmıştır. Olabilirlik oranı testinin p-değeri sıfır olarak kestirilmiştir ve bu da 2PLM'nin verilere 1PLM'den daha iyi uyduğu anlamına gelmektedir.

İki parametreli modelin tek parametreli modele göre daha iyi uyum sağladığı tespit edildikten sonra, aynı test 2PLM ile 3PLM'yi karşılaştırmak için de gerçekleştirilmiştir. Tablo 13 bu karşılaştırmanın sonuçlarını göstermektedir. AIC, SABIC ve BIC model uyum indekslerinin değerlerinde yine düşüşler olsa da bu düşüşler 1PLM ve 2PLM karşılaştırmasındaki kadar büyük değildir. Her ne kadar logLik değerindeki artış çok fazla olmasa da olabilirlik testinin p-değerinin anlamlı olması 3PLM'nin 2PLM'ye göre veriye daha iyi uyduğunu göstermektedir.

Yapı Geçerliğine İlişkin Bulunan Kanıtlar - DMF

Değişen madde fonksiyonu (DMF) analizleri Lojistik regresyon, Lord'un ki-kare testi ve Mantel-Haenszel yöntemleri ile gerçekleştirilmiştir. Her üç yöntem tarafından da DMF maddesi olarak işaretlenen 34 madde bulunmaktadır. Bu maddeler Tablo 14'te listelenmiş ve üç DIF yönteminin sonuçları EK-F'de verilen grafiklerde görselleştirilmiştir.

Lojistik regresyon ve Lord'un ki-kare yöntemlerinin sonuçlarına göre, önemli düzeyde DMF gösteren herhangi bir madde bulunmamaktadır. Ancak Mantel-Haenszel sonuçları, 34 madde arasında 24 maddenin ihmal edilebilir veya orta düzeyde DMF, 10 maddenin ise önemli düzeyde DMF gösterdiğini ortaya koymaktadır. Bir madde için hesaplanan Δ MH'nin mutlak değeri 1.50'den yüksekse, maddenin önemli düzeyde DMF sergilediği kabul edilir (Magis ve diğerleri, 2010). Önemli düzeyde DMF gösteren maddeler Tablo 14'te koyu renkle gösterilen 3, 7, 17, 20, 63, 72, 74, 98, 104 ve 138 numaralı maddelerdir. Ayrıca 34 madde için DMF istatistikleri de gösterilmektedir. DMF maddelerinin LRT istatistikleri, bu istatistiğe ilişkin p-değeri ve Nagelkerke'nin R^2 değeri (Nagelkerke, 1991) Lojistik regresyon yöntemi sonuçlarında verilmiştir. Lord'un ki-kare yönteminin sonuçlarında Lord'un χ^2 istatistiği ve bu istatistiğe ait p-değeri verilmiştir. Mantel-Haenszel sonuçlarında ise ki-kare ve p-değerlerinin yanı sıra α MH ve Δ MH değerleri de sunulmaktadır.

Tablo 14

DMF Gösteren Maddeler ve İstatistikleri

		Lojistik Regresyon		Lord'un ki-karesi				Mantel-Haenszel	
Madde	LRT İstatistiği	p-değeri	R^2	Lord's χ^2	p-değeri	MH χ^2	p-değeri	α MH	Δ MH
3	**20.1307**	**0.0000**	**0.0305**	**14.3176**	**0.0008**	**15.5451**	**0.0001**	**0.3939**	**2.1892**
7	**22.9162**	**0.0000**	**0.0288**	**10.7711**	**0.0046**	**15.7926**	**0.0001**	**0.4775**	**1.7370**
14	13.8811	0.0010	0.0096	8.9157	0.0116	9.9325	0.0016	0.6513	1.0078
16	10.7591	0.0046	0.0075	7.9392	0.0189	10.0675	0.0015	0.6629	0.9663
17	**14.9093**	**0.0006**	**0.0297**	**9.6579**	**0.0080**	**9.9347**	**0.0016**	**0.2531**	**3.2290**
20	**56.2356**	**0.0000**	**0.0381**	**33.9760**	**0.0000**	**49.5375**	**0.0000**	**0.4083**	**2.1048**
25	10.8871	0.0043	0.0139	9.1278	0.0104	9.7025	0.0018	0.5439	1.4310
31	12.1880	0.0041	0.0137	13.2703	0.0033	7.6450	0.0087	0.5606	1.3603
32	9.3430	0.0094	0.0060	10.8923	0.0043	8.2690	0.0040	1.4813	-0.9234
34	19.5076	0.0001	0.0107	21.1253	0.0000	15.2666	0.0001	1.7712	-1.3434
55	12.6630	0.0018	0.0074	9.8334	0.0073	7.9740	0.0047	0.6683	0.9472
59	14.4812	0.0007	0.0077	20.3274	0.0000	13.8536	0.0002	1.7509	-1.3163
62	11.0667	0.0040	0.0098	12.7461	0.0017	11.9657	0.0005	1.8575	-1.4552
63	**59.7718**	**0.0000**	**0.0313**	**42.5657**	**0.0000**	**48.6764**	**0.0000**	**0.3481**	**2.4799**
69	8.2869	0.0159	0.0047	10.1917	0.0061	6.0796	0.0137	1.4578	-0.8857
72	**39.5102**	**0.0000**	**0.0323**	**49.7523**	**0.0000**	**24.5789**	**0.0000**	**2.3489**	**-2.0068**
74	**19.3315**	**0.0001**	**0.0149**	**31.8527**	**0.0000**	**15.2801**	**0.0001**	**2.5784**	**-2.2258**
82	7.8601	0.0196	0.0037	13.1187	0.0014	5.7513	0.0165	1.4983	-0.9501
90	6.1641	0.0459	0.0056	7.7200	0.0211	6.1266	0.0133	1.4817	-0.9241
92	13.3572	0.0013	0.0079	10.6864	0.0048	13.1646	0.0003	0.5860	1.2560
93	9.6763	0.0079	0.0053	12.1010	0.0024	6.4793	0.0109	1.4568	-0.8842
98	**41.6161**	**0.0000**	**0.0234**	**30.4000**	**0.0000**	**38.0141**	**0.0000**	**0.3789**	**2.2806**
104	**31.2653**	**0.0000**	**0.0153**	**20.8562**	**0.0000**	**33.5489**	**0.0000**	**0.4063**	**2.1163**
109	17.3575	0.0002	0.0089	11.0125	0.0041	16.6053	0.0000	0.5471	1.4175
114	16.1258	0.0003	0.0080	8.8901	0.0117	14.5232	0.0001	0.5423	1.4380
115	9.4762	0.0088	0.0049	13.0773	0.0014	6.6750	0.0098	1.4819	-0.9243
116	11.5260	0.0031	0.0113	9.2755	0.0097	7.4143	0.0065	1.6128	-1.1232
122	9.6359	0.0081	0.0056	7.7227	0.0210	8.7972	0.0030	0.6240	1.1084
123	9.7038	0.0078	0.0050	5.9939	0.0499	7.7186	0.0055	0.6616	0.9709
124	10.1171	0.0064	0.0058	12.0558	0.0024	8.3024	0.0040	1.5783	-1.0724
128	11.6936	0.0029	0.0061	16.1651	0.0003	7.1759	0.0074	1.5005	-0.9537
130	6.0630	0.0482	0.0037	9.7588	0.0076	3.9663	0.0464	1.3197	-0.6519
138	**24.9563**	**0.0000**	**0.0156**	**23.2532**	**0.0000**	**20.7391**	**0.0000**	**1.9070**	**-1.5170**
140	10.6448	0.0049	0.0116	10.8352	0.0044	11.0913	0.0009	1.8808	-1.4845

Tablo 14'te gösterildiği gibi, DMF gösteren maddelerin her üç yöntemde de hesaplanan p değerleri .05'ten küçüktür. Erkekler ve kadınlar lehine DMF gösteren maddeleri Δ MH değerlerini inceleyerek karar verilebilir (Magis ve diğerleri, 2010). DeltaMH

(Δ MH) değerinin negatif olması referans grup lehine, pozitif olması ise odak grup lehine DMF olduğunu göstermektedir (Holland & Thayer, 1988). DMF analizi için yazılan kodlarda referans grup olarak kadınlar önceden belirlenmiştir. Önemli düzeyde DMF sergileyen maddeler arasında negatif Δ MH değerlerine sahip olan maddelerin 72, 74 ve 138 numaralı maddeler olduğu görülmektedir. Bu maddeler kadınlar lehine DMF sergilemektedir ve bu maddelerde yer alan kelimeler sırasıyla *palette, kindergarten* ve *erythrocyte*'dir. Ayrıca, bu DMF maddelerinin madde karakteristik eğrileri (MKK) EK-G'de gösterilmiştir. MKK'ler incelendiğinde, kadınlar için bu maddelere doğru yanıt verme olasılığının yetenek ölçeğinin hemen her düzeyinde daha yüksek olduğu görülmektedir. Pozitif Δ MH değerlerine sahip olan 3, 7, 17, 20, 63, 98 ve 104. maddeler erkekler lehine önemli düzeyde DMF göstermektedir. Bu maddelerde sorulan kelimeler sırasıyla *period, jump, pub, pro, stealth, crowbar* ve *counterclaim* olup bu maddelerin MKK'leri EK-Ğ'de verilmiştir.

Kapsam Geçerliğine İlişkin Bulunan Kanıtlar

3PLM'nin verilere en iyi uyduğu gözlemlendikten sonra, kapsam geçerliği kanıtı elde etmek için madde ve birey parametreleri üç parametreli lojistik model ile kestirilmiştir. EK-H'de, VST maddelerinin güçlük parametreleri ile cevaplayıcıların yetenek parametrelerinin aynı ölçek üzerinde yer aldığı birey-madde haritası (Wright-Map) verilmiştir. Bu haritada, VST'nin her yetenek parametresi düzeyinde yeterli sayıda maddeye sahip olduğu, yani 140 maddelik VST'nin hem düşük hem de yüksek yeterlilikteki bireylerin kelime bilgisini ölçebildiği görülmektedir. 140 madde arasında en kolay maddeler 6., 2. ve 1. maddelerdir. Bu maddeler için b parametreleri sırasıyla -6.50, -6.23 ve -5.45 olarak kestirilmiştir. En zor maddeler 96., 58. ve 68. maddelerdir ve bu maddeler için b parametreleri sırasıyla 3.14, 3.09 ve 3.01 olarak bulunmuştur. Bu maddelerin yerleri birey-madde haritasında görülebilir.

Daha önce de belirtildiği gibi, bir VST maddesinin sırası azaldıkça, bu maddede kullanılan kelimenin İngilizce metinlerde kullanım sıklığı artmaktadır. Bir kelimenin İngilizce metinlerde kullanım sıklığı arttığında ise o kelimenin sorulduğu bir test maddesi daha kolay bir madde olurken, kullanım sıklığı azaldıkça sorunun güçlüğü artmaktadır. Bu teorik

varsayım alan yazındaki birçok çalışma ile ampirik olarak büyük oranda doğrulanmıştır (Beglar, 2010; Laufer & Goldstein, 2004; Schmitt ve diğerleri, 2001)

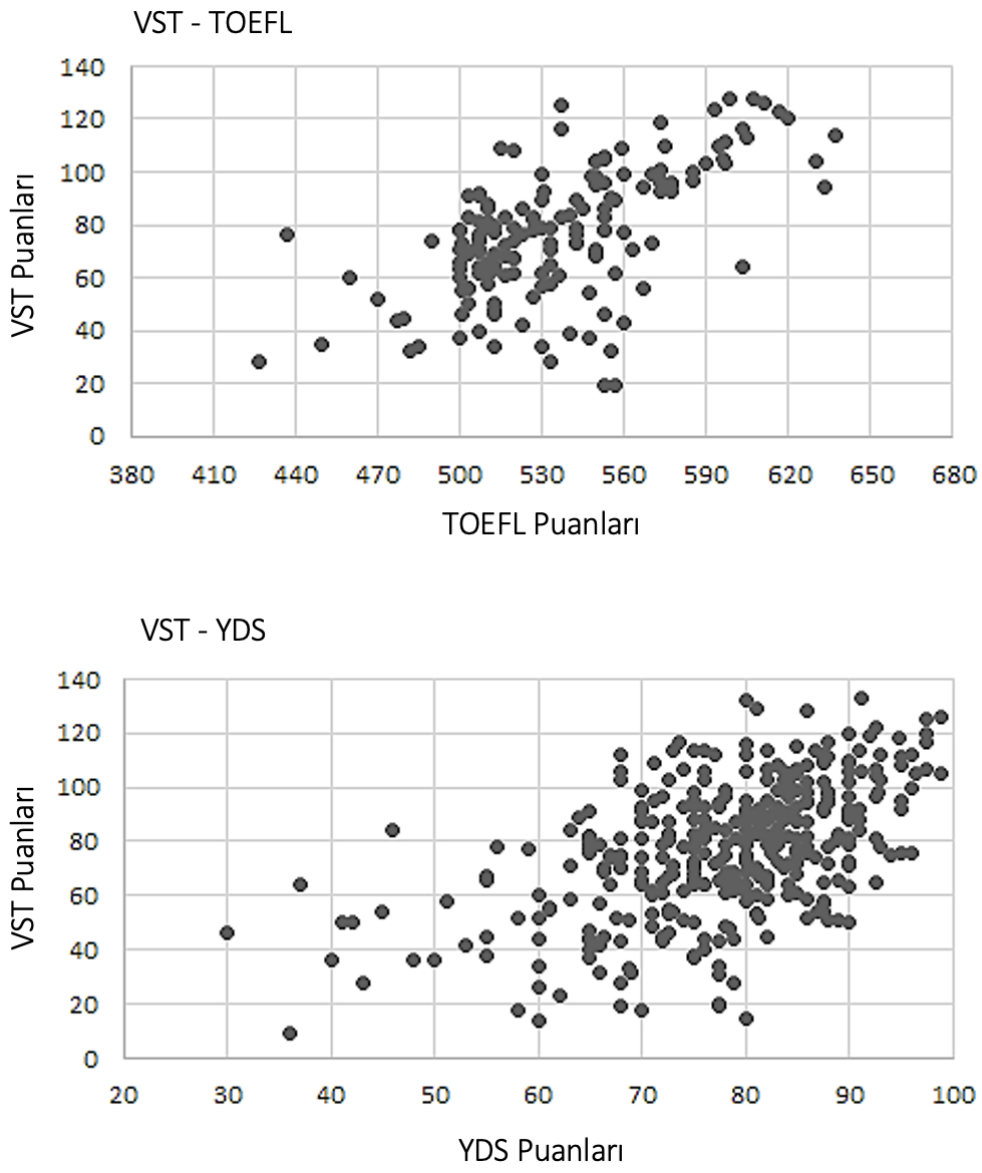
Birey-madde haritasında (EK-H) bu teorinin bir dereceye kadar geçerli olduğunu rahatlıkla görebiliriz. İlk sorular ölçeğin sol tarafında yer almakta ve maddelerin sıra numarası arttıkça giderek sağ tarafa kaymaktadır. Ancak bazı istisnalar da mevcuttur. İlk olarak, beklenenden daha zor olan bazı sorular vardır. Bunlar 4, 16, 58 ve 68. maddelerdir ve bu maddelerde kullanılan kelimeler sırasıyla *figure, nil, cavalier* ve *azalea*'dır. Bu dört madde, 10'ar kelimeden oluşan gruplarındaki diğer maddelere göre oldukça yüksek güçlük parametrelerine sahiptir. Örneğin, ilk 10 maddenin güçlük parametrelerinin ortalaması -3,80 iken, 4. maddenin b parametresi 0,55'tir. Kişi-madde haritasından 4. maddenin gruptan ayrılması açıkça görülebilmektedir. İkinci olarak, beklenenden daha kolay olan yaklaşık 20 soru vardır. Bunlar 35, 46, 47, 50, 54, 56, 61, 67, 70, 72, 74, 83, 88, 94, 103, 105, 107, 117 ve 126. maddelerdir ve bu maddelerde kullanılan kelimeler *quiz, cube, miniature, bacterium, accessory, thesis, olive, demography, yoghurt, palette, kindergarten, monologue, octopus, mystique, yoga, puma, aperitif, caffeine* ve *plankton*'dur. Olive ve kindergarten hariç, tüm bu kolay kelimeler alıntı veya ortak kelimelerdir (hem İngilizcede hem de Türkçede bulunmaktadır) ve dolayısıyla bu kelimeler kendi gruplarındaki diğer maddelere göre oldukça fazla sayıda doğru cevaplanmıştır. Örneğin, 126. madde plankton kelimesinin anlamını sormaktadır ve b parametresi -1.30 olarak hesaplanmıştır. Ancak aynı gruptaki diğer maddeler aynı parametre için 1,70 ortalamaya sahiptir. Benzer şekilde, bu maddelerin gruplarından sapmaları birey-madde haritasında da görülebilir (EK-H).

Uyum Geçerliğine İlişkin Bulunan Kanıtlar

VST ile veri toplandığı anda, katılımcılara bir İngilizce yeterlilik sınavından aldıkları en son puan da sorulmuştur. Bu soruya yaklaşık 600 öğrenci yanıt vermiştir. Yanıtlar, TOEFL ve YDS olmak üzere iki İngilizce yeterlilik sınavından alınan puanları içermektedir. VST puanları ile bu iki İngilizce yeterlilik testinden alınan puanlar arasındaki ilişki incelenmiştir.

Şekil 11

VST ve İki Dil Sınavına ait Saçılım Grafikleri



VST puanları ile TOEFL puanları ve YDS puanlarının saçılım grafiklerinin gösterildiği Şekil 11'de VST ve TOEFL puanları ile VST ve YDS puanları arasındaki ilişki görülebilir. Yüz altmış öğrenci TOEFL puanlarını vermiştir ve şekil 11'deki ilk grafikten de görülebileceği gibi, VST ve TOEFL puanları arasında pozitif bir korelasyon vardır. Bu değişkenler için korelasyon katsayısı .60 (%95 GA = .49, .69) olarak hesaplanmıştır. Katılımcıların 368'i YDS sonuçlarını bildirmiştir ve VST ile YDS sonuçları arasındaki pozitif korelasyon ikinci grafikte görülebilir. Korelasyon katsayısı bu ikisi için .53 (%95 GA = .45, .60) olarak bulunmuştur. İki pozitif yönlü ve manidar korelasyon VST puanlarının diğer İngilizce

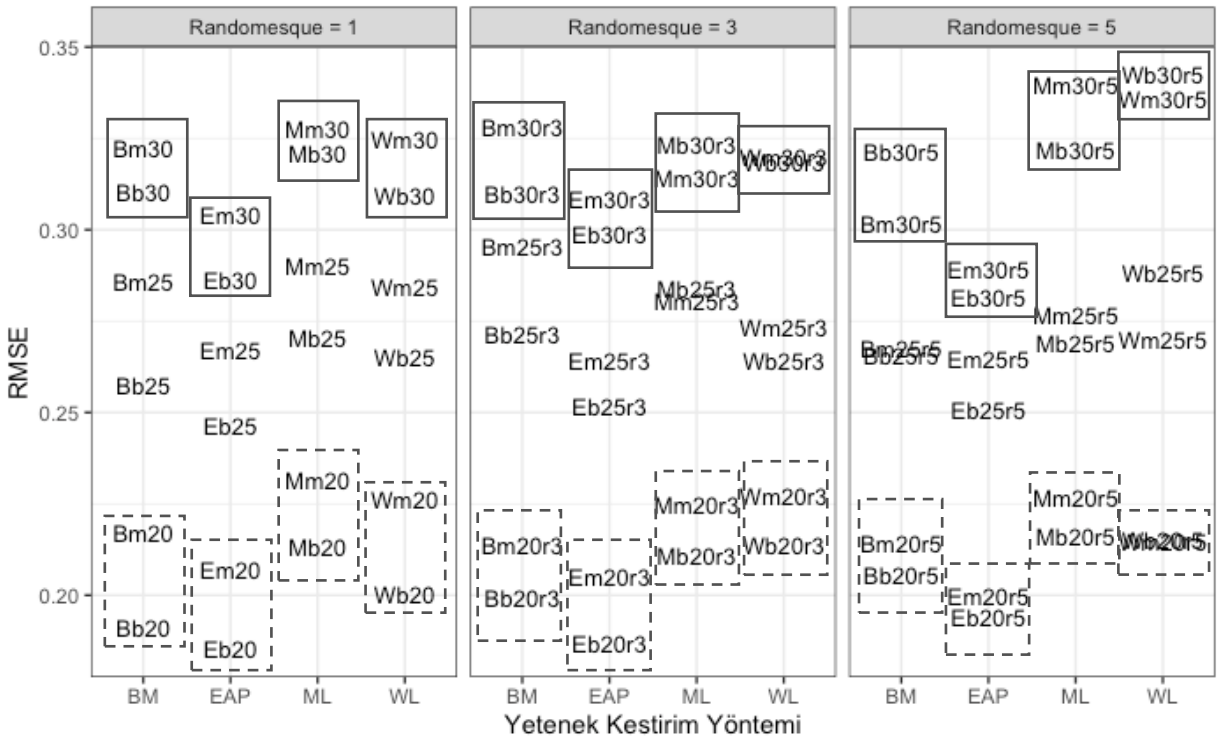
yeterlilik ölçümleriyle yakından ilişkili olduğunu gösterir ve bu da VST'nin geçerliği için uyum geçerliğine dair kanıt sağlar.

Post-hoc Simülasyon Bulguları

VST'nin İngilizce kelime bilgisini ölçme amacıyla kullanılabilecek geçerli bir ölçme aracı olduğunun tespitinden sonra, VST maddelerinden bir madde havuzu oluşturarak gerçek zamanlı bir BBT uygulaması geliştirme çalışmalarına başlanmıştır. BBT uygulaması için en uygun başlama, ilerleme, sonlandırma ve puanlama koşullarını belirleyebilmek için geçerlik çalışmalarında toplanan veri kullanılarak post-hoc simülasyonlar yapılmıştır. Simülasyon çalışmasından elde edilen sonuçlar genelden özele olacak şekilde bu bölümde verilmiştir. Öncelikle tüm koşullar için kestirilen RMSE, yanlılık, korelasyon ve ortalama test uzunluğu istatistikleri incelenmiştir.

Şekil 12

Tüm Koşullar için RMSE Değerleri



Şekil 12'de tüm koşullar için kestirilen RMSE değerleri gösterilmiştir. Grafiğin x ekseninde simülasyon koşullarına dahil edilen dört yetenek kestirim yöntemi, y ekseninde

ise RMSE değerlerine yer verilmiştir. Grafik, madde kullanım sıklığı kontrol yöntemine göre üç farklı yüzeyden oluşmaktadır. Grafikler oluşturulurken temel alınan simülasyon bulguları tablo halinde tüm koşullar için EK-I'da verilmiştir. Hem tablodan hem de grafikten görüleceği üzere tüm koşullarda kestirilen RMSE değerleri .1854 - .3423 aralığında yer almaktadır.

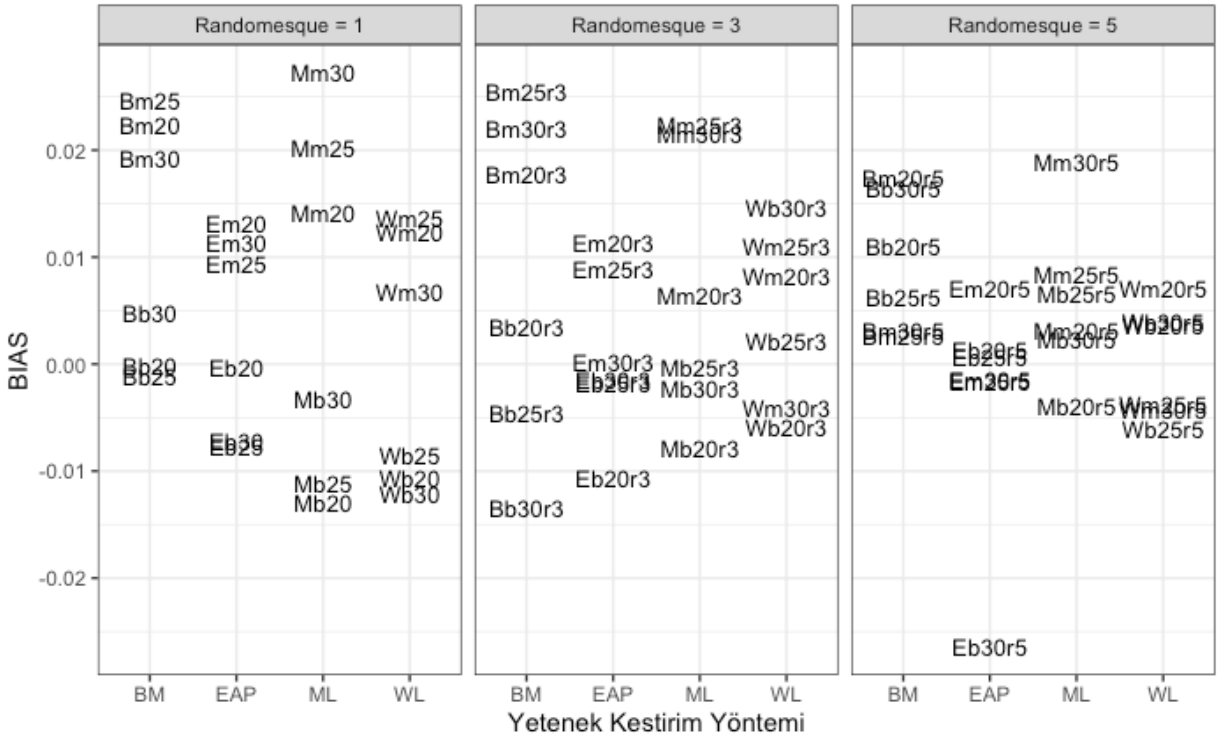
Grafiğe bakıldığında ilk olarak, RMSE değerlerinin, madde kullanım sıklığı kontrol yöntemine göre önemli bir şekilde farklılık göstermediği gözlenmektedir. Madde kullanım sıklığı kontrol yönteminin her düzeyinde, standart hata değeri .30 olan koşullar (Düz çizgili dörtgen içindeki koşullar) RMSE değeri olarak çoğunlukla .30 - .35 aralığında, hata değeri .25 olan koşullar (dörtgen içinde olmayan koşullar) RMSE değeri olarak çoğunlukla .25 - .30 aralığında ve hata değeri .20 olan koşulların (kesikli çizgili dörtgen içindeki koşullar) ise tamamı RMSE değeri olarak .25'in altında bulunmaktadırlar.

Beklendiği üzere standart hata kesme değeri .20 olan koşullar en düşük RMSE değerlerini üretirken, .30 olan koşullar en yüksek RMSE değerlerini üretmiştir. Diğer tüm koşulların aynı olduğu, sadece madde seçim yönteminin farklılaştığı durumlarda bOpt (koşul isminde ikinci karakteri b olan koşullar) yönteminin kullanıldığı koşullar MFI (koşul isminde ikinci karakteri m olan koşullar) yönteminin kullanıldığı koşullara göre çoğunlukla daha az RMSE değerlerine sahiptir. Fakat ilerleyen bölümlerde de değinileceği üzere, bunun sebebi bOpt yönteminin MFI yöntemine göre çok daha fazla madde kullanmasıdır. Grafikte yetenek kestirim yöntemlerinin performansına bakıldığında EAP yöntemini içeren koşulların (koşul isminde ilk karakteri E olan koşullar) daha düşük RMSE değerlerine sahip olduğu görülmektedir.

Gerçek zamanlı BBT uygulamasında hangi senaryonun tercih edileceğini seçebilmek için öncelikle genel RMSE, yanlılık (bias), ortalama test uzunluğu değerlerine bakılmış, sonra da alan yazında da tavsiye edildiği şekilde bireyler, yetenek düzeylerine göre 10 eşit gruba ayrılarak, her grup için hesaplanan RMSE, yanlılık ve ortalama test uzunluğu değerlerine bakılmıştır.

Şekil 13

Tüm Koşullar için Yanlılık Değerleri

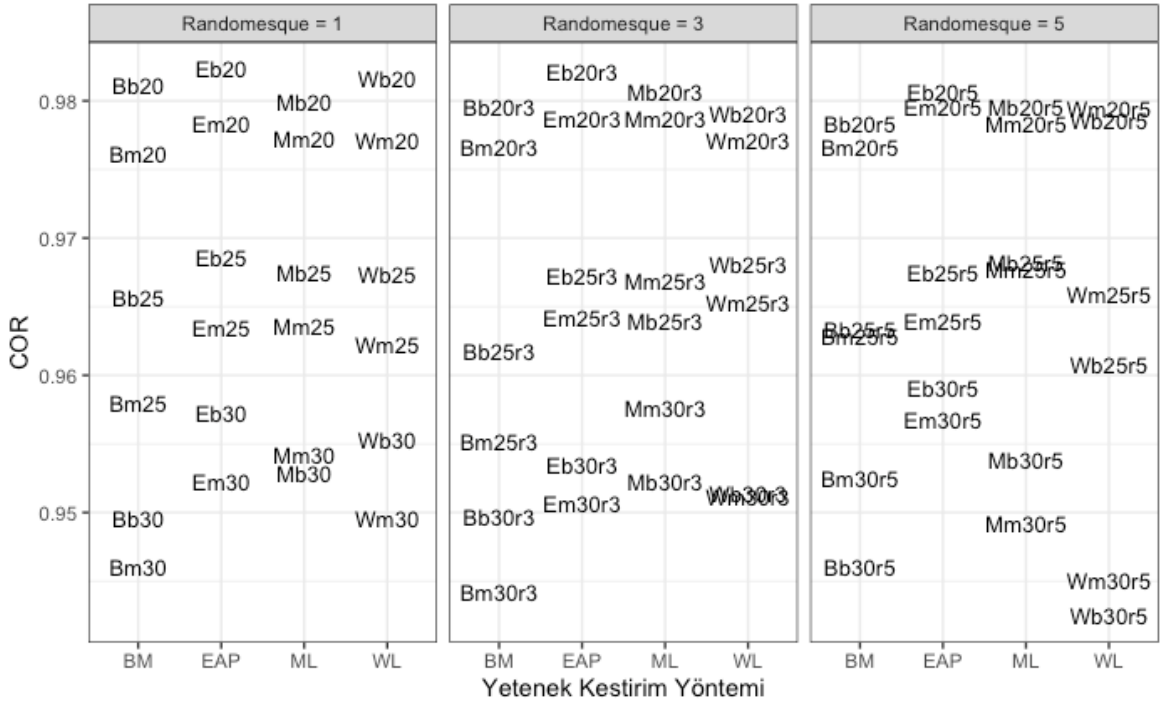


Şekil 13'te tüm koşullar için ortalama yanlılık (BIAS) değerleri verilmiştir. Şekil 12'de olduğu gibi grafiğin x ekseninde simülasyon koşullarına dahil edilen dört yetenek kestirim yöntemi vardır, y ekseninde ise bu sefer BIAS değerlerine yer verilmiştir. Grafik, aynı şekilde madde kullanım sıklığı kontrol yöntemine göre üç farklı yüzeyden oluşmaktadır. EK-I'daki tabloya bakıldığında tüm koşullar için yanlılık değerlerinin - .0264 ile .0272 aralığında yer aldığı görülmektedir.

"randomesque = 1" ve "randomesque = 3" yüzeylerinde MFI madde seçim yöntemi ile gerçekleştirilen simülasyon koşullarının (koşul isminde ikinci karakteri m olan koşullar) bireyin yeteneğini genellikle olduğundan biraz daha yüksek kestirdiği (overestimate) görülmektedir. bOpt yöntemi kullanıldığında ise (koşul isminde ikinci karakteri b olan koşullar) yine genellikle bireyin yeteneği olduğundan biraz daha düşük kestirildiği (underestimate) dikkati çekmektedir. Üçüncü yüzeyde ise böyle bir ayrım görülmemektedir ve yanlılık değerleri birbirine çok yakın çıkmıştır. Yanlılık kestirimleri mutlak değer olarak dikkate alınırsa eğer, koşullar arasında büyük farklar olmadığı görülecektir.

Şekil 14

Tüm Koşullar için Gerçek ve Kestirilen Yetenek Değerleri Arasındaki Korelasyon

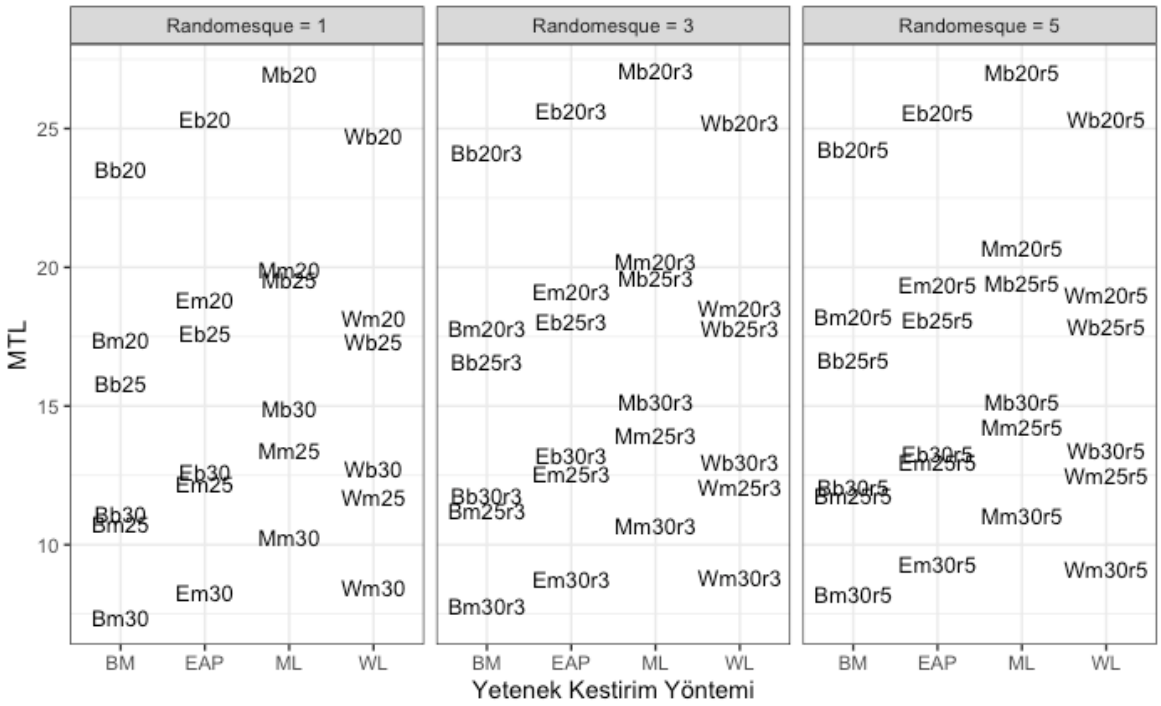


Şekil 14'te ise bireylerin kestirilen yetenek düzeyleri ile gerçek yetenek düzeyleri arasındaki korelasyonlar verilmiştir. Tüm koşullar için korelasyon değerleri oldukça yüksek çıkmıştır ve EK-I'daki yer alan tabloya bakıldığında .9425 - .9823 aralığında yer aldıkları görülmektedir. Grafikten de anlaşılacağı üzere hata değeri olarak .20 seçilen koşullar daha yüksek korelasyonlar üretmiştir. Hata değeri .30 olan koşullar ise bir istisna hariç (Mm30r3 > Bm25r3) en düşük korelasyon değerlerine sahiptir.

Şekil 15'te tüm koşullarda ortalama test uzunluğu değerleri verilmiştir. Grafiğin geneline bakıldığında hata değeri olarak .20 seçilen koşulların, hata değerini kesme noktasının altına indirip testi sonlandırabilmek için çok daha fazla maddeye ihtiyaç duyduğu gözlemlenmektedir. Hata değeri .30 olan koşulların ise çok az madde ile testi sonlandırabildiği görülmektedir. Ayrıca madde seçim yöntemine göre ortalama test uzunluklarının oldukça farklılaştığı görülmektedir. Bu sebeple ayrı bir tablo (Tablo 15) ile bu durum daha ayrıntılı bir şekilde incelenmiştir.

Şekil 15

Tüm Koşullar için Ortalama Test Uzunluğu



Tablo 15

Sonlandırma Kuralı ve Madde Seçim Yöntemine Göre Ortalama Test Uzunluğu

SH kesme değeri	SH = .30			SH = .25			SH = .20		
Madde seçim yöntemi	Min	Maks	Ort	Min	Maks	Ort	Min	Maks	Ort
MFI	7.38	11.04	**9.01**	10.76	14.28	**12.45**	17.41	20.70	**18.94**
bOpt	11.09	15.14	**13.20**	15.82	19.59	**17.87**	23.54	27.07	**25.41**

Tablo 15'de tüm koşullar madde seçim yöntemi ve standart hata kesme değerine göre gruplandırılarak ortalama test uzunluğuna ait minimum, maksimum ve ortalama istatistikleri verilmiştir. Her grupta 12 koşul vardır. Madde seçme yöntemi MFI, hata değeri .30 olan 12 koşulun ortalama test uzunluğu değerlerine ait minimum değer 7.38, maksimum değer 11.04 ve ortalama değer de 9.01 olarak bulunmuştur. Aynı koşullarda madde seçim yöntemi bOpt seçildiğinde ise ortalama test uzunluğuna ait ortalama test uzunluğu 13.2 olarak hesaplanmıştır. Hata değerinin diğer değerlerinde de aynı durum gözlenmektedir.

Eğer madde seçim yöntemi olarak bOpt seçilirse, istenen sonlandırma kuralını gerçekleştirebilmek için yaklaşık %30 daha fazla madde kullanımına ihtiyaç duyulmaktadır.

Tüm koşullarda hesaplanan RMSE, yanlılık, korelasyon ve ortalama test uzunluğu istatistiklerine bakıldığında, madde kullanım sıklığı kontrol yöntemine göre bu istatistiklerin önemli bir düzeyde farklılaşmadığı görülmektedir. Bu sebeple gerçek zamanlı BBT uygulaması için seçilecek koşulun belirlenmesi sürecinde yapılacak ilk eleme madde kullanım sıklığı kontrol yöntemi hiç kullanılmayan, yani "randomesque = 1" olan koşullar ve "randomesque = 3" olan koşullardır. Toplam 48 koşul böylece elenmiştir. Geriye kalan 24 koşulun hepsinde "randomesque = 5"tir. Yani bireyin anlık olarak kestirilen yetenek düzeyine göre beş tane en uygun madde belirlenmektedir ve bu beş madde arasından bir tanesi rastgele seçilip bireye uygulanmaktadır. Böylece "a" parametresi çok yüksek olan bazı maddelerin tekrar tekrar kullanımının önüne geçilmekte ve madde kullanım sıklığı bu şekilde kontrol edilmektedir.

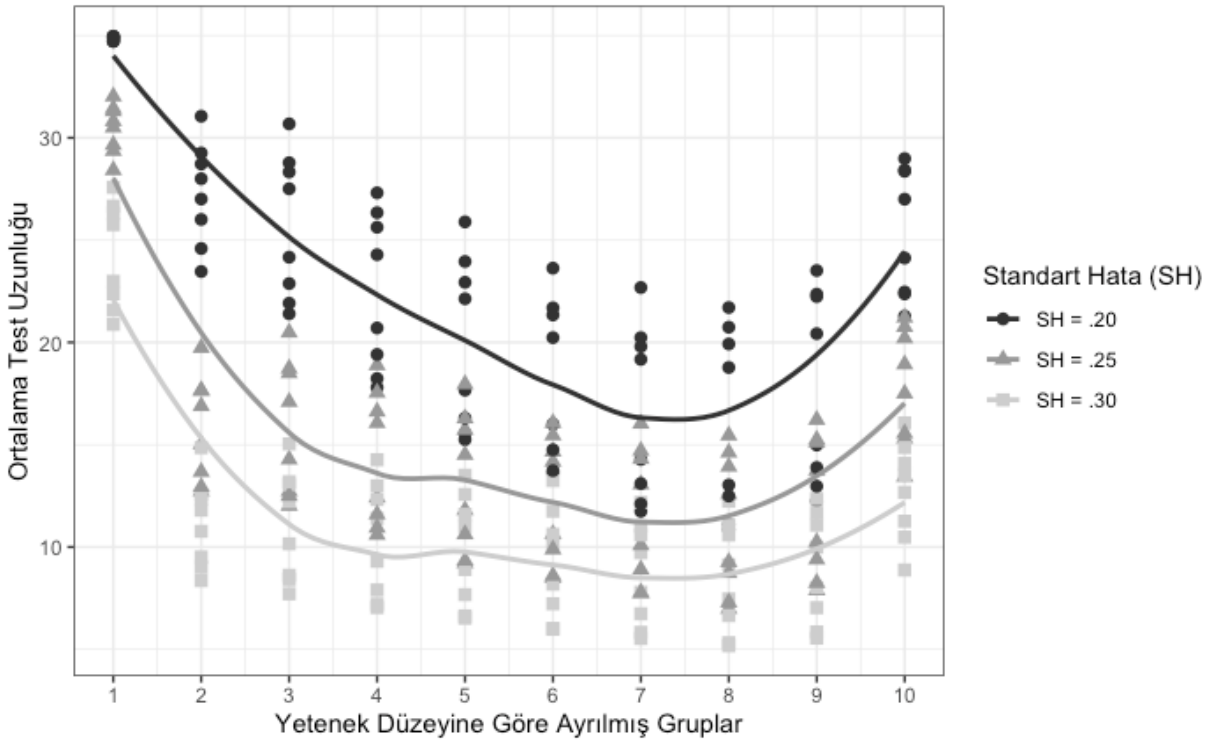
Şekil 16

"randomesque = 5" Olan Koşullarda Ortalama Test Uzunluğu

Geriye kalan 24 koşula ait ortalama test uzunluğu grafiği Şekil 16'da verilmiştir. Hata değeri .30 olan koşulların çok az madde ile testi sonlandırdığı görülmektedir. Beklendiği üzere de .20 olan koşullar ise belirlenen kesme değerinin altına inebilmek için daha fazla maddeye ihtiyaç duymaktadır.

Şekil 17

Yetenek Düzeyine Göre Ayrılmış Gruplarda Ortalama Test Uzunluğu



Şekil 17'de yetenek düzeylerine göre gruplara ayrılmış bireyler 10 grup olarak x ekseninde verilmiştir. Ortalama test uzunluğu ise y ekseninde verilmiştir. EK-İ, EK-J ve EK-K'de sırasıyla yetenek düzeyine göre oluşturulan gruplardaki ortalama RMSE, yanlılık ve ortalama test uzunluğu değerleri verilmiştir. Yapılan simülasyonlarda tüm koşullarda gerçek veri kullanıldığı için, koşullarda kullanılan birey yetenek parametreleri hepsinde aynıdır. Bu sebeple bireyler yetenek düzeylerine göre 10 eşit gruba bölündüklerinde, her simülasyon koşulunda her gruba aynı bireyler düşmektedir. Böylece her grubun yetenek parametresine ait ortalama, her simülasyon koşulunda aynıdır. Alttaki grafikte x ekseninde gruplar sırasıyla

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 olarak adlandırılmıştır. Gruplara ait ortalama θ ($\bar{\theta}$) değerleri sırasıyla -2.27, -0.88, -0.36, -0.04, 0.14, 0.30, 0.44, 0.61, 0.85 ve 1.31'dir.

Bu grafikten yetenek ölçeğinde her iki uçtaki bireylerin İngilizce kelime bilgisini kestirebilmek için daha fazla sayıda maddeye ihtiyaç duyulduğu görülmektedir. Bu sayı özellikle yeteneği düşük bireylerde daha fazladır. Hatta hata değerinin .20 olduğu 8 koşulun çoğunda ortalama test uzunluğu 35'e çok yakın çıkmıştır (EK-K'deki tablodan ayrıntılı incelenebilir). Yani yeteneği en düşük grupta bireylerin çoğu için hata değerini .20'nin altına indirmek mümkün olmamış ve bu sebeple test 35 madde uygulanınca sonlanmıştır.

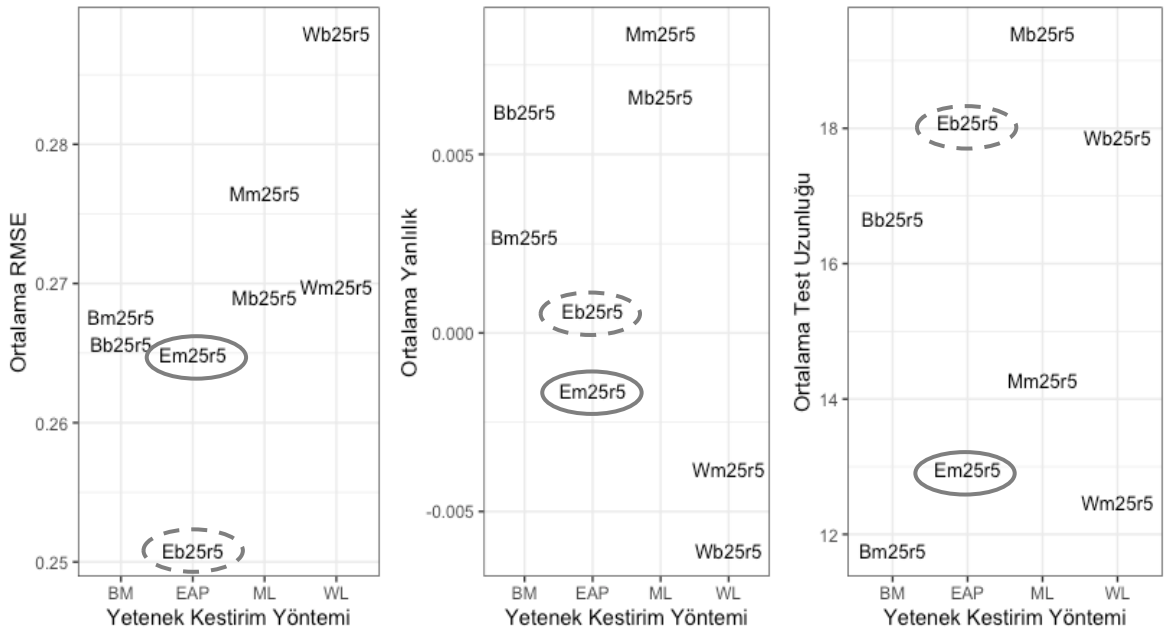
Hata değeri .20 olan koşulların nispeten fazla maddeye ihtiyaç duyması ve yetenek değeri düşük bireylerde belirlenen kesme noktasının altına inmenin çoğu zaman mümkün olmaması sebebiyle, hata değeri .20 olarak belirlenen koşullar da elenmiştir. Hata değeri .30 olan koşullar ise çok az madde kullanması sebebiyle elenmiştir. Yetenek düzeyinde orta gruplarda yer alan bireyler için testin genellikle 10 maddenin altında hatta bazı koşullarda ortalama 5-6 madde ile sonlandığı bulunmuştur. Bu sebeple bahsedilen her iki hata değerine kıyasla daha makul ortalama test uzunluğuna sahip, hata değeri .25 olan koşullar arasından gerçek BBT uygulamasında kullanılacak yöntemlerin seçilmesine karar verilmiştir.

Şekil 18'de madde kullanım sıklığı kontrol yöntemi "randomesque = 5" ve hata değeri .25 olan koşulların ortalama RMSE, yanlılık ve test uzunluğu değerleri tek bir grafikte verilmiştir. Ortalama RMSE değerlerinde en düşük RMSE değerlerine sahip iki koşul Eb25r5 (kesik çizgili daire ile gösterilmiştir) ve Em25r5 (düz çizgili daire ile gösterilmiştir) koşullarıdır. Yine aynı şekilde mutlak değer olarak en düşük yanlılık değerleri de bu iki koşula aittir. İki koşul arasındaki tek fark madde seçim yöntemidir. Eb25r5'te madde seçim yöntemi bOpt, Em25r5'te MFI'dır. Her iki koşulun ortalama test uzunluğu değerlerine bakıldığında madde seçim yöntemi olarak bOpt kullanan koşulun %40 civarı daha fazla madde kullandığı görülmektedir. Yani Eb25r5 koşulu bu kadar düşük ölçme kesinliği elde etmek için, yani düşük RMSE ve yanlılık değerlerine sahip olabilmek için diğer koşula

kıyasla %40 daha fazla maddeye ihtiyaç duymaktadır. BBT uygulamalarının en temel amaçlarından biri aynı ölçme kesinliğini daha az madde ile elde etmek olduğu için gerçek zamanlı BBT uygulamasında tercih edilecek koşul Em25r5 olarak belirlenmiştir. Yani nihai olarak seçilen koşulda yetenek kestirim yöntemi EAP, madde seçim yöntemi MFI, sonlandırma kuralı olarak hata değeri .25 ve madde kullanım sıklığı kontrol yöntemi ise "randomesque = 5"tir.

Şekil 18

"randomesque = 5" ve Hata Değeri .25 Olan Koşulların Ortalama RMSE, Yanlılık ve Test Uzunluğu Değerleri



"randomesque = 5" ve hata değeri .25 olan 8 koşulun yetenek düzeylerine göre 10 gruba ayrılmış bireylerdeki ortalama RMSE, yanlılık ve test uzunluğu performansları EK-İ, EK-J ve EK-K'deki tablolardan ve EK-L'de verilen grafiklerden ayrıntılı incelenebilir. Ayrıca EK-M'de *Eb25r5* ve *Em25r5*'in ikili olarak karşılaştırıldığı grafikler de verilmiştir.

BBT Bulguları

Gerçek zamanlı BBT uygulamasında BBT kuralları olarak hangi koşulların kullanılacağı post-hoc simülasyonlar sonucu belirlendikten sonra Concerto platformu ile BBT uygulaması geliştirilmiştir. Toplam 56 kişiden toplanan verilerin sonucunda elde edilen

bulgulara göre öğrencilerin VST'nin kâğıt-kalem versiyonundan aldıkları toplam puanlar (KTK'ya göre kestirilen) ile BBT versiyonu sonucunda kestirilen yetenek düzeyleri arasındaki korelasyon katsayısı .82 (%95 GA = .72, .89) olarak hesaplanmıştır. VST'nin kâğıt-kalem versiyonuna verdikleri cevaplar ile 3PLM'ye göre yine EAP yöntemi kullanılarak kestirilen yetenek düzeyleri (MTK'ya göre kestirilen) ile BBT versiyonu sonucunda kestirilen yetenek düzeyleri arasındaki korelasyon katsayısı ise .83 (%95 GA = .73, .90) olarak bulunmuştur.

Tablo 16

BBT İstatistikleri

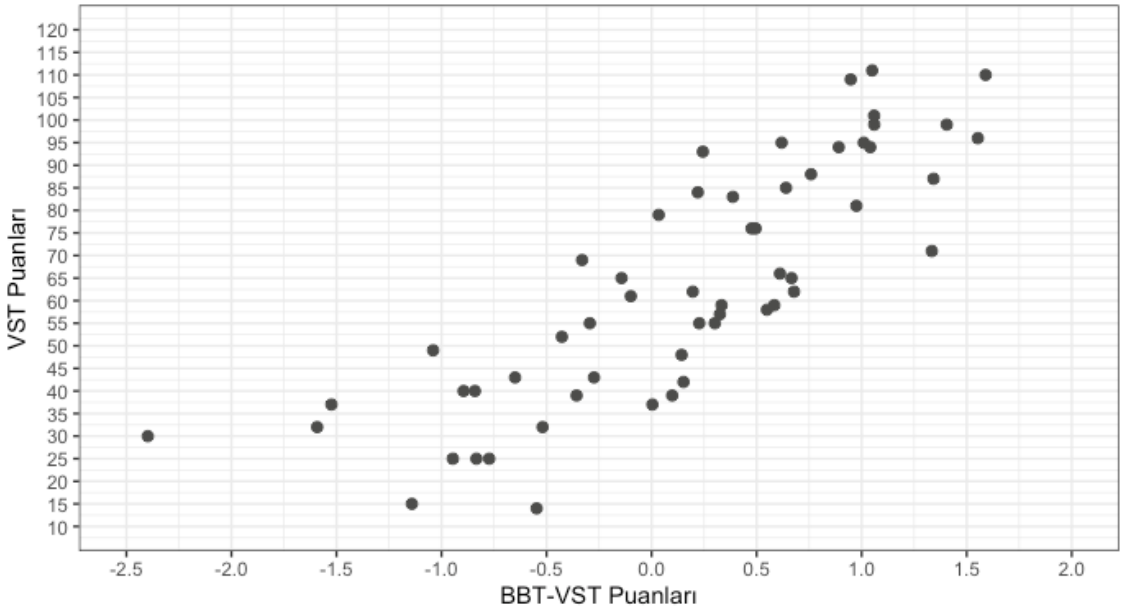
ID	VP	VT	BT	VS	BS	F	TU	ID	VP	VT	BT	VS	BS	F	TU
Ö1	111	1.14	1.05	1	8	-7	10	**Ö29**	61	-0.11	-0.10	29	37	-8	11
Ö2	110	1.11	1.59	2	1	1	18	**Ö30**	59	-0.18	0.58	31	20	11	9
Ö3	109	1.06	0.95	3	12	-9	10	**Ö31**	59	-0.21	0.33	33	25	8	8
Ö4	101	0.86	1.06	5	7	-2	11	**Ö32**	58	-0.19	0.55	32	21	11	9
Ö5	99	0.83	1.06	4	3	1	10	**Ö33**	57	-0.08	0.33	28	26	2	10
Ö6	99	0.89	1.41	7	6	1	16	**Ö34**	55	-0.32	-0.29	35	27	8	12
Ö7	96	0.86	1.55	6	2	4	19	**Ö35**	55	-0.32	0.30	36	29	7	9
Ö8	95	0.73	0.62	10	10	0	7	**Ö36**	55	-0.36	0.23	34	40	-6	19
Ö9	95	0.72	1.01	9	18	-9	9	**Ö37**	52	-0.44	-0.43	38	43	-5	13
Ö10	94	0.74	0.89	11	9	2	7	**Ö38**	49	-0.41	-1.04	37	52	-15	13
Ö11	94	0.66	1.04	8	13	-5	9	**Ö39**	48	-0.47	0.14	39	33	6	11
Ö12	93	0.57	0.24	13	28	-15	11	**Ö40**	43	-0.77	-0.65	40	39	1	12
Ö13	88	0.57	0.76	14	14	0	7	**Ö41**	43	-0.62	-0.27	41	46	-5	12
Ö14	87	0.62	1.34	12	4	8	13	**Ö42**	42	-0.77	0.15	42	32	10	9
Ö15	85	0.48	0.64	16	17	-1	9	**Ö43**	40	-0.86	-0.84	44	49	-5	14
Ö16	84	0.50	0.22	15	30	-15	9	**Ö44**	40	-1.14	-0.90	50	50	0	12
Ö17	83	0.45	0.39	18	24	-6	10	**Ö45**	39	-0.92	0.10	45	34	11	10
Ö18	81	0.47	0.98	17	11	6	9	**Ö46**	39	-0.78	-0.36	43	42	1	16
Ö19	79	0.34	0.03	19	35	-16	10	**Ö47**	37	-1.05	0.00	47	36	11	9
Ö20	76	0.25	0.50	21	22	-1	9	**Ö48**	37	-1.06	-1.52	49	54	-5	23
Ö21	76	0.32	0.48	20	23	-3	11	**Ö49**	32	-0.94	-0.52	46	44	2	13
Ö22	71	0.23	1.33	22	5	17	13	**Ö50**	32	-1.06	-1.59	48	55	-7	22
Ö23	69	0.19	-0.33	23	41	-18	12	**Ö51**	30	-1.52	-2.40	53	56	-3	35
Ö24	66	0.10	0.61	24	19	5	8	**Ö52**	25	-1.25	-0.83	54	47	7	12
Ö25	65	0.02	-0.14	25	16	9	12	**Ö53**	25	-1.57	-0.77	51	48	3	12
Ö26	65	0.05	0.67	26	38	-12	8	**Ö54**	25	-1.50	-0.95	52	51	1	12
Ö27	62	-0.02	0.20	30	15	15	9	**Ö55**	15	-2.45	-1.14	56	53	3	13
Ö28	62	-0.13	0.68	27	31	-4	10	**Ö56**	14	-2.13	-0.55	55	45	10	16

Tablo 16'da BBT uygulamasına katılan öğrencilerin sırasıyla öğrenci ID'leri (ID), VST toplam puanları (VP), VST cevaplarından kestirilen yetenek düzeyleri (VT), BBT-VST

uygulamasında kestirilen yetenek düzeyleri (BT), VT'ye göre yüksekten düşüğe doğru sıra değerleri (VS), BT'ye göre yüksekten düşüğe doğru sıra değerleri (BS), iki sıra değeri arasındaki fark (F) ve test uzunluğu (TU) istatistikleri verilmiştir.

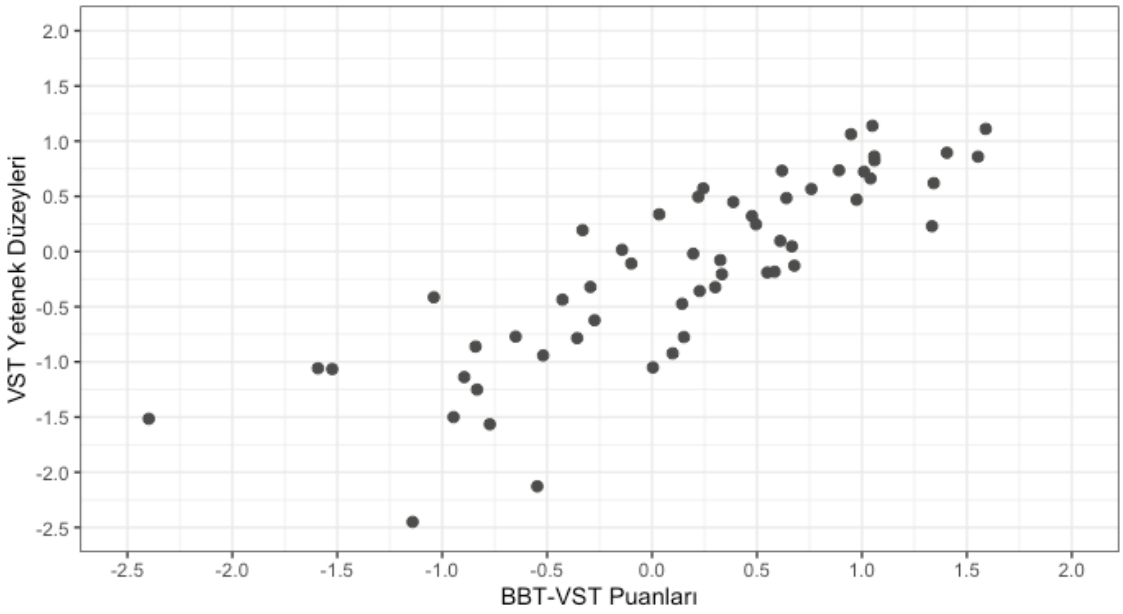
Şekil 19

VST ve BBT-VST Puanları Arasındaki İlişki



Şekil 20

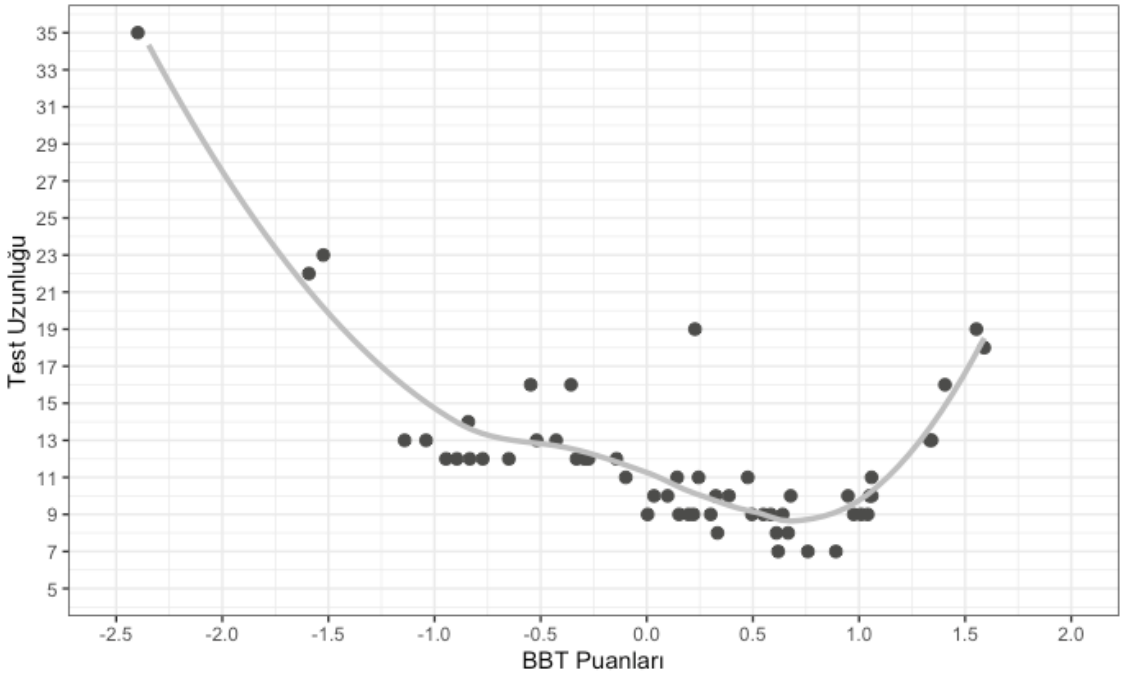
VST Yetenek Düzeyleri ve BBT-VST Puanları Arasındaki İlişki



Şekil 19'da öğrencilerin VST toplam puanları ile BBT-VST puanları arasındaki ilişkiyi gösteren saçılım grafiği verilmiştir. Şekil 20'de ise bireylerin tüm madde havuzuna verdikleri cevaplardan kestirilen yetenek düzeyleri ile BBT-VST puanları arasındaki ilişkiyi gösteren grafik verilmiştir. Aradaki yüksek ilişki her iki grafikten de gözlenebilmektedir. VST puanlarının ortalaması 63.59, VST yetenek düzeylerinin ortalaması -0.16 ve BBT-VST puanlarının ortalaması ise 0.15 olarak hesaplanmıştır. Önceden de belirtildiği gibi VST puanları ve BBT-VST puanlarına ait histogram ve yoğunluk grafikleri EK-A ve EK-B'de verilmiştir. VST cevaplarından kestirilen yetenek düzeylerine ait histogram ve yoğunluk grafikleri ise EK-N'de verilmiştir.

Şekil 21

BBT Puanları ve Test Uzunluğu Arasındaki İlişki



Şekil 21'de bireylerin BBT puanları ile BBT uygulamasında karşılaştıkları madde sayısı arasındaki ilişkiyi gösteren grafik verilmiştir. BBT uygulamasında ortalama test uzunluğu 12 olmuştur. En az uygulanan madde sayısı 7 olurken, en çok uygulanan madde sayısı 35'tir. Yetenek düzeyinin her iki ucunda daha fazla madde uygulanmıştır.

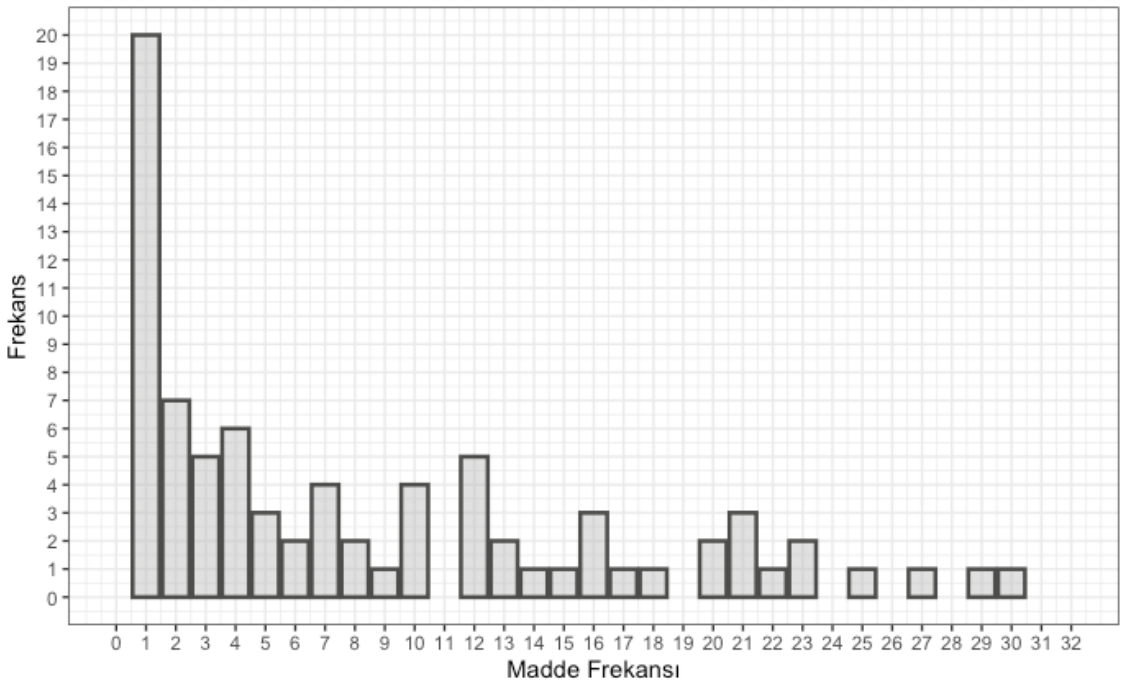
Tablo 17

Madde Kullanım Sıklığı

Madde	F	Sıklık	Madde	F	Sıklık	Madde	F	Sıklık	Madde	F	Sıklık
M2	1	0.02	M42	17	0.30	M74	20	0.36	M107	12	0.21
M3	1	0.02	M43	16	0.29	M75	20	0.36	M108	2	0.04
M5	1	0.02	M46	4	0.07	M76	16	0.29	M109	5	0.09
M7	1	0.02	M47	3	0.05	M79	27	0.48	M110	4	0.07
M9	1	0.02	M48	7	0.13	M80	29	0.52	M114	9	0.16
M10	1	0.02	M50	10	0.18	M82	30	0.54	M115	1	0.02
M13	1	0.02	M51	10	0.18	M83	2	0.04	M116	5	0.09
M17	1	0.02	M53	18	0.32	M84	1	0.02	M117	2	0.04
M19	1	0.02	M54	3	0.05	M85	13	0.23	M118	12	0.21
M21	1	0.02	M56	16	0.29	M87	2	0.04	M119	2	0.04
M22	1	0.02	M57	21	0.38	M88	25	0.45	M120	1	0.02
M25	1	0.02	M60	6	0.11	M89	7	0.13	M121	4	0.07
M27	1	0.02	M61	10	0.18	M94	23	0.41	M123	1	0.02
M29	15	0.27	M62	21	0.38	M95	3	0.05	M125	1	0.02
M31	1	0.02	M64	6	0.11	M96	7	0.13	M126	12	0.21
M34	4	0.07	M65	21	0.38	M97	22	0.39	M127	2	0.04
M35	4	0.07	M67	2	0.04	M101	8	0.14	M128	1	0.02
M36	3	0.05	M70	12	0.21	M103	14	0.25	M129	4	0.07
M38	3	0.05	M72	8	0.14	M104	10	0.18	M132	23	0.41
M40	7	0.13	M73	12	0.21	M105	13	0.23	M134	5	0.09

Şekil 22

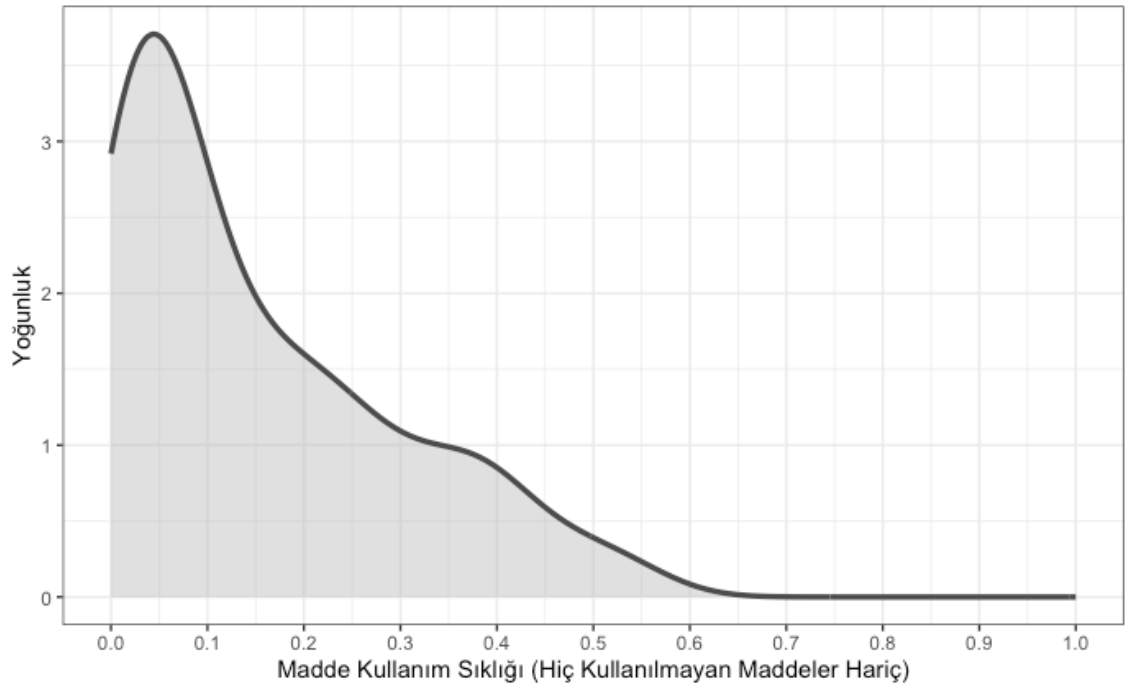
BBT-VST Maddelerinin Uygulanma Sayısı ve Bu Sayıya Ait Frekanslar



140 maddelik havuzdan 80 madde uygulanırken, 60 madde hiç uygulanmamıştır. Tablo 17'de uygulanan 80 maddeye ait frekans ve madde kullanım sıklığı oranları verilmiştir. Şekil 22'de de frekanslar grafikle gösterilmiştir. Gösterim kolaylığından dolayı hiç kullanılmayan maddelere ait frekanslar grafikte de gösterilmemiştir. Tabloda ve şekilde gösterilen bulgulara göre 20 madde sadece 1 kere uygulanmıştır. En sık uygulanan maddeler 80. ve 82. maddeler olmuştur. Bu maddelerin madde kullanım sıklığı oranı sırasıyla .52 ve .54'tür. Yani 80. madde BBT uygulamasına katılan öğrencilerin %52'sine uygulanırken, 82. madde öğrencilerin %54'üne uygulanmıştır.

Şekil 23

Madde Kullanım Sıklığına Ait Yoğunluk Grafiği



Şekil 23'te madde kullanım sıklığına ait yoğunluk grafiği verilmiştir. Bu yoğunluk grafiğinde de hiç kullanılmayan maddeler dahil edilmemiştir. Görüldüğü üzere yine de oldukça çarpık bir dağılım vardır. Bu dağılama MFI yönteminin oldukça seçici olması ve bu çalışmada kullanılan madde kullanım sıklığı kontrol yönteminin bu seçiciliğe yeterince çözüm getirememesi başlıca sebep olabilir. Diğer bir sebep de ortalama test uzunluğunun

düşük olmasıdır. Ortalama 12 madde ile test uzunluğunda %91'lik bir azalma vardır. Bu da kullanılan madde sayısını ve frekanslarını düşürmüştür.

Madde kullanım sıklığının dengeli bir şekilde dağılıp dağılmadığını belirleyebilmek için Hsu ve Wang (2022), çalışmalarında bir ki-kare değerinin hesaplanmasını tavsiye etmişlerdir. Yapılan hesap sonucu elde edilen daha küçük bir ki-kare değerinin, madde kullanım sıklığı oranlarının daha tekdüze (uniform) bir dağılım gösterdiğine ve madde havuzundan daha fazla madde kullanıldığına işaret ettiğini belirtmişlerdir. Ki-kare değerinin hesaplanmasında kullanılan formül, eşitlik 10'da verilmiştir. Eşitlikte r_j, j maddesine ait madde kullanım sıklığı oranını, $\bar{r}$ ise madde havuzundaki tüm maddelere ait madde kullanım sıklığı değerlerinin ortalamasını göstermektedir.

$$\mathrm{X}^2(r) = \sum_{j=1}^{J} \frac{(r_j - \bar{r})^2}{\bar{r}} \qquad \text{…eşitlik 10}$$

Bu değer, yapılan post-hoc simülasyon çalışmasında her bir koşul için hesaplanmış ve EK-I'da verilen simülasyon bulgularında "CHI_ER" isimli sütunda verilmiştir. Aynı tablodaki "MER" isimli sütun madde havuzundaki maddeler için hesaplanan en düşük kullanım sıklığı oranını (minimum exposure rate), "IWMER" ise en düşük orana sahip kaç tane madde olduğunu (number of items with minimum exposure rate) göstermektedir. Çalışmada BBT uygulamasının kuralları, Em25r5 koşuluna göre oluşturulmuştu. Tablodan bu koşula ait simülasyon bulgularına bakıldığında en düşük kullanım sıklığı oranının sıfır olduğunu ve bu orana sahip toplam 29 madde olduğu görülmektedir. Yani madde havuzundan 29 madde hiç kullanılmamıştır. Bu koşul için hesaplanan ki-kare değeri ise 24.75'tir. Aynı hesaplama, BBT uygulamasında kullanılan maddelere ait kullanım sıklığı oranları kullanılarak yapılmıştır. BBT uygulaması için ki-kare değeri 27.96 olarak bulunmuştur ve bilindiği üzere de hiç kullanılmayan madde sayısı 60'tır. Simülasyon çalışmasında elde edilen daha küçük ki-kare değeri, madde kullanım sıklığı oranlarının daha tekdüze dağıldığını göstermektedir. Daha küçük "CHI_ER" ve "IWMER" değerlerinin

sebebi simülasyon çalışmasında simüle edilen birey sayısının (n = 1287) gerçek zamanlı BBT uygulamasındaki birey sayısından (n = 56) oldukça fazla olması olabilir. randomesque yönteminin madde kullanım sıklığını kontrol etmede pek etkili olmasa bile hiçbir şekilde olumlu bir katkısının olmadığı da söylenemez. Konuyu daha detaylı ele almak için için Tablo 18 oluşturulmuştur. Tabloda gerçek zamanlı BBT uygulaması ve BBT uygulamasında kullanılan yetenek kestirim yöntemi, madde seçim yöntemi ve sonlandırma kuralının aynısını içeren simülasyon koşullarına ait ortalama test uzunluğu (OTL), korelasyon (KOR), RMSE, yanlılık, MER, IWMER ve CHI_ER değerleri verilmiştir. Koşullar arasındaki tek fark randomesque yönteminde belirlenen madde sayısıdır. Bilindiği üzere “Em25” koşulunda madde kullanım sıklığı kontrol edilmemektedir, yani randomesque değeri 1’e eşittir. “Em25r3” koşulunda bu değer 3’e, “Em25r5” koşulunda ise 5’e eşittir.

Tablo 18

BBT Uygulaması ve İlgili Üç Koşulun Karşılaştırılması

Koşullar	OTL	KOR	RMSE	Yanlılık	MER	IWMER	CHI_ER
Em25	12.23	0.9635	0.2672	0.0093	0	41	33.57
Em25r3	12.57	0.9642	0.2643	0.0089	0	38	28.42
Em25r5	12.96	0.9640	0.2649	-0.0016	0	29	24.75
BBT-VST	12.00	0.8313	0.3335	-0.3083	0	60	27.96

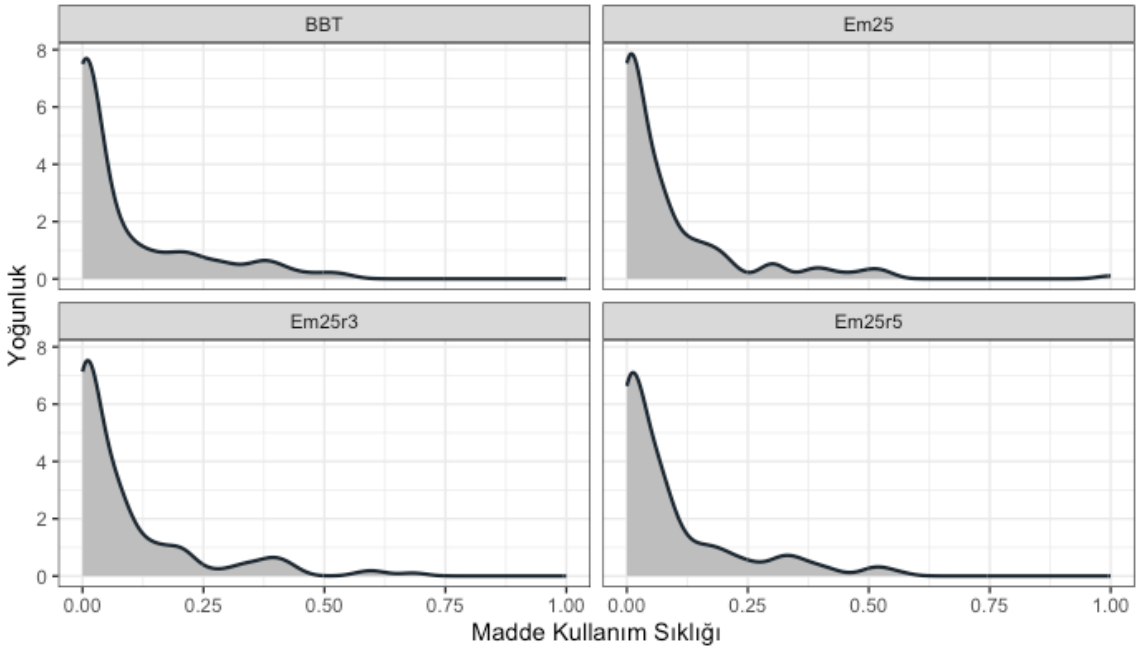
Bilindiği üzere Tablo 18’deki koşulların hepsinde yetenek kestirim yöntemi EAP, madde seçim yöntemi MFI, testi sonlandırma kuralı standart hatanın .25 veya daha düşük bir değer almasıdır ve BBT-VST, Em25r5 ile aynı BBT kurallarına sahiptir. Önceden de bahsedildiği gibi simülasyon koşullarındaki korelasyon, RMSE ve yanlılık değerleri madde kullanım sıklığı kontrol yönteminin aldığı değere göre önemli bir değişim göstermemektedir.

Madde kullanım sıklığı oranının dağılımının tekdüzeliğine dair hesaplanan ki-kare değerlerine bakıldığında ise, randomesque yönteminin aldığı değer arttıkça, ki-kare değerinin giderek azaldığı görülmektedir. Bu da göstermektedir ki randomesque yöntemi, madde kullanım sıklığının dağılımındaki çarpıklığı çok engelleyemese de sıklık oranlarının daha tekdüze bir dağılım göstermesini sağlamaktadır. randomesque değeri 1 olduğunda,

yani madde seçimi işleminde anlık yetenek kestirimine göre en uygun madde seçildiğinde, madde havuzundan 41 madde hiç kullanılmamıştır. Hiç kullanılmayan madde sayısı randomesque değeri arttıkça azalmıştır. Em25r5 koşulunda hiç kullanılmayan madde sayısı 29'a düşmüştür.

Şekil 24

Madde Kullanım Sıklığı Oranlarına Ait Yoğunluk Grafiklerinin Karşılaştırılması



Şekil 24'te, Tablo 18'de verilen üç koşul ve gerçek zamanlı BBT uygulamasındaki madde kullanım sıklığı oranlarının yoğunluk grafikleri verilmiştir. Görüldüğü üzere randomesque değeri arttıkça dağılımın daha da tekdüzeleşmesi grafikten pek anlaşılamamaktadır. Bu sebeple randomesque yönteminin madde havuzunu dengeli kullanmada etkili olup olmadığını inceleme aşamasında grafikten ziyade ki-kare değerlerine bakmak daha doğru olacaktır. Grafiklerde belli olmasa da, her bir koşul için hesaplanan ki-kare değerlerine bakıldığında randomesque yönteminin madde kullanım sıklığını kontrol etmede faydalı olduğu gözlemlenmektedir.

Bölüm 5

Sonuç, Tartışma ve Öneriler

Sonuç ve Tartışma

Bu çalışmada ölçme kesinliğinden ödün vermeden İngilizce kelime bilgisinin bireyselleştirilmiş bilgisayarlı test uygulamaları ile ne derece ölçülebildiği araştırılmıştır. Bu amaçla öncelikle BBT uygulamasında madde havuzunu oluşturacak İngilizce kelime bilgisi testine (VST) ait geçerlik kanıtları, Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014) ilkeleri temel alınarak araştırılmıştır. VST ile ilk önce 1622 kişilik lisans ve lisansüstü eğitimleri devam eden bir öğrenci grubundan veri toplanmıştır. Geçerlik çalışmaları bu veri ile gerçekleştirilmiştir.

Geçerlik analizlerinde en başta yapı geçerliğine ilişkin kanıtlara bakılmıştır. Bu amaçla öncelikle boyutluluk ve yerel bağımsızlık incelemeleri yapılmış, sonrasında verinin hangi madde tepki kuramı (MTK) modeli ile en iyi uyum gösterdiği analiz edilmiş ve son olarak da VST maddeleri içinde değişen madde fonksiyonu (DMF) gösteren maddeler olup olmadığı araştırılmıştır. Yapılan boyutluluk analizlerinde baskın bir boyut olduğu gözlemlenmiştir. Açımlayıcı faktör analizi sonucunda hesaplanan özdeğerlere bakıldığında ilk boyutun ikinci boyuttan yaklaşık olarak altı kat daha büyük bir özdeğere sahip olduğu görülmüştür. Bu bulguya rağmen tek boyutlu yapıyı destekleyen kanıtları arttırmak amacıyla tek boyutlu model ve çok boyutlu modellerden (iki, üç ve dört boyutlu) hangisinin veriye daha iyi uyum gösterdiği araştırılmıştır. Boyut sayısı arttıkça model uyum indekslerinin daha da iyileştiği, yani modelin veriye daha iyi uyum gösterdiği gözlemlenmiştir. Fakat çok boyutlu modellerde her boyut altında yük veren maddelerin güçlük parametreleri ayrıca incelendiğinde, bu çok boyutlu bulguların sebebinin *güçlük faktörleri* olduğu anlaşılmıştır. Güçlük faktörleri yapay bir şekilde oluşan sahte faktörler olduğu için bu bulgunun da VST'nin tek boyutluluğuna bir gösterge olduğu kabul edilmiştir. VST'ye ait alan yazında ilk ve tek geçerlik çalışması olan Beglar'ın (2010) Rasch temelli çalışmasında da VST, bu çalışmada olduğu gibi, tek boyutlu olarak bulunmuştur. Ayrıca alan yazında, İngilizce yazma, okuma,

dinleme ve konuşma gibi beceriler hakkında yapılan boyutluluk çalışmalarının bulgularında herhangi bir ittifak görünmemektedir. Yani çalışmaların yaklaşık yarısı bu becerileri tek boyutlu bir yapı olarak bulurken, diğer yarısı ise çok boyutlu olarak bulmaktadır. Fakat İngilizce kelime bilgisine ait boyutluluk çalışmalarında ise büyük oranda tek boyutlu yapıya ait bulguların varlığı görülmektedir (Min & Aryadoust, 2021). Bu noktadan da bu çalışmanın bulgularının genel bulgulara bir paralellik gösterdiği söylenebilir.

Yerel bağımsızlığı test etmek için Yen'in (1993) Q3 istatistiğine bakılmıştır. Kesme noktasından yüksek Q3 değerine sahip madde çiftleri incelenmiştir. Yerel bağımlılığa neden olabilecek benzer soru kökü ya da benzer seçenekler gibi herhangi bir durumla karşılaşılmamıştır. Alan yazında yerel bağımsızlığının ihlalinin olası farklı sebepleri araştırılmıştır. Ackerman'ın bir çalışmasında (1987) yerel bağımsızlığa yol açabilecek sebeplerde test maddelerinin madde güçlüğüne göre sıralanmasını da saydığı görülmüştür. Bu çalışmada da VST maddeleri kolaydan zora doğru sıralandığı için bazı maddelerde kesme değerinin üzerinde Q3 istatistiklerinin bulunmasının sebebinin bu durum olabileceği değerlendirilmiştir. Olası yerel bağımlı madde çiftleri incelendiğinde de, çoğunlukla birbirine komşu maddeler olduğu gözlemlenmiştir.

Boyutluluk ve yerel bağımsızlığın test edilmesinden sonra madde havuzumuzu oluşturacak VST'nin maddelerinin hangi MTK modeli ile kalibre etmenin daha uygun olduğu araştırılmıştır. Bu amaçla olabilirlik oran testi ile öncelikle bir parametreli lojistik model (1PLM) ile iki parametreli lojistik modele (2PLM) ait kestirimler karşılaştırılmıştır. 2PLM'nin veriye daha iyi uyum sağladığı görüldükten sonra, aynı karşılaştırma 2 PLM ile üç parametreli lojistik model (3PLM) arasında da yapılmıştır. 3PLM daha iyi uyum indeksleri üretmiştir. Bu sebeple VST'nin maddeleri 3PLM'ye göre kalibre edilmiştir. 3PLM'nin daha iyi uyum göstermesinden ayrıca anlaşılmaktadır ki, VST maddelerini doğru cevaplamada şans başarısı önemli bir etkendir. Bu bulgumuz, VST maddelerini doğru cevaplamada şans başarısının önemli bir etken olduğunu gösteren diğer çalışmalar (Stewart, 2014; Stoeckel ve Sukigara, 2018; Zhang, 2013) ile aynı doğrultudadır.

DMF analizleri Lojistik regresyon, Lord'un ki-kare testi ve Mantel-Haenszel yöntemleri kullanılarak yapılmıştır. Her üç yönteme göre DMF gösteren madde sayısı 34 olarak bulunmuştur. Fakat Lojistik regresyon ve Lord'un ki-kare yöntemlerinden elde edilen bulgulara göre bu 34 madde içinde önemli düzeyde DMF gösteren herhangi bir madde yoktur. Sadece Mantel-Haenszel yöntemine göre 10 maddenin önemli düzeyde DMF gösterdiği görülmüştür. Bu bulgu diğer yöntemler tarafından desteklenmediği için bu maddeler için uzman görüşü almak gibi ileri birtakım işlemler yapılmamıştır. Alan yazında VST'nin DMF analizi sadece Beglar (2010) tarafından yapılmıştır. Fakat Beglar, DMF analizlerini sadece ilk 80 soru ile yapabilmiştir. O çalışmada 80 soru içinden 2 maddenin DMF gösterdiği belirtilmiştir, fakat DMF'nin önem derecesi hakkında bir bilgi verilmemiştir.

VST'nin kapsam geçerliğine dair kanıt elde etmek için ise birey-madde haritası oluşturulmuş ve VST'de yeterli sayıda madde olup olmadığını ve bunların yetenek ölçeğinde eşit düzeyde dağılıp dağılmadığı kontrol edilmiştir. Sonuçlar, VST maddelerinin yetenek ölçeğinin neredeyse her düzeyinde yer alan çok çeşitli güçlük parametrelerine sahip olduğunu, yani VST'nin yüksek yeterliliğe sahip öğrencileri düşük yeterliliğe sahip olanlardan ayırdığını ve her θ seviyesi için uygun sayıda soruya sahip olduğunu göstermiştir. Fakat, birey-madde haritası ayrıntılı incelendiğinde, bazı kelimelere ait güçlük parametrelerinin İngilizce metinlerde kullanım sıklığı nispeten daha fazla olan maddelerin güçlük parametrelerine göre daha düşük değerler aldığı görülmüştür. Yani İngilizce metinlerde daha az geçen bazı maddelerin sorulduğu sorular, daha zor maddeler olması gerekirken, daha kolay sorular olarak çalışmışlardır. Bu maddelere tek tek bakıldığında, bu maddelerde anlamı sorulan kelimenin çoğunun *quiz, cube, yoghurt, yoga, puma*, gibi alıntı ve ortak kelimeler (hem İngilizcede hem de Türkçede olan kelimeler) olduğu görülmüştür. VST'den elde edilen toplam puan 100 ile çarpılarak testi alan bireyin İngilizce kelime bilgisi hesaplanır (Nation & Beglar, 2007). Bu gibi alıntı ve ortak kelimelerin fazlalığı bireylerin İngilizce kelime bilgisinin olduğundan yüksek kestirilmesine, yani aşırıtahmin (overestimation) sorununa neden olmaktadır (Harris, 2019; Stewart, 2014). VST'nin

Türkçeden başka diğer dillerde de alıntı ve ortak kelimeler sebebiyle (Elgort, 2013; Harris, 2019; Karami, 2012; Zhao & Ji, 2018) aşırı tahmin gibi problemlere yol açtığı görülmüştür.

Uyum geçerliğine kanıt olarak VST sonuçlarının iki dil sınavı (TOEFL ve YDS) ile olan ilişkisi araştırılmıştır. VST sonuçlarının TOEFL ile .60 (%95 GA = .49, .69) korelasyon gösterirken, YDS sonuçları ile .53 (%95 GA = .45, .60) korelasyon gösterdiği bulunmuştur. Buna göre VST puanlarının İngilizce yeterliği ölçen iki önemli test ile gösterdiği pozitif yönlü anlamlı ilişki, VST'nin uyum geçerliğine dair kanıt oluşturmaktadır. Mclean vd. de (2014) VST puanları ile TOEFL puanları arasında bu çalışmada bulunan değere yakın bir korelasyon bulmuştur ($r = .58$, $p < .001$). Drummond (2018) ise çalışmasında IELTS (The International English Language Testing System) puanları ile VST puanları arasındaki ilişkiye bakmıştır ve daha yüksek bir korelasyon değeri bulmuştur ($r = .68$, $p < .01$). Bu bulgulara göre İngilizce öğrenen bireylerin VST'den aldıkları puanların onların İngilizce yeterlikleri hakkında önemli ölçüde bilgi verebileceği söylenebilir. Korelasyonun nedenselliğe bir delil olmadığı malumdur fakat alan yazında İngilizce kelime bilgisinin dil becerileri üzerinde önemli düzeyde etkisi olduğunu bulan ve bu bulgularına yapısal eşitlik modeli ya da yol analizi kullanarak ulaşan çalışmalar (Li & Zhang, 2019; Vafaee & Suzuki, 2020; Vandergrift & Baker, 2015; Wallace & Lee, 2020) da vardır.

VST'nin çeşitli geçerlik kanıtlarının elde edilmesinden sonra, en uygun BBT koşullarının belirlenebilmesi amacı ile post-hoc simülasyonlar yapılmıştır. Dört farklı yetenek kestirim yöntemi (BM, ML, WL ve EAP), iki madde seçim yöntemi (MFI ve bOpt), üç sonlandırma kuralı ($SH \leq .20$, $SH \leq .25$ ve $SH \leq .30$) ve üç de madde kullanım sıklığı kontrol yöntemi (randomesque = 1, randomesque = 3 ve randomesque = 5) olarak toplam 72 koşulda simülasyonlar gerçekleştirilmiştir.

Alan yazında ikili puanlanan maddeler ile çalışılırken *randomesque* yöntemi kullanıldığında ölçme kesinliğinde azalmanın olduğu çalışmalar olduğu gibi (Moyer ve diğerleri, 2012), ölçme kesinliğinde önemli düzeyde herhangi bir değişimin olmadığı çalışmalar (Leroux ve diğerleri, 2013) da bulunmaktadır. Bu çalışmada ise randomesque yönteminin aldığı değere bağlı olarak ölçme kesinliğinde önemli farklar hesaplanmamıştır.

Bu sebeple randomesque = 1 ve randomesque = 3 olan koşullar, uygun BBT koşullarını belirleme sürecinde elenmiştir. Geriye madde kullanım sıklığı koşulu randomesque = 5 olan 24 koşul kalmıştır. Bu koşullar içinde sonlandırma kuralı olarak hata değeri .20 olan koşulların çok sayıda madde kullanması, hata değeri .30 olan koşulların ise çok az madde kullanması sebebiyle bu iki sonlandırma kuralını içeren koşullar da elenmiştir. Geriye sonlandırma kuralı olarak hata değeri .25 ve madde kullanım sıklığı kontrol yöntemi randomesque = 5 olan ve sadece yetenek kestirim ve madde seçim yöntemi olarak farklılaşan sekiz koşul kalmıştır. Bu sekiz koşul içinden de ölçme kesinliği olarak en iyi değerlerden birini üreten ve bu değerleri daha az madde ile sağlayan Em25r5 koşulu gerçek zamanlı BBT uygulaması için en uygun BBT kurallarını içeren koşul olarak seçilmiştir. Yani simülasyon çalışmasının bulgularına göre yetenek kestirim yöntemi olarak EAP, madde seçim yöntemi olarak MFI, sonlandırma kuralı olarak hata değeri .25 ve madde kullanım sıklığı kontrol yöntemi olarak da randomesque = 5 en uygun BBT kuralları olarak belirlenmiştir. Alan yazında yetenek kestirim yöntemi ve madde seçim yöntemi için genellikle tavsiye edilen ikili EAP ve MFI olmuştur (Erdem Kara & Doğan, 2022; van der Linden, 2008; van der Linden & Pashley, 2010). Bu çalışmanın bulguları da bu tavsiyeleri desteklemektedir.

Uygun BBT şartlarının belirlenmesinden sonra, gerçek zamanlı BBT uygulaması (BBT-VST) Concerto Platform kullanılarak geliştirilmiştir. Elli altı kişiden oluşan çalışma grubuna hem VST'nin kâğıt-kalem formu hem de yeni geliştirilen BBT-VST herhangi bir sıralama gözetmeden ve ara vermeden peş peşe uygulanmıştır. Kâğıt-kalem formuna verilen cevaplardan kestirilen yetenek parametreleri ile BBT formunda kestirilen yetenek parametreleri arasındaki korelasyon katsayısı .83 (%95 GA = .73, .90) olarak hesaplanmıştır. Bu çalışmada bulunan ilişki katsayısı İngilizce kelime bilgisinin BBT ile ölçüldüğü diğer çalışmalarda elde edilen katsayılar ile kıyaslandığında görülmektedir ki; bu çalışmada hesaplanan korelasyon katsayısı, Kezer ve Koç'un (2014) çalışmasında bulunan değere (r = .86) oldukça yakın çıkarken, Tseng'in (2016) çalışmasında hesaplanan korelasyon değerlerinden küçük bulunmuştur. Tseng'in çalışmasında altı farklı BBT

koşulunda altı farklı korelasyon hesaplanmıştır. Altı koşulun hepsinde de .90'nın üzerinde korelasyon katsayıları elde edilmiştir. Bu çalışmanın madde havuzu 140 maddeden oluşmaktadır ve gerçekleştirilen BBT uygulamasında ortalama test uzunluğu 12 olarak bulunmuştur. Yani %91'lik bir tasarruf gerçekleşmiştir, yani %91 oranında daha az madde kullanılmıştır. Kezer ve Koç'un çalışmasında ise madde havuzu 72 maddeden oluşmuş, BBT uygulamasında ortalama test uzunluğu 16 olarak bulunmuş ve %78'lik bir tasarruf yapılmıştır. Tseng'in çalışmasında ise bu oran, sonlandırma kuralı olarak hata değerinin .30 olduğu koşulda %90, .20 olduğu koşulda ise %76 olarak bulunmuştur. Bilindiği üzere BBT uygulamaları ölçme kesinliğinden ödün vermeden test uzunluğunda ortalama olarak %50 oranında bir azalma sağlar (Weiss, 1985). Hem bu çalışmanın bulgularına hem de İngilizce kelime bilgisini BBT uygulamaları ile ölçen diğer çalışmaların bulgularına bakıldığında bu oranın çok daha yüksek olduğu görülmektedir. Buna göre, BBT uygulamalarının hem zamanda hem de test uzunluğunda sağladığı tasarruf dikkate alındığında İngilizce kelime bilgisini belirleme noktasından çok etkili bir ölçme yöntemi olduğu söylenebilir.

BBT uygulamasında kullanılan madde kullanım sıklığı kontrol yöntemlerinden biri olan *randomesque* yönteminin madde kullanım sıklığının dağılımındaki çarpıklığa engel olamadığı gözlemlenmiştir. 140 maddeden 60 madde hiç kullanılmamış, 20 madde de sadece 1 kere kullanılmıştır. Çok kategorili maddeler ile yapılan çalışmalarda madde kullanım sıklığı kontrol yöntemi olarak randomesque kullanıldığında ölçme kesinliğindeki düşüş önemsiz düzeyde kalmış, madde havuzu daha verimli bir şekilde kullanılmıştır (Davis, 2004; Lee & Dodd, 2012; Leroux ve diğerleri, 2019). Fakat bu çalışmada olduğu gibi randomesque yönteminin ikili puanlanan maddeler için madde kullanım sıklığını kontrol etmede bazen iyi bir performans göstermediği bilinmektedir (Leroux ve diğerleri, 2013). Hsu ve Wang'ın (2022) çalışmalarında da belirttiği üzere, madde kullanım sıklığına ait dağılımın normal değil de tekdüze bir dağılım göstermesi daha istenilir bir durumdur. Tekdüze dağılımın ne oranda gerçekleştiğini belirleyebilmek için Hsu ve Wang'ın (2022) önerdiği ki-kare istatistiği hesaplandığında, randomesque yönteminin dağılımı tekdüzeliğe yaklaştırdığı bulunmuştur.

Öneriler

Bu bölümde öncelikle VST'nin geçerlik çalışmasından, simülasyon çalışmalarından, ve gerçek zamanlı BBT uygulamasından elde edilen sonuçlara göre uygulamaya yönelik öneriler listelenmiştir. Sonrasında ise araştırmaya yönelik önerilerden bahsedilmiştir.

Uygulamaya Yönelik Öneriler

Uygulamaya yönelik öneriler şöyledir:

1. VST'nin boyutluluk analizleri yapıldığında çok boyutlu modellerin veriye daha iyi uyum sağladığı bulunmuştur. Fakat çok boyutlu çıktılarda farklı boyutlarda yer alan maddelerin güçlük parametreleri incelendiğinde, sahte ya da yapay faktörlere sebep olan *güçlük faktörleri* probleminin olduğu görülmüştür. Bu sebeple teorik olarak tek boyutlu olduğu düşünülen ya da alan yazındaki boyutluluk çalışmalarında genellikle tek boyutlu olarak çıktılar veren yapılar ile çalışıldığında çok boyutlu bulgulara denk gelinirse, bu çalışmada olduğu gibi farklı boyutlardaki madde güçlüğü değerleri analiz edilerek *güçlük faktörleri* sorununun olup olmadığının test edilmesi önerilir.

2. Bu çalışmada yerel bağımsızlık varsayımı test edilirken, birbirine komşu maddelerin yerel bağımsızlık istatistiklerinin daha yüksek değerler aldığı gözlemlenmiştir. Ackerman'a (1987) göre yerel bağımlılığın sebeplerinden birisi test maddelerin zordan kolaya ya da kolaydan zora doğru sıralanması olabilir. Bu sebeple ileride VST'nin kullanılacağı araştırmalarda sorular karışık olarak ya da en azından sayfalar karışık olarak verilebilir. Böylece hem olası yerel bağımlığın önüne geçilmiş olunur hem de VST'nin özellikle ikinci yarısında daha zor maddelerin bulunmasından kaynaklı testte ilerledikçe gelen olası motivasyon düşüklüğüne karşı önlem alınmış olunur.

3. VST'de oldukça fazla sayıda alıntı ve ortak kelime (Hem Türkçede hem de İngilizcede bulunan kelime) vardır. VST'de bu alıntı ve ortak kelimelerin sorulduğu maddelerin yerine yeni maddeler eklenmelidir. Çünkü alıntı ve ortak kelimeler VST'nin daha zor olması gerektiği düşünülen seviyelerinde kendine yer bulmaktadır. Fakat testi çözen öğrencilerin ana dilinde de bu kelimeler olduğu için öğrenciler bu soruları çözmekte

zorlanmamaktadır ve öğrencilerin kelime bilgileri bu sebeple olduğundan yüksek kestirilmektedir. Bu durum ayrıca, bu çalışmanın bulgularının da gösterdiği üzere, kolay maddelerin sayısını arttırmakta, zor maddelerin sayısını azaltmaktadır. Bunun neticesinde yetenek ölçeğinin her düzeyinde yeterli sayıda madde bulunmama olasılığı artmaktadır.

4. Madde kullanım sıklığı oranlarına ait dağılımın tekdüze bir dağılım göstermesi, madde havuzunun daha etkin ve dengeli kullanımına işaret etmektedir. Tekdüze bir dağılımı elde etmek de ancak etkili madde kullanım sıklığı kontrolü ile mümkün olmaktadır. Concerto platformu arka planda catR paketini kullanmaktadır ve catR paketi de madde kullanım sıklığı kontrol yöntemi olarak da sadece randomesque yöntemini içermektedir. randomesque yönteminin tekdüze bir dağılımı sağlamada başarısının sınırlı olduğu görülmüştür. catR paketine ve Concerto platformuna randomesque yönteminden başka madde kullanım sıklığı kontrol yöntemlerinin eklenmesi BBT araştırmaları için çok faydalı olacaktır.

5. Bu çalışmada İngilizce kelime bilgisinin bireyselleştirilmiş bilgisayarlı test uygulamaları ile % 91 oranında daha az madde ile güvenilir bir şekilde kestirilebildiği gözlemlenmiştir. Yetenek düzeyinin her iki ucu hariç, düşük, orta ve yüksek İngilizce kelime bilgisine sahip öğrencilerin kelime bilgileri ortalama 12 soru ile ölçülebilmektedir. Daha güvenilir sonuçlar elde edilmek istendiğinde sonlandırma kuralı olarak standart hata değeri .20, hatta .15 gibi çok düşük değerler olacak şekilde belirlenebilir. Fakat çok düşük ya da çok yüksek İngilizce kelime bilgisine sahip bireyler için testin çok sayıda madde uygulandığı halde sonlanmaması ihtimaline göre de, hatanın yanında sabit madde uzunluğu da bir sonlandırma kuralı olarak BBT algoritmasına eklenmelidir.

Araştırmaya yönelik öneriler

Araştırmaya yönelik öneriler şunlardır;

1. Bu çalışmada VST maddelerinin kalibrasyonunda ve BBT uygulamalarında üç parametreli lojistik model kullanılmıştır. Dört parametreli model kullanılarak yapılacak yeni çalışmalar alan yazına katkıda bulunabilir.

2. DMF analizleri lojistik regresyon, Lord'un ki-kare testi ve Mantel-Haenszel yöntemleri ile gerçekleştirilmiştir. Bu çalışmada kullanılan yöntemlerden farklı olarak MTK

temelli ya da MTK temelli olmayan birçok DMF test etme yöntemi vardır ve hala daha yeni yöntemler geliştirilmektedir (Henninger ve diğerleri, 2023; Lim ve diğerleri, 2022). Bu yöntemler ile de VST'nin maddeleri DMF yönünden incelenebilir. Bu çalışmada yanlılık gösterme ihtimali olan maddeler tespit edilmiş, daha da ileriye gidilmemiştir. Fakat ileride gerçekleştirilecek VST'ye yönelik DMF analizlerinde uzman görüşlerine de başvurularak eğer varsa yanlılık gösteren maddelerin tespit edilip, VST'den bu maddelerin çıkarılarak yeni maddeler eklenmesi sağlanabilir.

3. Madde havuzunun dengeli ve etkin kullanımı amacıyla tercih edilen randomesque yöntemi 3 ve 5 değerlerini aldığında tek başına bu amacı gerçekleştiremediği bulunmuştur. Gelecek çalışmalarda randomesque değerinin arttırılması (örn. randomesque = 10) ya da madde kullanım sıklığını kontrol etmede randomesque yönteminin yanında, bu kontrolün farklı madde seçim yöntemleri ile de desteklenerek gerçekleştirilmesi önerilmektedir.

4. Alan yazında VST'den elde edilen puanların şans başarısından dolayı aşırıtahmin problemini içerdiği gösterilmiştir. Bu problemin önüne geçmek için *bilmiyorum* cevabını içeren bir seçenek eklenmesi ya da yanlış cevaba ceza verilmesi gibi yöntemler denenmiştir. VST'nin kâğıt-kalem formunda denenen bu yöntemler aynen BBT versiyonunda da denenebilir. "bilmiyorum" seçenekli ya da yanlış cevaba ceza verilen BBT versiyonu ile orijinal BBT versiyonu sonuçları karşılaştırılabilir.

Kaynaklar

Ackerman, T. A. (1987, April). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum. https://doi.org/10.5040/9781474212151

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

Aybek, E. C. (2016). Concerto: Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulamaları için Bir Platform. *Journal of Measurement and Evaluation in Education and Psychology, 7*(2), 251-271. https://doi.org/10.21031/epod.267198

Baker, F. (2001). *The basics of Item response theory.* ERIC Clearinghouse on Assessment and Evaluation.

Barrada, J.R., Olea, J., Ponsoda, V. & Abad F.J. (2009). Item Selection Rules in Computerized Adaptive Testing: Accuracy and Security. *Methodology, 5*(1), 7–17. https://doi.org/10.1027/1614-2241.5.1.7

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Barnard, John J. (2018). From Simulation to Implementation: Two CAT Case Studies. *Practical Assessment, Research & Evaluation, 23*(14). Available online: http://pareonline.net/getvn.asp?v=23&n=14

Birnbaum A. (1968) Some Latent Trait Models, In Lord F.M., & Novick M.R. (eds.), *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Camilli, G. (1994). Origin of the Scaling Constant d = 1.7 in Item Response Theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293-295. https://doi.org/10.2307/1165298

Cattell, R. B. (1966). The Scree Test for The Number of Factors. *Multivariate Behavioral Research, 1*(2), 245– 276. https://doi.org/10.1207/s15327906mbr0102_10

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19*, 273–299. https://doi.org/10.1017/S0267190599190147

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chang H. H. (2015). Psychometrics behind Computerized Adaptive Testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods, 49*, 502-512. https://doi.org/10.3758/s13428-016-0712-6

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the Parallel Analysis Procedure With Polychoric Correlations. *Educational and Psychological Measurement, 69*(5), 748–759. https://doi.org/10.1177/0013164409332229

Cronbach, L. J. (1990). *Essentials of psychological testing (5th ed.)*. Harper and Row.

Çepni, Z. & Kelecioğlu, H. (2021). Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods. *Journal of Measurement and Evaluation in Education and Psychology, 12*(3), 267-285. https://doi.org/10.21031/epod.988879

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667268

Davey, T., Parshall, C. G. (1995). *New Algorithms for Item Selection and Exposure Control with Computerized Adaptive Testing*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. https://eric.ed.gov/?id=ED421525

Davis, L. L. (2004). Strategies for Controlling Item Exposure in Computerized Adaptive Testing With the Generalized Partial Credit Model. *Applied Psychological Measurement, 28*(3), 165–185. https://doi.org/10.1177/0146621604264133

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

de la Torre, J., & Deng, W. (2008). Improving Person-Fit Assessment by Correcting the Ability Estimate and Its Reference Distribution. *Journal of Educational Measurement, 45*(2), 159–177. http://www.jstor.org/stable/20461887.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Mahwah: Lawrence Erlbaum Associates.

Drummond, A. (2018). Investigating the relationship between IELTS scores and receptive vocabulary size. *Journal of the Foundation Year Network, 1*, 113–125. https://jfyn.co.uk/index.php/ukfyn/article/view/13

Economides, A. A. & Roupas, C. (2007). Evaluation of computer adaptive testing systems. *International Journal of Web-Based Learning and Teaching Technologies, 2*(1), 70–87. https://doi.org/10.4018/jwltt.2007010104

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing, 30*(2), 253–272. https://doi.org/10.1177/0265532212459028

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum. https://doi.org/10.4324/9781410605269

Erdem Kara, B. & Doğan, N. (2022). The Effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education, 9*(3), 682-696. https://doi.org/10.21449/ijate.1105769

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributionsfor item selection in CAT. *Journal of Educational and Behavioral Statistics,37*(5), 655–670. https://doi.org/10.3102/1076998611422912

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Gaertner, K., & McBride, Y. (2017). Detecting unexpected changes in pass rates: A comparison of two statistical approaches. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 262–279). Routledge.

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

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Han, K. T. (2018). Conducting simulation studies for computerized adaptive testing using SimulCAT: an instructional piece. *Journal of Educational Evaluation for Health Professions, 15, 20*. https://doi.org/10.3352/jeehp.2018.15.20

Harris, J. (2019). Issues related to the presence of Japanese loanwords of English origin on vocabulary size tests. *The Asian Journal of Applied Linguistics, 6*(1), 1–13. https://caes.hku.hk/ajal/index.php/ajal/article/view/568

Harrison, C., Loe, B. S., Lis, P., & Sidey-Gibbons, C. (2020). Maximizing the Potential of Patient-Reported Assessments by Using the Open-Source Concerto Platform With Computerized Adaptive Testing and Machine Learning. *Journal of Medical Internet Research, 22*(10). https://doi.org/10.2196/20950

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hau, K., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*(3), 249–266. https://doi.org/10.1111/j.1745-3984.2001.tb01126.x

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing, 2*(2), 141–154. https://doi.org/10.1177/026553228500200203

Henninger, M., Debelak, R., & Strobl, C. (2023). A New Stopping Criterion for Rasch Trees Based on the Mantel–Haenszel Effect Size Measure for Differential Item Functioning. *Educational and Psychological Measurement*, *83*(1), 181–212. https://doi.org/10.1177/00131644221077135

Hertzman, M. (1936). The effects of the relative difficulty of mental tests on patterns of mental organization. *Archives of Psychology, 197*.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). American Psychological Association. https://doi.org/10.1037/10244-014

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Lawrence Erlbaum.

Horn, J. L. (1965). A Rationale and Test for The Number of Factors in Factor Analysis. *Psychometrika, 30*, 179–185. https://doi.org/10.1007/BF02289447

Hsu, C.-L., & Wang, W.-C. (2022). Reducing the Misclassification Costs of Cognitive Diagnosis Computerized Adaptive Testing: Item Selection With Minimum Expected Risk. *Applied Psychological Measurement, 46*(3), 185–199. https://doi.org/10.1177/01466216211066610

Huang, Y. M., Lin, Y.T., Cheng, S.C. (2009). An adaptive testing system for supporting versatile educational assessment. Computers & Education, *52*(1), 53-67. https://doi.org/10.1016/j.compedu.2008.06.007

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education, 16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal, 43*(1), 53–67. https://doi.org/10.1177/0033688212439359

Kaya, E. (2022). *A Comparability and Classification Analysis of Computerized Adaptive and Conventional Paper-Based Versions of an English Language Proficiency Reading Subtest* (Doctoral dissertation). İhsan Doğramacı Bilkent University, Ankara.

Kezer, F. & Koç, N. (2014). A comparison of computerized adaptive testing strategies. *Journal of Educational Sciences Research, 4*(1), 145-174. http://dx.doi.org/10.12973/jesr.2014.41.8

Khoshsima, H. and Toroujeni, S. M. H. (2017). Computer Adaptive Testing (Cat) Design; Testing Algorithm and Administration Mode Investigation. *European Journal of Education Studies, 3*(5), 764-795. https://doi.org/10.5281/zenodo.576047

Kıbrıslıoğlu Uysal, N., & Atalay Kabasakal, K. (2017). The effect of background variables on gender related differential item functioning. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 373-390. https://doi.org/10.21031/epod.333451

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6

Koyuncu, İ., & Kılıç, A. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Education and Science, 44*(198). http://dx.doi.org/10.15390/EB.2019.7665

Kremmel, B. & Pellicer-Sánchez, A. (2021). Measuring vocabulary development. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of SLA and Language Testing* (pp. 211-222). Routledge. https://doi.org/10.4324/9781351034784

Laufer, B., & Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and *Computer Adaptiveness. Language Learning, 54*. 399-436. http://dx.doi.org/10.1111/j.0023-8333.2004.00260.x

Lee, H., & Dodd, B. G. (2012). Comparison of Exposure Controls, Item Pool Characteristics, and Population Distributions for CAT Using the Partial Credit Model. *Educational and Psychological Measurement, 72*(1), 159–175. https://doi.org/10.1177/0013164411411296

Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A Comparison of Exposure Control Procedures in CATs Using the 3PL Model. *Educational and Psychological Measurement, 73*(5), 857–874. https://doi.org/10.1177/0013164413486802

Leroux, A. J., Waid-Ebbs, J. K., Wen, P.-S., Helmer, D. A., Graham, D. P., O'Connor, M. K., & Ray, K. (2019). An Investigation of Exposure Control Methods With Variable-

Length CAT Using the Partial Credit Model. *Applied Psychological Measurement, 43*(8), 624–638. https://doi.org/10.1177/0146621618824856

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2003). Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods. *The Journal of Technology, Learning and Assessment, 2*(5), 3-15. https://ejournals.bc.edu/index.php/jtla/article/view/1665

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*, 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, C. H. (2019). Using a Listening Vocabulary Levels Test to explore the effect of vocabulary knowledge on GEPT listening comprehension performance. *Language Assessment Quarterly, 16*(3), 328–344. https://doi.org/10.1080/15434303.2019.1648474

Li, Y. & Zhang, X. (2019). L2 Vocabulary Knowledge and L2 Listening Comprehension: A Structural Equation Model. *Canadian Journal of Applied Linguistics 22*(1). https://doi.org/10.7202/1060907ar

Lim, H., Choe, E.M. & Han, K.T. (2022), A Residual-Based Differential Item Functioning Detection Framework in Item Response Theory. *Journal of Educational Measurement, 59*: 80-104. https://doi.org/10.1111/jedm.12313

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In S. Chea, U. Kang, & J. M. Linacre (Eds.), *Development of computerized middle school achievement test*. Komesa Press.

Linacre, J. M. (2007). A user's guide to WINSTEPS. Chicago: winsteps.com.

Lord, F. M. (1952*). A theory of test scores*. Psychometric Monographs, 7, x, 84.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*(2), 247–264. https://doi.org/10.1007/BF02291471

Lord, F. M. (1980). Application of item response theory to practical testing problems. Erlbaum. https://doi.org/10.4324/9780203056615

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Magis, D., & Barrada, J. R. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software, Code Snippets, 76*(1), 1–19. https://doi.org/10.18637/jss.v076.c01

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software, 48*(8), 1–31. https://doi.org/10.18637/jss.v048.i08

Magis, D., Raîche, G., & Béland, S. (2012). A Didactic Presentation of Snijders's l_z* Index of Person Fit With Emphasis on Response Model Selection and Ability Estimation. *Journal of Educational and Behavioral Statistics*, *37*(1), 57–81. https://doi.org/10.3102/1076998610396894

Magis, D., Yan, D. & von-Davier, A. (Eds.). (2017). *Computerized adaptive and multistage testing with R: Using packages catr and mstr.* Springer. https://doi.org/10.1007/978-3-319-69218-0

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27–45. https://doi.org/10.1037/edu0000205

McClarty, K. L., Sperling, R. A., & Dodd, B. G. (2006). *A variant of the progressive-restricted item exposure control procedure in computerized adaptive testing systems based*

on the 3PL and partial credit models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*(1), 82–99. https://doi.org/10.1111/j.2044-8317.1974.tb00530.x

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' english vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction, 3* (2), 47-55. http://dx.doi.org/10.7820/vli.v03.2.mclean.et.al

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching*, 13(3-4), 221-246. https://doi.org/10.1017/S0261444800008879

Messick S (1989). Validity. In R. L., Linn (Ed), *Educational measurement* (3rd ed) (pp. 13–103). New York: Macmillan

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741. https://doi.org/10.1037/0003- 066X.50.9.741

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Maters. https://doi.org/10.21832/9781847692092

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer. (Eds.), *L2 vocabulary acquisition, knowledge, and use: New perspectives on assessment and corpus analysis* (pp. 57-78). Eurosla Monographs Series. https://www.eurosla.org/monographs/EM02/Milton.pdf

Min, S. & Aryadoust, V. (2021). A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation, 68*. https://doi.org/10.1016/j.stueduc.2020.100963.

Miralpeix, I. & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching, 56*(1), 1-24. https://doi.org/10.1515/iral-2017-0016

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the word part levels test. *Language Testing, 36*(1), 101–123. https://doi.org/10.1177/0265532217725776

Moyer, E. L., Galindo, J. L., & Dodd, B. G. (2012). Balancing flexible constraints and measurement precision in computerized adaptive testing. *Educational and Psychological Measurement, 72*(4), 629-648. https://doi.org/10.1177/0013164411431838

Muthén, B., du Toit, S.H.C. & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished technical report. https://www.statmodel.com/download/Article_075.pdf

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692. https://doi.org/10.1093/biomet/78.3.691

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13. https://jalt- publications.org/tlt/issues/2007-07_31.7

Nation, I. S. P. (2013). *Learning vocabulary in another language (2nd ed.).* Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal, 42*(1), 86–99. https://doi.org/10.1177/0033688210390264

Noreillie, A. S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and

French as foreign languages. *International Journal of Applied Linguistics, 169*(1), 212-231. https://doi.org/10.1075/itl.00013.nor

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [20.11.2021] from www.psych.umn.edu/psylabs/CATCentral/

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Parshall, C.G., Davey, T., & Nering, M.L. (1998). *Test Development Exposure Control for Adaptive Testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA. https://eric.ed.gov/?id=ED421526

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, 25(3), 193–203. https://doi.org/10.1111/j.1745-3984.1988.tb00302.x

Reshetar, R. (1990). *Computer adaptive testing: Development and application* (Laboratory of Psychometric and Evaluative Research Report No. 204). University of Massachusetts, School of Education.

Sari, H. (2020). Testing Multistage Testing Configurations: Post-Hoc vs. Hybrid Simulations . *International Journal of Psychology and Educational Studies, 7*(1), 27–37. https://doi.org/10.17220/ijpes.2020.01.003

Sanz, S., Luzardo, M., García, C., & Abad, F. J. (2020). Detecting Cheating Methods on Unproctored Internet Tests. Psicothema, 32(4), 549–558. https://doi.org/10.7334/psicothema2020.86

Sasao, Y., & Webb, S. (2017). The Word Part Levels Test. *Language Teaching Research, 21*(1), 12–30. https://doi.org/10.1177/1362168815586083

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *The*

British journal of mathematical and statistical psychology, 68(3), 478–496. https://doi.org/10.1111/bmsp.12057

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*(1), 109-120. https://doi.org/10.1017/S0261444819000326

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Segall, D. O. (2004). A Sharing Item Response Theory Model for Computerized Adaptive Testing. *Journal of Educational and Behavioral Statistics, 29*(4), 439–460. https://doi.org/10.3102/10769986029004439

Shin, C. D. (2017). Conditional Randomesque Method for Item Exposure Control in CAT. *International Journal of Intelligent Technologies and Applied Statistics, 10*(3), 145-155. https://doi.org/10.6148/IJITAS.2017.1003.02

Shin, D., Yuehmei, C., Way, D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATS*. Pearson Research Report, Pearson, Iowa, IA.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden: Nieuwsbrief voor Toegepaste Statistiek en Operationele Research, 7*(22), 131-145.

Sijtsma, K., & Meijer, R. R. (1992). A Method for Investigating the Intersection of Item Response Functions in Mokken's Nonparametric IRT Model. *Applied Psychological Measurement, 16*(2), 149–157. https://doi.org/10.1177/014662169201600204

Snijders, T.A.B. Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika 66*, 331–342 (2001). https://doi.org/10.1007/BF02294437

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* MaCmillan.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly, 11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

Stocking, M.L., & Lewis, C. (1995). *A New Method of Controlling Item Exposure in Computerized Adaptive Testing*. ETS Research Report Series, 1995: i-29. https://doi.org/10.1002/j.2333-8504.1995.tb01660.x

Stocking, M. L., & Lewis, C. (1998). Controlling Item Exposure Conditional on Ability in Computerized Adaptive Testing. *Journal of Educational and Behavioral Statistics, 23*(1), 57–75. https://doi.org/10.3102/10769986023001057

Stoeckel, T., & Sukigara, T. (2018). A serial multiple-choice format designed to reduce overestimation of meaning-recall knowledge on the Vocabulary Size Test. *TESOL Quarterly, 52*(4), 1050-1062. https://doi.org/10.1002/tesq.429

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R Package for Person-Fit Analysis in IRT. *Journal of Statistical Software, 74*(5), 1–27. https://doi.org/10.18637/jss.v074.i05

Thompson, N. A., & Weiss, D. J. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research, and Evaluation, 16*(1). 1-9. https://doi.org/10.7275/wqzt-9427

Tian, J., Miao, D., Zhu, X., & Gong, J. (2007). An introduction to the Computerized Adaptive Testing. *US-China Education Review, 4(*1), 72–81. https://eric.ed.gov/?id=ED497385

Tran, U. S., & Formann, A. K. (2009). Performance of Parallel Analysis in Retrieving Unidimensionality in the Presence of Binary Data. *Educational and Psychological Measurement, 69*(1), 50–61. https://doi.org/10.1177/0013164408318761

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education, 97*, 69–85. https://doi.org/10.1016/j.compedu.2016.02.018

Urry, V. W. (1970). *A Monte Carlo Investigation of Logistic Test Models*. (Doktora tezi). Purdue University, West Lafayette, IN.

Uysal, İ., Ertuna, L., Ertaş, F., G. & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 133-148. https://doi.org/10.21031/epod.534312

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition, 42*(2), 383–410. https://doi.org/10.1017/S0272263119000676

van der Linden, W.J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*(1), 5-20. https://doi.org/10.3102/1076998607302626

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

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

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), *Elements of adaptive testing* (p. 1-25). Springer. https://doi.org/10.1007/978-0-387-85461-8

Vandergrift, L., & Baker, S. C. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning, 65*, 390–416. https://doi.org/10.1111/lang.12105

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*(1), 15–20. https://doi.org/10.1111/j.1745-3992.1993.tb00519.x

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., & Mislevy, R.J. (2000*). Computerized Adaptive Testing: A Primer* (2nd ed.). Routledge. https://doi.org/10.4324/9781410605931

Wallace, M. P., & Lee, K. (2020). Examining second language listening, vocabulary, and executive functioning. *Frontiers in Psychology, 11*, Article 1122. https://doi.org/10.3389/fpsyg.2020.01122

Walter, O.B. (2009). Adaptive Tests for Measuring Anxiety and Depression. In: van der Linden, W., Glas, C. (Eds) Elements of Adaptive Testing. Statistics for Social and Behavioral Sciences. Springer. https://doi.org/10.1007/978-0-387-85461-8_6

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774–789. https://doi.org/10.1037/0022-006X.53.6.774

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70. https://doi.org/10.1080/07481756.2004.11909751

Weng, L.-J., & Cheng, C.-P. (2005). Parallel Analysis with Unidimensional Binary Data. *Educational and Psychological Measurement, 65*(5), 697–716. https://doi.org/10.1177/0013164404273941

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing, 2*(1), 1–17. https://doi.org/10.7333/1401-0201001

Wolfe, E. W., & Smith, E. V., Jr (2007a). Instrument development tools and activities for measure validation using Rasch models: part I - instrument development tools. *Journal of Applied Measurement, 8*(1), 97–123.

Wolfe, E. W., & Smith, E. V., Jr (2007b). Instrument development tools and activities for measure validation using Rasch models: part II--validation activities. *Journal of Applied Measurement, 8*(2), 204–234.

Xiao, J., & Bulut, O. (2022). Item Selection With Collaborative Filtering in On-The-Fly Multistage Adaptive Testing. *Applied psychological measurement, 46*(8), 690–704. https://doi.org/10.1177/01466216221124089

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods, 47*(3), 756–772. https://doi.org/10.3758/s13428-014-0499-2

Yen, W.M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zhang, X. (2013). The “i don’t know” option in the vocabulary size test. *TESOL Quarterly, 47*(4), 790–811. https://doi.org/10.1002/tesq.98

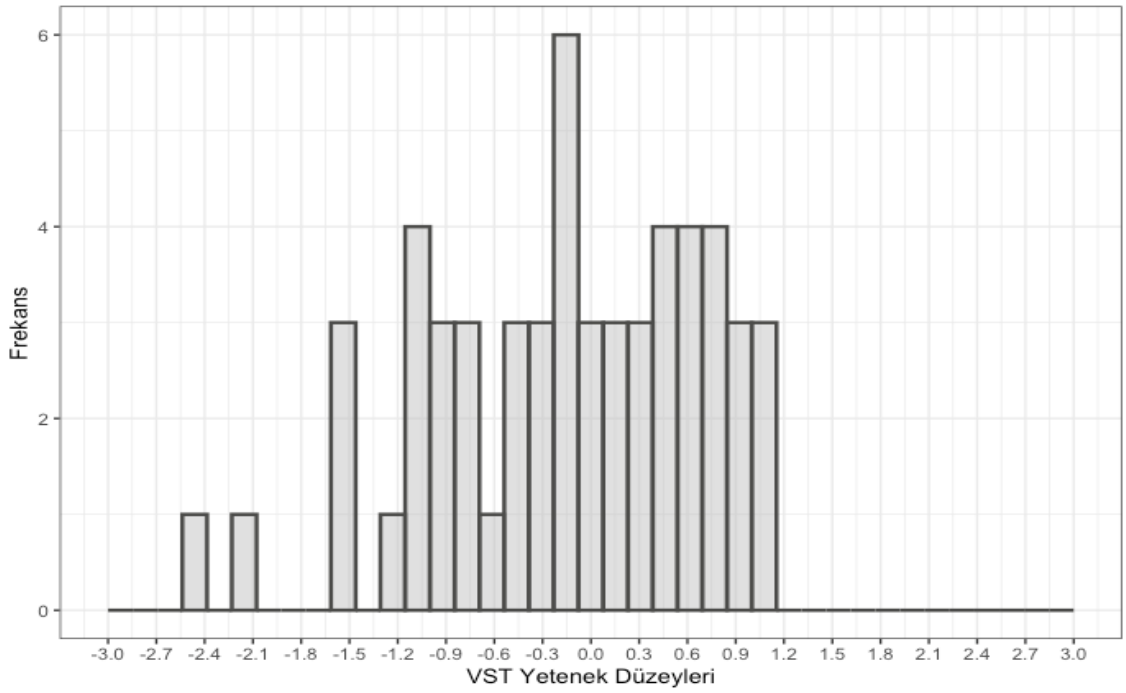
Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research,26*(4), 696–725. https://doi.org/10.1177/1362168820913998

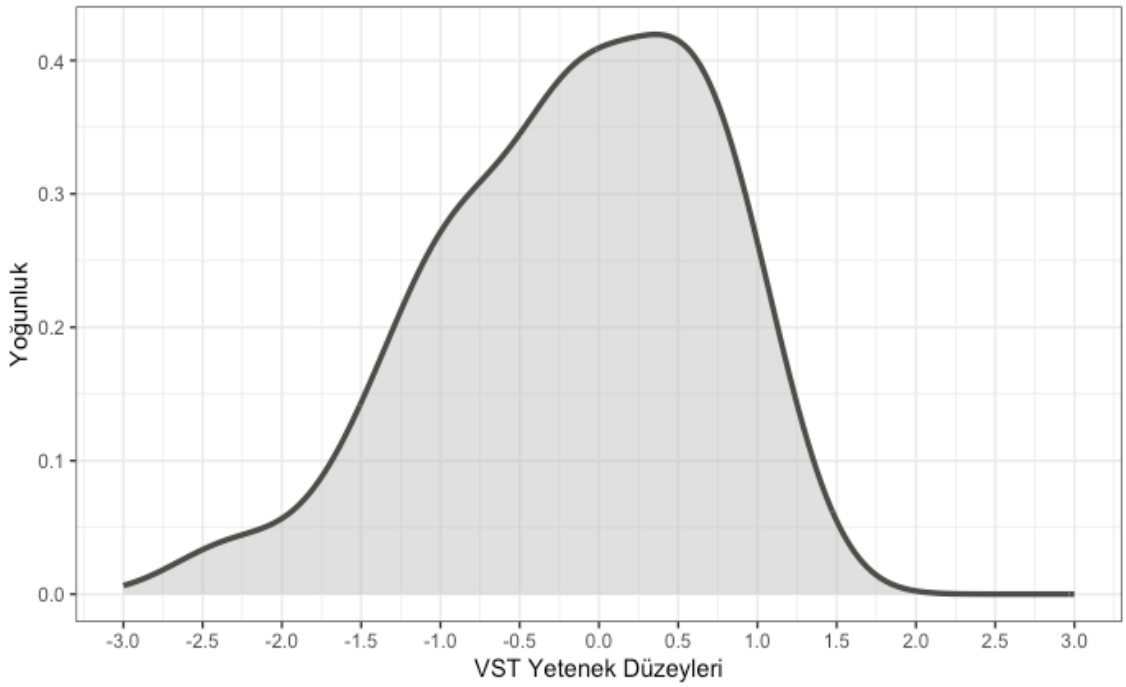
Zhao, P., & Ji, X. (2018). Validation of the Mandarin version of the vocabulary size test. *RELC Journal, 49*(3), 308–321. https://doi.org/10.1177/0033688216639761

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 25–46). Routledge.

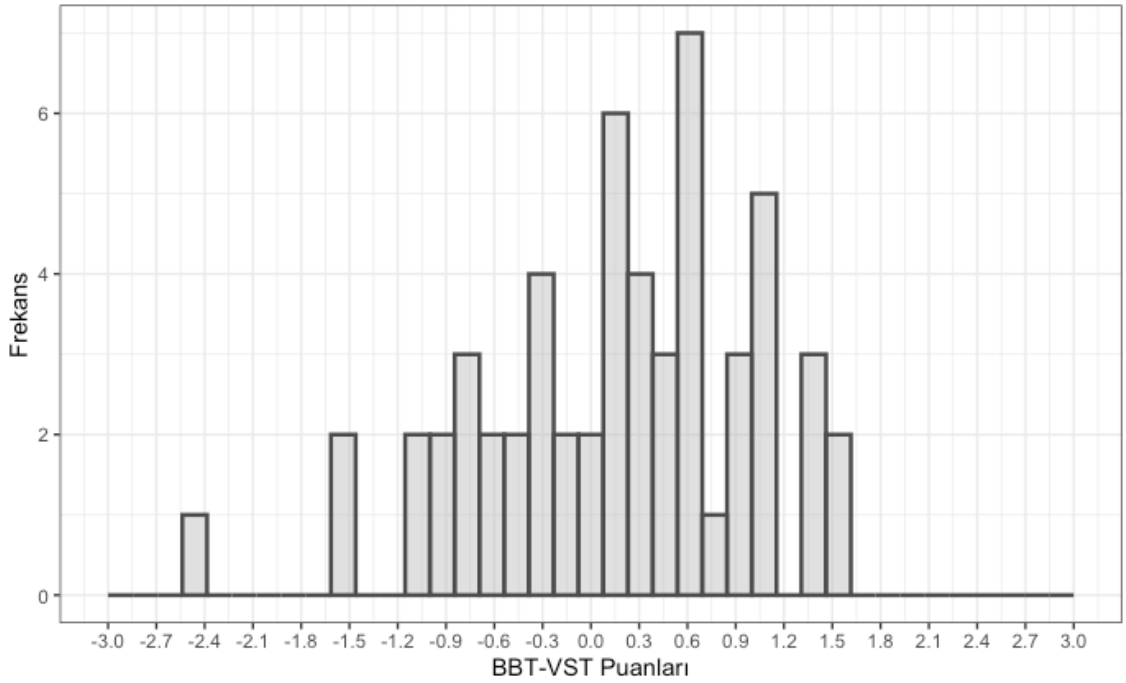
Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233. https://doi.org/10.1080/15434300701375832

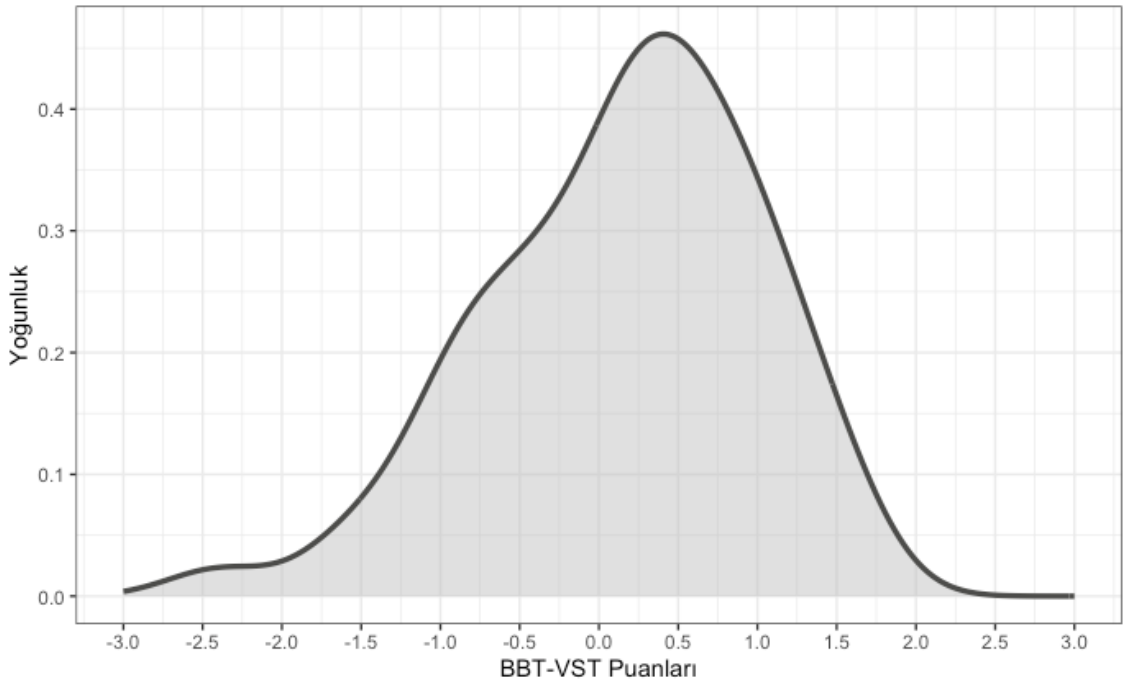
EK-A: BBT Aşamasında VST'nin Kâğıt-Kalem Versiyonundan Elde Edilen Yetenek Değerlerine Ait Histogram ve Yoğunluk Grafikleri





EK-B: BBT Aşamasında VST'nin BBT Versiyonunda Kestirilen Yetenek Değerlerine Ait Histogram ve Yoğunluk Grafikleri





EK-C: VST (İlk Sayfa)

Vocabulary Size Test[1]

Circle the letter a-d with the closest meaning to the key word in the question.

1. SEE: They **saw** it.
   a. cut
   b. waited for
   c. looked at
   d. started

2. TIME: They have a lot of **time**.
   a. money
   b. food
   c. hours
   d. friends

3. PERIOD: It was a difficult **period**.
   a. question
   b. time
   c. thing to do
   d. book

4. FIGURE: Is this the right **figure**?
   a. answer
   b. place
   c. time
   d. number

5. POOR: We are **poor**.
   a. have no money
   b. feel happy
   c. are very interested
   d. do not like to work hard

6. DRIVE: He **drives** fast.
   a. swims
   b. learns
   c. throws balls
   d. uses a car

7. JUMP: She tried to **jump**.
   a. lie on top of the water
   b. get off the ground suddenly
   c. stop the car at the edge of the road
   d. move very fast

8. SHOE: Where is your **shoe**?
   a. the person who looks after you
   b. the thing you keep your money in
   c. the thing you use for writing
   d. the thing you wear on your foot

9. STANDARD: Her **standards** are very high.
   a. the bits at the back under her shoes
   b. the marks she gets in school
   c. the money she asks for
   d. the levels she reaches in everything

10. BASIS: This was used as the **basis**.
    a. answer
    b. place to take a rest
    c. next step
    d. main part

[1] The test is created by Paul Nation, Victoria University of Wellington, and found at http://www.lextutor.ca/. This test is freely available and can be used by teachers and researchers for a variety of purposes.

Second 1000

1. MAINTAIN: Can they **maintain** it?
   a. keep it as it is
   b. make it larger
   c. get a better one than it
   d. get it

2. STONE: He sat on a **stone**.
   a. hard thing
   b. kind of chair
   c. soft thing on the floor
   d. part of a tree

3. UPSET: I am **upset**.
   a. tired
   b. famous
   c. rich
   d. unhappy

4. DRAWER: The **drawer** was empty.
   a. sliding box
   b. place where cars are kept
   c. cupboard to keep things cold
   d. animal house

5. PATIENCE: He has no **patience**.
   a. will not wait happily
   b. has no free time
   c. has no faith
   d. does not know what is fair

6. NIL: His mark for that question was **nil**.
   a. very bad
   b. nothing
   c. very good
   d. in the middle

7. PUB: They went to the **pub**.
   a. place where people drink and talk
   b. place that looks after money
   c. large building with many shops
   d. building for swimming

8. CIRCLE: Make a **circle**.
   a. rough picture
   b. space with nothing in it
   c. round shape
   d. large hole

9. MICROPONE: Please use the **microphone**.
   a. machine for making food hot
   b. machine that makes sounds louder
   c. machine that makes things look bigger
   d. small telephone that can be carried around

10. PRO: He's a **pro**.
    a. someone who is employed to find out important secrets
    b. a stupid person
    c. someone who writes for a newspaper
    d. someone who is paid for playing sport etc

EK-Ç: Uç Değerlere Ait İstatistikler

1. Uç Değerlere ait Betimleyici İstatistikler (Frekans ve Oranlar)

	Kadın Uç Değer	Kadın Toplam	Erkek Uç Değer	Erkek Toplam	Toplam Uç Değer	Toplam
Hazırlık Frekans ve Oran	4 (%15)	27	3 (%14)	22	7 (%14)	49
Lisans Frekans ve Oran	76 (%19)	395	44 (%15)	295	120 (%17)	690
Y. Lisans Frekans ve Oran	4 (%4)	93	3 (%3)	88	7 (%4)	181
Doktora Frekans ve Oran	22 (%7)	308	14 (%6)	229	36 (%7)	537
Toplam Frekans ve Oran	106 (%13)	823	64 (%10)	634	170 (%12)	1457

2. Uç Değerlere ait Betimleyici İstatistikler (VST Puan Ortalamaları)

	Kadın Uç değer	Kadın Toplam	Erkek Uç Değer	Erkek Toplam	Toplam Uç Değer	Toplam
Hazırlık Puan Ortalaması	73	67	50	63	63	65
Lisans Puan Ortalaması	50	61	53	63	51	61
Y. Lisans Puan Ortalaması	50	74	76	73	61	73
Doktora Puan Ortalaması	64	77	62	79	63	78
Toplam Puan Ortalaması	54	68	56	70	54	69

EK-D: Tek Boyutlu Modelde Faktör Yükleri

Madde	Faktör Yükü	Madde	Faktör Yükü	Madde	Faktör Yükü	Madde	Faktör Yükü	Madde	Faktör Yükü
M1	0.269	M29	0.473	M57	0.700	M85	0.734	M113	0.688
M2	0.782	M30	0.720	M58	0.234	M86	0.461	M114	0.765
M3	0.476	M31	0.552	M59	0.687	M87	0.582	M115	0.718
M4	0.301	M32	0.504	M60	0.583	M88	0.859	M116	0.619
M5	0.785	M33	0.421	M61	0.862	M89	0.585	M117	0.718
M6	0.700	M34	0.668	M62	0.647	M90	0.593	M118	0.674
M7	0.413	M35	0.978	M63	0.684	M91	0.492	M119	0.557
M8	0.759	M36	0.706	M64	0.659	M92	0.631	M120	0.525
M9	0.507	M37	0.692	M65	0.692	M93	0.680	M121	0.582
M10	0.554	M38	0.972	M66	0.484	M94	0.926	M122	0.649
M11	0.527	M39	0.654	M67	0.671	M95	0.695	M123	0.714
M12	0.428	M40	0.607	M68	0.475	M96	0.590	M124	0.660
M13	0.757	M41	0.556	M69	0.654	M97	0.742	M125	0.609
M14	0.403	M42	0.698	M70	0.932	M98	0.659	M126	0.809
M15	0.467	M43	0.721	M71	0.669	M99	0.382	M127	0.533
M16	0.375	M44	0.663	M72	0.767	M100	0.537	M128	0.710
M17	0.852	M45	0.522	M73	0.785	M101	0.640	M129	0.605
M18	0.502	M46	0.791	M74	0.915	M102	0.666	M130	0.578
M19	0.637	M47	0.758	M75	0.823	M103	0.959	M131	0.543
M20	0.398	M48	0.753	M76	0.743	M104	0.747	M132	0.783
M21	0.919	M49	0.524	M77	0.643	M105	0.941	M133	0.579
M22	0.715	M50	0.932	M78	0.532	M106	0.664	M134	0.561
M23	0.593	M51	0.665	M79	0.768	M107	0.820	M135	0.560
M24	0.551	M52	0.516	M80	0.798	M108	0.647	M136	0.546
M25	0.416	M53	0.714	M81	0.572	M109	0.738	M137	0.674
M26	0.555	M54	0.725	M82	0.764	M110	0.565	M138	0.544
M27	0.593	M55	0.613	M83	0.738	M111	0.707	M139	0.481
M28	0.481	M56	0.926	M84	0.628	M112	0.654	M140	0.544

EK-E: En Yüksek Q3 Değerine Sahip Madde Çiftleri

1. Çift (2 ve 6)

(Q3 = .73)

2. TIME: They have a lot of **time**.

a. money

b. food

c. hours

d. friends

6. DRIVE: He **drives** fast.

a. swims

b. learns

c. throws balls

d. uses a car

2. Çift (103 ve 105)

(Q3 = .62)

103. YOGA: She has started **yoga**.

a. handwork done by knotting thread

b. a form of exercise for body and mind

c. a game where a cork stuck with feathers is hit between two players

d. a type of dance from eastern countries

105. PUMA: They saw a **puma**.

a. small house made of mud bricks

b. tree from hot, dry countries

c. very strong wind that sucks up anything in its path

d. large wild cat

3. Çift (35 ve 38)

(Q3 = .51)

103. QUIZ: We made a **quiz**.

a. thing to hold arrows

b. serious mistake

c. set of questions

d. box for birds to make nests in

105. VOCABULARY: You will need more **vocabulary**.

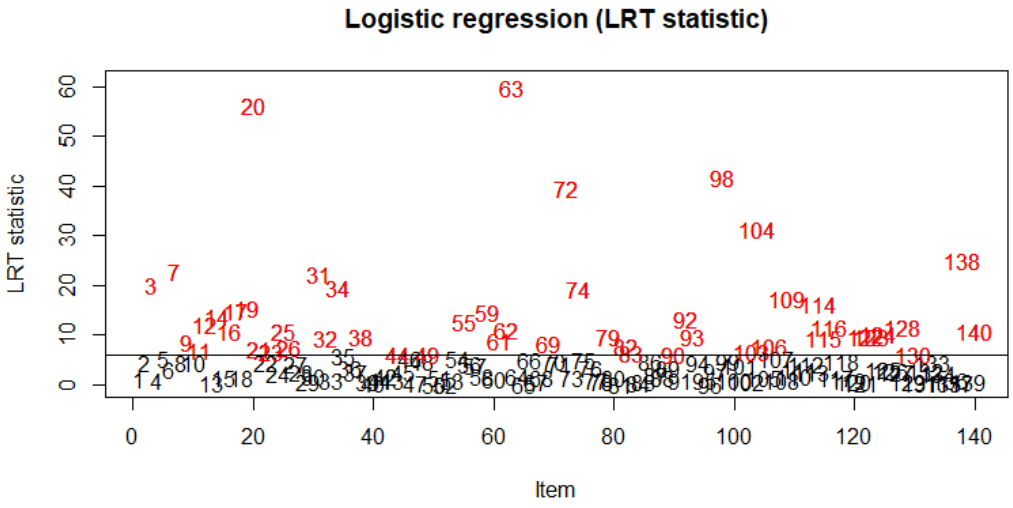
a. words

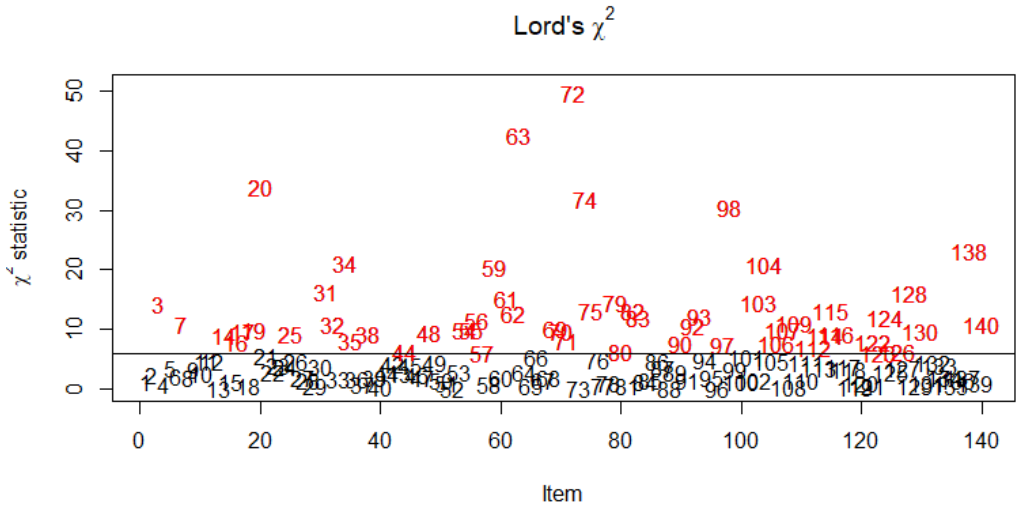
b. skill

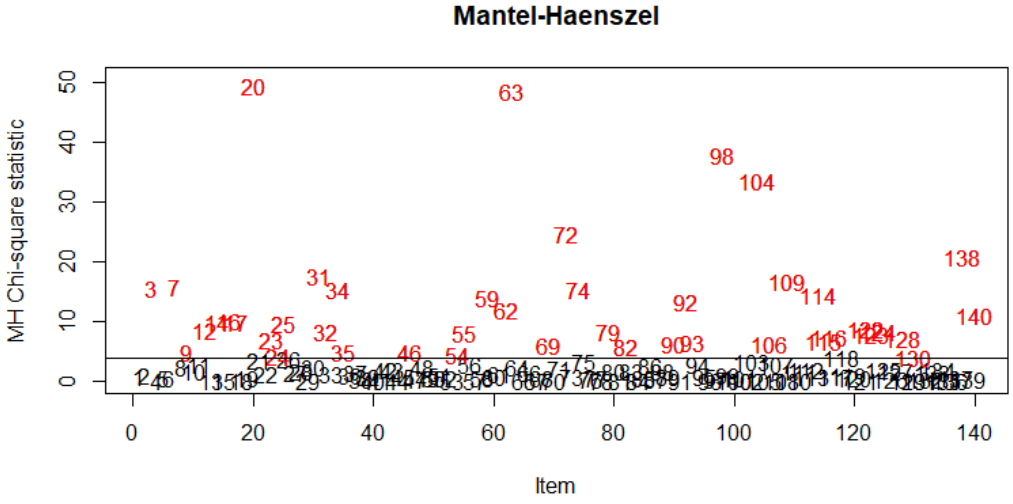
c. money

d. guns

EK-F: DMF Bulgularının Grafikleri

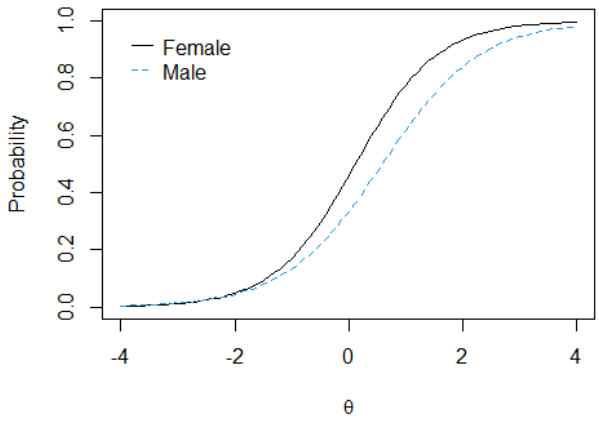




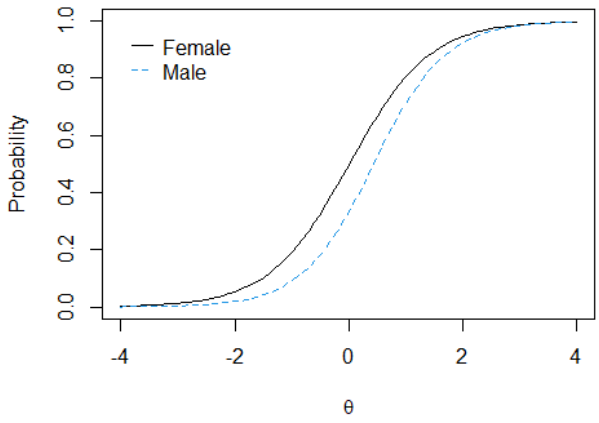


EK-G: Kadınlar lehine DMF Gösteren Maddelerin Madde Karakteristik Eğrileri

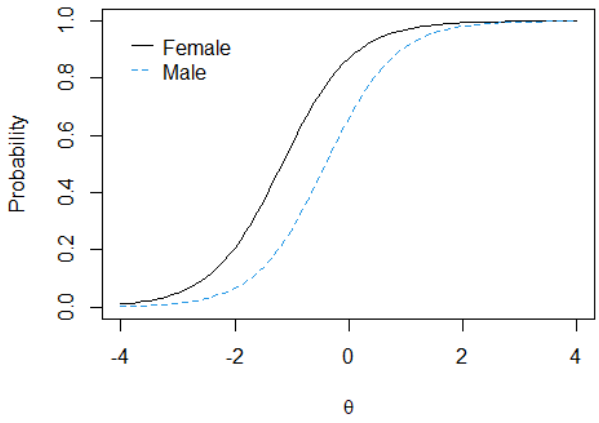
ITEM34



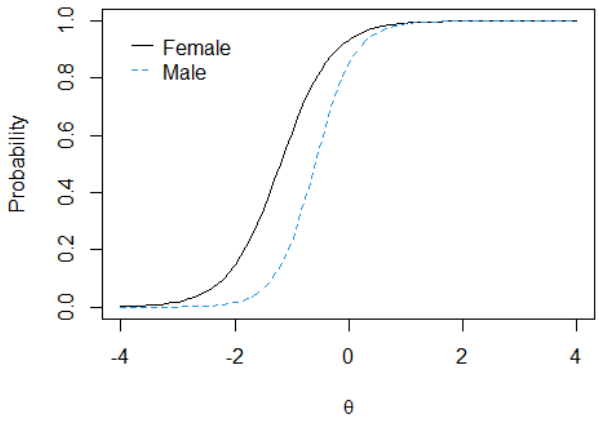
ITEM59



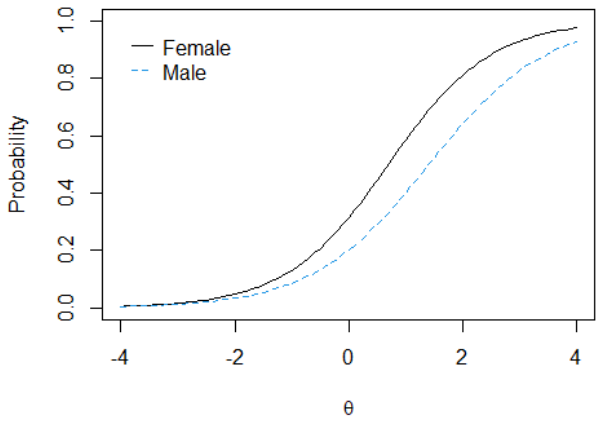
ITEM72



ITEM74

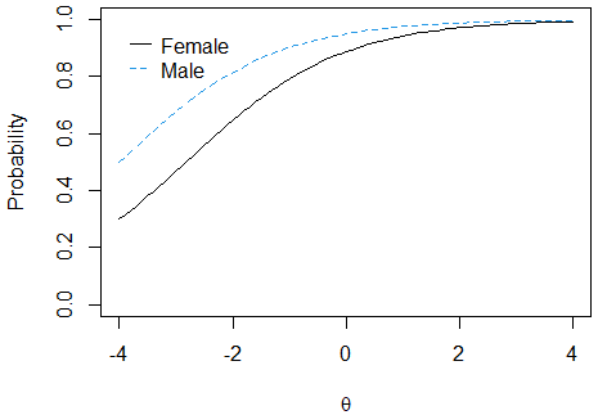


ITEM138

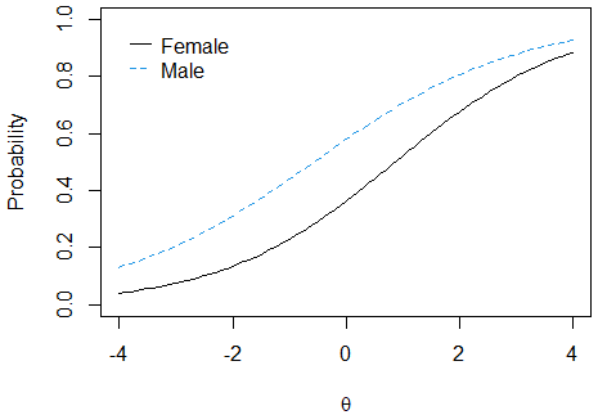


EK-Ğ: Erkekler lehine DMF Gösteren Maddelerin Madde Karakteristik Eğrileri

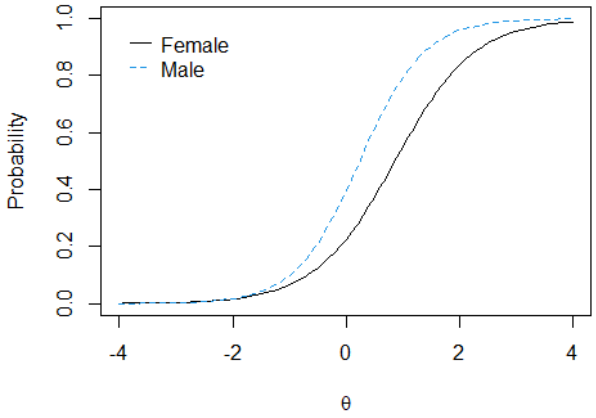
ITEM3



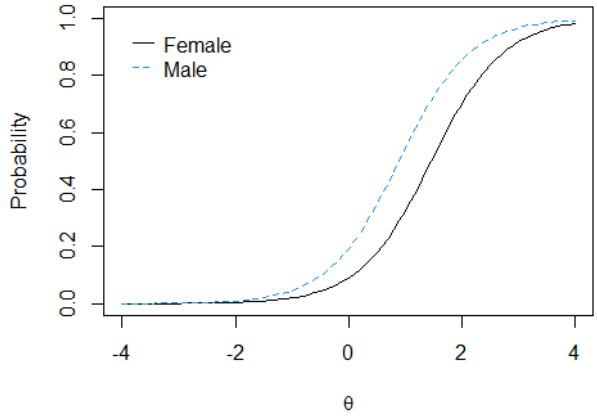
ITEM20



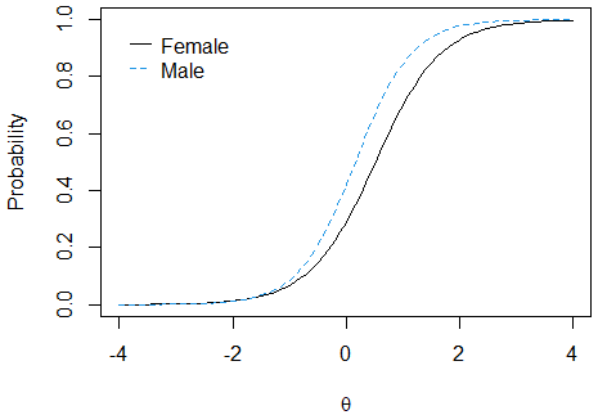
ITEM63



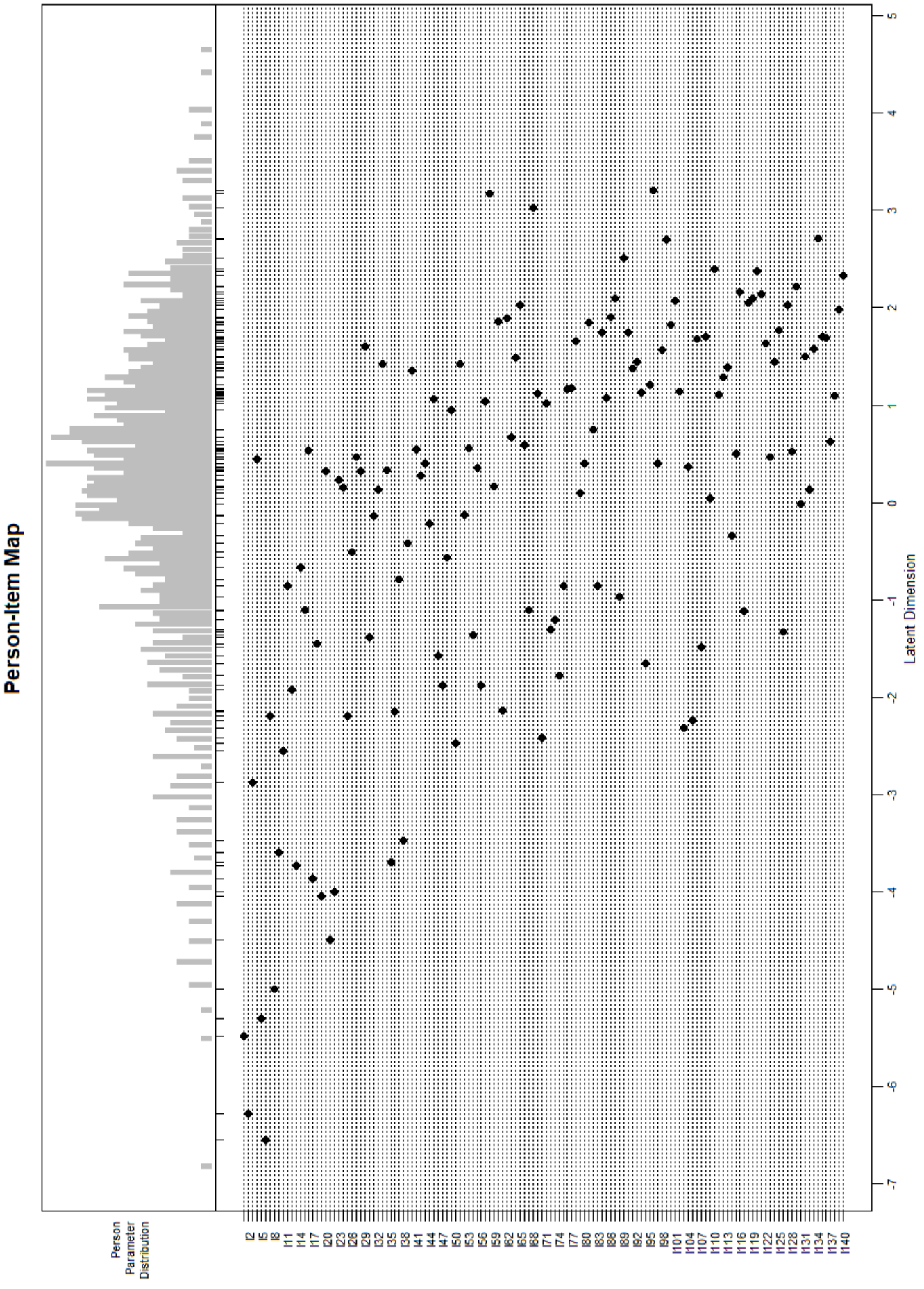
ITEM98



ITEM104



EK-H: Birey – Madde Haritası



EK-I: Tüm Koşullara Ait Genel Simülasyon Bulguları

KOŞUL	ORT	KOR	RMSE	BIAS	MER	IWMER	OVERLAP	CHI_ER
Bm30	7,38	0,9460	0,3225	0,0192	0,0000	61	0,3553	42,3643
Bm25	10,76	0,9580	0,2860	0,0247	0,0000	46	0,3224	34,3783
Bm20	17,41	0,9761	0,2171	0,0224	0,0000	30	0,3404	30,2505
Bb30	11,09	0,9496	0,3104	0,0048	0,0000	11	0,2511	24,0591
Bb25	15,82	0,9657	0,2573	-0,0011	0,0000	2	0,2447	18,4412
Bb20	23,54	0,9811	0,1913	-0,0002	0,0008	1	0,2800	15,6702
Mm30	10,25	0,9542	0,3276	0,0272	0,0000	43	0,3264	35,4446
Mm25	13,38	0,9636	0,2903	0,0202	0,0000	37	0,3196	31,3625
Mm20	19,92	0,9772	0,2316	0,0141	0,0000	24	0,3424	28,0181
Mb30	14,94	0,9528	0,3210	-0,0033	0,0008	2	0,2618	21,7102
Mb25	19,56	0,9675	0,2702	-0,0111	0,0023	1	0,2620	17,1171
Mb20	26,98	0,9799	0,2131	-0,0129	0,0101	1	0,2933	14,0846
Wm30	8,45	0,9496	0,3245	0,0067	0,0000	52	0,3437	39,6670
Wm25	11,72	0,9623	0,2846	0,0137	0,0000	44	0,3268	34,0337
Wm20	18,17	0,9771	0,2261	0,0124	0,0000	29	0,3455	30,2016
Wb30	12,74	0,9553	0,3092	-0,0122	0,0000	6	0,2357	20,2558
Wb25	17,31	0,9673	0,2650	-0,0085	0,0008	2	0,2369	15,8605
Wb20	24,75	0,9816	0,2001	-0,0107	0,0039	1	0,2759	13,8708
Em30	8,31	0,9522	0,3041	0,0114	0,0000	51	0,3499	40,6762
Em25	12,23	0,9635	0,2672	0,0093	0,0000	41	0,3272	33,5773
Em20	18,84	0,9783	0,2068	0,0131	0,0000	21	0,3491	30,0380
Eb30	12,59	0,9573	0,2864	-0,0072	0,0000	5	0,2431	21,4442
Eb25	17,64	0,9686	0,2465	-0,0077	0,0000	1	0,2463	16,8411
Eb20	25,37	0,9823	0,1854	-0,0003	0,0008	1	0,2896	15,1681
Bm30r3	7,77	0,9442	0,3279	0,0220	0,0000	54	0,2575	28,2859
Bm25r3	11,22	0,9552	0,2952	0,0255	0,0000	42	0,2768	27,5372
Bm20r3	17,78	0,9766	0,2138	0,0178	0,0000	28	0,3252	27,7471
Bb30r3	11,76	0,9497	0,3100	-0,0134	0,0008	1	0,1568	10,1958
Bb25r3	16,63	0,9618	0,2713	-0,0045	0,0016	1	0,1929	10,3739
Bb20r3	24,14	0,9795	0,1993	0,0034	0,0008	1	0,2532	11,3091
Mm30r3	10,71	0,9576	0,3142	0,0215	0,0000	39	0,2360	22,3335
Mm25r3	13,97	0,9669	0,2803	0,0223	0,0000	34	0,2603	22,4783
Mm20r3	20,24	0,9787	0,2249	0,0065	0,0000	20	0,3120	23,4323
Mb30r3	15,14	0,9522	0,3233	-0,0022	0,0109	1	0,1658	8,0737
Mb25r3	19,59	0,9639	0,2837	-0,0003	0,0093	1	0,1988	8,2435
Mb20r3	27,07	0,9806	0,2106	-0,0079	0,0101	1	0,2490	7,7965
Wm30r3	8,79	0,9512	0,3198	-0,0041	0,0000	48	0,2519	26,4695
Wm25r3	12,08	0,9653	0,2734	0,0111	0,0000	38	0,2798	27,1012
Wm20r3	18,55	0,9771	0,2272	0,0083	0,0000	21	0,3269	27,2099
Wb30r3	12,99	0,9514	0,3184	0,0147	0,0016	1	0,1563	8,8846
Wb25r3	17,81	0,9681	0,2643	0,0022	0,0016	1	0,1921	9,0853
Wb20r3	25,21	0,9791	0,2137	-0,0059	0,0047	1	0,2522	10,0989
Em30r3	8,78	0,9506	0,3085	0,0002	0,0000	45	0,2759	29,8439

Em25r3	12,57	0,9642	0,2643	0,0089	0,0000	38	0,2928	28,4220
Em20r3	19,11	0,9787	0,2049	0,0114	0,0000	20	0,3380	28,2155
Eb30r3	13,22	0,9535	0,2986	-0,0014	0,0000	1	0,1681	10,3104
Eb25r3	18,03	0,9672	0,2518	-0,0018	0,0000	2	0,2043	10,5650
Eb20r3	25,66	0,9821	0,1867	-0,0107	0,0016	1	0,2760	12,9828
Bm30r5	8,25	0,9525	0,3017	0,0032	0,0000	47	0,2165	22,0583
Bm25r5	11,76	0,9629	0,2676	0,0027	0,0000	34	0,2542	23,8329
Bm20r5	18,24	0,9766	0,2140	0,0174	0,0000	23	0,3174	26,1905
Bb30r5	12,10	0,9460	0,3214	0,0164	0,0000	1	0,1470	8,4775
Bb25r5	16,67	0,9633	0,2657	0,0062	0,0000	1	0,1865	9,4379
Bb20r5	24,24	0,9783	0,2054	0,0110	0,0008	1	0,2472	10,3643
Mm30r5	11,04	0,9492	0,3397	0,0189	0,0000	31	0,1975	16,6003
Mm25r5	14,28	0,9677	0,2765	0,0084	0,0000	29	0,2332	18,3774
Mm20r5	20,70	0,9783	0,2265	0,0031	0,0000	17	0,2982	21,0583
Mb30r5	15,14	0,9538	0,3218	0,0024	0,0109	1	0,1519	6,1348
Mb25r5	19,41	0,9682	0,2690	0,0066	0,0109	1	0,1838	6,3200
Mb20r5	27,04	0,9795	0,2159	-0,0040	0,0132	1	0,2453	7,2959
Wm30r5	9,11	0,9451	0,3357	-0,0043	0,0000	40	0,2211	21,8347
Wm25r5	12,48	0,9659	0,2698	-0,0038	0,0000	29	0,2593	23,8303
Wm20r5	19,00	0,9794	0,2147	0,0070	0,0000	17	0,3164	25,2943
Wb30r5	13,39	0,9425	0,3423	0,0040	0,0016	1	0,1428	6,5950
Wb25r5	17,88	0,9608	0,2880	-0,0061	0,0031	1	0,1841	7,8986
Wb20r5	25,33	0,9786	0,2150	0,0036	0,0016	1	0,2449	8,9557
Em30r5	9,28	0,9567	0,2894	-0,0014	0,0000	38	0,2322	23,2242
Em25r5	12,96	0,9640	0,2649	-0,0016	0,0000	29	0,2693	24,7456
Em20r5	19,36	0,9795	0,2000	0,0070	0,0000	18	0,3238	25,9800
Eb30r5	13,31	0,9591	0,2815	-0,0264	0,0000	1	0,1630	9,5094
Eb25r5	18,08	0,9675	0,2508	0,0006	0,0016	2	0,1933	8,9766
Eb20r5	25,56	0,9806	0,1939	0,0013	0,0016	1	0,2567	10,3776

EK-İ: Yetenek Düzeyine Göre Oluşturulan Gruplardaki Ortalama RMSE Değerleri

KOŞUL	RMSE_D1	RMSE_D2	RMSE_D3	RMSE_D4	RMSE_D5	RMSE_D6	RMSE_D7	RMSE_D8	RMSE_D9	RMSE_D10
Bm30	0,3730	0,3270	0,3710	0,3870	0,3330	0,2960	0,2720	0,2500	0,2570	0,3240
Bm25	0,3290	0,3070	0,3250	0,3650	0,2950	0,2690	0,2470	0,2210	0,2280	0,2360
Bm20	0,1200	0,2730	0,2470	0,2890	0,2260	0,2240	0,2030	0,1680	0,1770	0,1890
Bb30	0,3720	0,2610	0,2960	0,3140	0,3670	0,3320	0,2830	0,2810	0,2660	0,3080
Bb25	0,2850	0,2270	0,2640	0,2800	0,3110	0,2770	0,2270	0,2320	0,2210	0,2290
Bb20	0,1300	0,1700	0,2050	0,2350	0,2240	0,2120	0,1800	0,1810	0,1860	0,1670
Mm30	0,2780	0,3220	0,4020	0,4190	0,3610	0,3240	0,3060	0,2680	0,2760	0,2810
Mm25	0,2610	0,2960	0,3480	0,3870	0,2990	0,2980	0,2620	0,2160	0,2380	0,2560
Mm20	0,2440	0,2640	0,2410	0,2940	0,2300	0,2410	0,2030	0,1750	0,1850	0,2140
Mb30	0,2890	0,2870	0,3460	0,3730	0,3630	0,3590	0,2900	0,2810	0,2910	0,3130
Mb25	0,2510	0,2540	0,2830	0,3160	0,3180	0,2780	0,2490	0,2550	0,2420	0,2440
Mb20	0,2320	0,2070	0,2200	0,2440	0,2270	0,2270	0,1810	0,1890	0,2000	0,1940
Wm30	0,4300	0,3260	0,3610	0,3850	0,3180	0,3010	0,2720	0,2290	0,2520	0,3180
Wm25	0,3450	0,3100	0,3410	0,3600	0,2780	0,2720	0,2400	0,1990	0,2200	0,2300
Wm20	0,2350	0,2650	0,2510	0,2950	0,2200	0,2230	0,2020	0,1660	0,1740	0,1970
Wb30	0,3290	0,2530	0,3080	0,3800	0,3490	0,2950	0,2840	0,2770	0,3130	0,2840
Wb25	0,3230	0,2330	0,2730	0,3140	0,3170	0,2480	0,2260	0,2070	0,2460	0,2340
Wb20	0,2210	0,1660	0,2090	0,2420	0,2140	0,2110	0,1850	0,1570	0,1940	0,1850
Em30	0,3570	0,3250	0,3350	0,3420	0,3030	0,2900	0,2880	0,2530	0,2530	0,2740
Em25	0,2930	0,3010	0,2750	0,3170	0,2600	0,2690	0,2640	0,2210	0,2190	0,2340
Em20	0,1270	0,2640	0,2150	0,2630	0,2080	0,2110	0,2040	0,1760	0,1690	0,1940
Eb30	0,3210	0,2710	0,2640	0,3070	0,3000	0,2950	0,2850	0,2610	0,2660	0,2860
Eb25	0,2780	0,2390	0,2420	0,2760	0,2660	0,2430	0,2320	0,2170	0,2310	0,2350
Eb20	0,1300	0,1650	0,2000	0,2170	0,1970	0,1970	0,1850	0,1810	0,1830	0,1860
Bm30r3	0,4350	0,3460	0,3430	0,3800	0,3420	0,3080	0,2700	0,2470	0,2610	0,3010
Bm25r3	0,4120	0,3390	0,2980	0,3440	0,2830	0,2530	0,2400	0,1880	0,2390	0,2900
Bm20r3	0,1870	0,2570	0,2270	0,2710	0,2160	0,2060	0,2090	0,1720	0,1820	0,1880
Bb30r3	0,3630	0,3390	0,3070	0,3270	0,3210	0,2990	0,2870	0,2210	0,3160	0,2980
Bb25r3	0,3310	0,2300	0,2680	0,3210	0,3070	0,2630	0,2380	0,2440	0,2390	0,2480
Bb20r3	0,1340	0,1860	0,2220	0,2400	0,2180	0,2150	0,1840	0,1860	0,2190	0,1670
Mm30r3	0,2910	0,2920	0,3830	0,3760	0,3290	0,2980	0,3040	0,2530	0,2910	0,3030
Mm25r3	0,2700	0,2370	0,3260	0,3640	0,3020	0,2790	0,2530	0,2060	0,2400	0,2930
Mm20r3	0,2450	0,2240	0,2500	0,2720	0,2290	0,2110	0,2030	0,1810	0,1780	0,2370
Mb30r3	0,2560	0,3680	0,3490	0,3840	0,3530	0,3490	0,2970	0,2650	0,2800	0,3020
Mb25r3	0,3090	0,2730	0,2870	0,3160	0,3220	0,2800	0,2650	0,2590	0,2690	0,2470
Mb20r3	0,2290	0,1830	0,2060	0,2340	0,2350	0,2150	0,2070	0,1750	0,2100	0,2030
Wm30r3	0,4330	0,2970	0,3560	0,3650	0,3390	0,2850	0,2600	0,2280	0,2670	0,3150

Wm25r3	0,3530	0,2600	0,3080	0,3430	0,3010	0,2450	0,2320	0,1900	0,2210	0,2310
Wm20r3	0,2390	0,2220	0,2430	0,2930	0,2240	0,2220	0,1980	0,1510	0,1790	0,2670
Wb30r3	0,4030	0,2830	0,3110	0,3800	0,3490	0,3100	0,2690	0,2710	0,3050	0,2720
Wb25r3	0,2550	0,2670	0,2640	0,2960	0,3290	0,2530	0,2530	0,2250	0,2400	0,2460
Wb20r3	0,2250	0,2570	0,2010	0,2680	0,1820	0,1920	0,1860	0,1950	0,2250	0,1860
Em30r3	0,3770	0,3370	0,3070	0,3450	0,3080	0,3040	0,2850	0,2380	0,2650	0,2940
Em25r3	0,2860	0,2680	0,2720	0,3050	0,2660	0,2720	0,2490	0,2130	0,2290	0,2700
Em20r3	0,1300	0,2620	0,2250	0,2510	0,1990	0,1960	0,2140	0,1660	0,1730	0,1970
Eb30r3	0,3460	0,3240	0,2810	0,3100	0,3480	0,2940	0,2740	0,2470	0,2580	0,2850
Eb25r3	0,2730	0,2270	0,2620	0,2840	0,2830	0,2640	0,2380	0,2460	0,2180	0,2110
Eb20r3	0,1320	0,1590	0,1990	0,2220	0,1830	0,1930	0,1970	0,1760	0,2140	0,1750
Bm30r5	0,4050	0,2600	0,3450	0,3330	0,2800	0,2630	0,2810	0,2430	0,2580	0,3100
Bm25r5	0,3280	0,2500	0,2670	0,3320	0,2630	0,2570	0,2450	0,2020	0,2290	0,2760
Bm20r5	0,1960	0,2420	0,2410	0,2820	0,2180	0,2080	0,1820	0,1630	0,1840	0,1980
Bb30r5	0,3950	0,4550	0,2720	0,3120	0,2990	0,3070	0,2430	0,2790	0,3140	0,2800
Bb25r5	0,2940	0,3250	0,2690	0,3000	0,2880	0,2540	0,2220	0,2250	0,2290	0,2310
Bb20r5	0,2070	0,1730	0,2280	0,2540	0,2230	0,2130	0,1820	0,1850	0,1990	0,1760
Mm30r5	0,4580	0,3600	0,3410	0,3740	0,3250	0,3260	0,3010	0,2390	0,2970	0,3320
Mm25r5	0,2660	0,2680	0,2820	0,3570	0,2740	0,2910	0,2280	0,2060	0,2910	0,2750
Mm20r5	0,2570	0,2410	0,2570	0,2720	0,2190	0,2080	0,2160	0,1550	0,1790	0,2330
Mb30r5	0,3170	0,3600	0,2750	0,3330	0,3870	0,3220	0,2770	0,3380	0,3190	0,2700
Mb25r5	0,2470	0,2690	0,2790	0,3390	0,2880	0,2660	0,2610	0,2240	0,2730	0,2250
Mb20r5	0,2430	0,2120	0,2140	0,2580	0,2190	0,2160	0,1820	0,1920	0,2060	0,2070
Wm30r5	0,4650	0,3140	0,3720	0,3970	0,3240	0,2970	0,2860	0,2230	0,2350	0,3700
Wm25r5	0,3430	0,2690	0,2920	0,3250	0,2740	0,2620	0,2220	0,2100	0,2170	0,2490
Wm20r5	0,2200	0,2150	0,2310	0,2770	0,2170	0,2090	0,2000	0,1730	0,1750	0,2120
Wb30r5	0,4940	0,3810	0,3410	0,3380	0,3400	0,2940	0,2680	0,2830	0,3350	0,2920
Wb25r5	0,3840	0,3250	0,2880	0,3090	0,2940	0,2780	0,2510	0,2280	0,2480	0,2380
Wb20r5	0,2300	0,2510	0,2070	0,2590	0,2010	0,2100	0,1630	0,1760	0,2110	0,2220
Em30r5	0,3360	0,2910	0,3160	0,3230	0,3170	0,2730	0,2810	0,2390	0,2480	0,2510
Em25r5	0,2940	0,2840	0,2590	0,2550	0,2660	0,2520	0,2650	0,2370	0,2320	0,2970
Em20r5	0,1740	0,2270	0,2250	0,2320	0,2130	0,2030	0,2050	0,1580	0,1790	0,1670
Eb30r5	0,2310	0,3000	0,2710	0,3050	0,2930	0,2690	0,2730	0,2990	0,2920	0,2750
Eb25r5	0,2590	0,2510	0,2530	0,2830	0,2700	0,2370	0,2660	0,2290	0,2220	0,2280
Eb20r5	0,1520	0,2180	0,1910	0,2380	0,2030	0,1990	0,1710	0,1790	0,1980	0,1760

EK-J: Yetenek Düzeyine Göre Oluşturulan Gruplardaki Ortalama Yanlılık Değerleri

KOŞUL	MBIAS_D1	MBIAS_D2	MBIAS_D3	MBIAS_D4	MBIAS_D5	MBIAS_D6	MBIAS_D7	MBIAS_D8	MBIAS_D9	MBIAS_D10
Bm30	0,1500	0,1100	0,0890	0,0520	0,0190	0,0190	-0,0520	-0,0210	-0,0670	-0,1070
Bm25	0,1200	0,1010	0,0730	0,0660	0,0220	0,0200	-0,0510	-0,0250	-0,0410	-0,0390
Bm20	0,0570	0,0900	0,0390	0,0480	0,0430	0,0230	-0,0260	-0,0250	-0,0180	-0,0060
Bb30	0,1830	0,1130	0,0550	-0,0130	-0,0020	-0,0310	-0,0180	-0,0330	-0,0700	-0,1360
Bb25	0,1120	0,0820	0,0450	-0,0120	-0,0300	-0,0400	-0,0250	-0,0400	-0,0380	-0,0660
Bb20	0,0680	0,0590	0,0370	-0,0280	-0,0180	-0,0410	-0,0170	-0,0140	-0,0130	-0,0350
Mm30	-0,1170	0,0210	0,0780	0,0730	0,0650	0,0560	-0,0070	0,0180	0,0300	0,0560
Mm25	-0,1360	0,0470	0,0610	0,0650	0,0530	0,0460	-0,0180	0,0140	0,0070	0,0620
Mm20	-0,1550	0,0560	0,0190	0,0380	0,0570	0,0330	-0,0080	0,0060	0,0200	0,0750
Mb30	-0,0670	0,0180	0,0630	0,0150	-0,0390	-0,0050	0,0130	-0,0360	-0,0070	0,0120
Mb25	-0,1240	0,0190	0,0470	-0,0120	-0,0540	-0,0120	0,0120	-0,0220	-0,0070	0,0420
Mb20	-0,1450	0,0190	0,0430	-0,0210	-0,0220	-0,0080	-0,0070	-0,0090	-0,0120	0,0330
Wm30	-0,0500	0,0490	0,0640	0,0720	0,0420	0,0340	-0,0360	-0,0030	-0,0460	-0,0590
Wm25	-0,0850	0,0650	0,0700	0,0450	0,0420	0,0310	-0,0240	0,0000	-0,0170	0,0100
Wm20	-0,1300	0,0600	0,0350	0,0520	0,0600	0,0300	-0,0150	-0,0080	0,0060	0,0330
Wb30	-0,0590	0,0200	0,0340	0,0160	-0,0710	-0,0090	0,0250	-0,0250	-0,0350	-0,0180
Wb25	-0,0790	0,0170	0,0330	-0,0100	-0,0590	0,0010	0,0400	0,0000	-0,0270	-0,0020
Wb20	-0,1190	0,0300	0,0300	-0,0020	-0,0300	-0,0040	0,0060	0,0030	-0,0250	0,0030
Em30	0,1340	0,1060	0,0550	0,0300	-0,0120	-0,0170	-0,0940	-0,0490	-0,0280	-0,0100
Em25	0,0870	0,0970	0,0310	0,0110	-0,0110	-0,0270	-0,0710	-0,0220	-0,0160	0,0150
Em20	0,0420	0,0880	0,0160	0,0070	0,0220	-0,0230	-0,0350	-0,0210	0,0030	0,0330
Eb30	0,1450	0,0950	0,0550	-0,0140	-0,0650	-0,0370	-0,0630	-0,0560	-0,0450	-0,0860
Eb25	0,0960	0,0650	0,0270	-0,0210	-0,0450	-0,0400	-0,0610	-0,0280	-0,0250	-0,0440
Eb20	0,0520	0,0590	0,0290	-0,0140	-0,0130	-0,0400	-0,0390	-0,0240	-0,0030	-0,0100
Bm30r3	0,1660	0,1230	0,0680	0,0750	-0,0100	-0,0010	-0,0390	-0,0160	-0,0370	-0,1090
Bm25r3	0,1350	0,1130	0,0560	0,0550	0,0430	-0,0150	-0,0240	-0,0350	-0,0350	-0,0380
Bm20r3	0,0670	0,0840	0,0320	0,0300	0,0460	-0,0030	-0,0300	-0,0290	-0,0250	0,0060
Bb30r3	0,1700	0,1280	0,0680	-0,0260	-0,0670	-0,0680	-0,0520	-0,0700	-0,0970	-0,1200
Bb25r3	0,1230	0,0740	0,0780	0,0080	-0,0830	-0,0300	-0,0490	-0,0370	-0,0670	-0,0620
Bb20r3	0,0670	0,0580	0,0530	-0,0150	-0,0040	-0,0350	-0,0100	-0,0140	-0,0380	-0,0270
Mm30r3	-0,1130	0,0140	0,0950	0,0720	0,0150	0,0290	-0,0110	0,0260	0,0250	0,0630
Mm25r3	-0,1290	0,0170	0,0500	0,0590	0,0480	0,0240	0,0040	0,0100	0,0490	0,0900
Mm20r3	-0,1520	0,0130	0,0190	0,0180	0,0540	0,0200	-0,0060	0,0120	0,0260	0,0620
Mb30r3	-0,0960	0,0650	0,0920	-0,0100	-0,0130	-0,0020	-0,0310	0,0010	-0,0120	-0,0140
Mb25r3	-0,1040	0,0430	0,0670	-0,0330	-0,0440	0,0340	-0,0010	0,0030	0,0060	0,0250
Mb20r3	-0,1410	0,0090	0,0250	-0,0190	-0,0090	-0,0050	0,0220	0,0250	-0,0180	0,0320
Wm30r3	-0,0450	0,0160	0,0650	0,0160	0,0310	0,0160	-0,0220	-0,0230	-0,0390	-0,0540
Wm25r3	-0,0770	0,0390	0,0330	0,0270	0,0830	0,0160	-0,0080	-0,0020	-0,0170	0,0160
Wm20r3	-0,1300	0,0390	0,0240	0,0460	0,0550	0,0060	-0,0030	-0,0110	0,0160	0,0410
Wb30r3	-0,0130	0,0840	0,0330	0,0660	0,0010	0,0100	-0,0120	0,0050	-0,0110	-0,0140
Wb25r3	-0,0950	0,0490	0,0400	-0,0380	-0,0110	0,0120	0,0420	0,0170	-0,0100	0,0160
Wb20r3	-0,1260	0,0470	0,0430	0,0040	-0,0040	-0,0240	-0,0060	-0,0050	-0,0090	0,0200

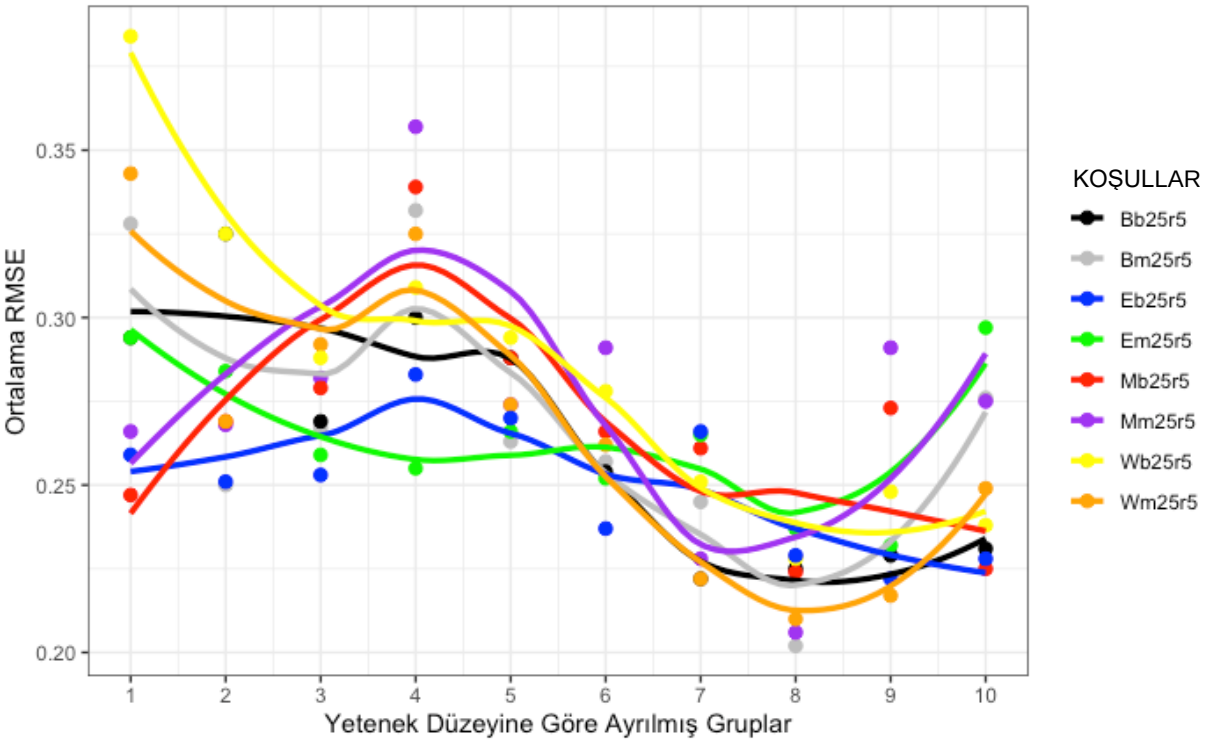
Em30r3	0,1440	0,1170	0,0390	-0,0120	-0,0190	-0,0480	-0,0840	-0,0730	-0,0410	-0,0220
Em25r3	0,0870	0,0800	0,0360	-0,0080	-0,0190	-0,0300	-0,0320	-0,0140	-0,0090	-0,0020
Em20r3	0,0420	0,0830	0,0280	0,0100	-0,0120	-0,0160	-0,0380	-0,0220	0,0110	0,0280
Eb30r3	0,1590	0,1010	0,0650	-0,0620	-0,0690	-0,0230	-0,0310	-0,0230	-0,0290	-0,1030
Eb25r3	0,0920	0,0700	0,0730	-0,0230	-0,0760	-0,0310	-0,0440	-0,0220	-0,0430	-0,0160
Eb20r3	0,0540	0,0490	0,0240	-0,0460	-0,0350	-0,0420	-0,0180	-0,0160	-0,0610	-0,0160
Bm30r5	0,1610	0,0690	0,0640	0,0360	-0,0200	-0,0380	-0,0450	-0,0350	-0,0570	-0,1020
Bm25r5	0,1180	0,0600	0,0240	-0,0170	0,0340	-0,0170	-0,0410	-0,0260	-0,0350	-0,0730
Bm20r5	0,0740	0,0780	0,0340	0,0340	0,0370	-0,0200	-0,0290	-0,0220	-0,0100	-0,0020
Bb30r5	0,1820	0,2050	0,0540	-0,0220	-0,0400	-0,0220	0,0020	-0,0200	-0,0930	-0,0810
Bb25r5	0,1280	0,1330	0,0790	-0,0470	-0,0190	-0,0680	-0,0530	-0,0230	-0,0400	-0,0270
Bb20r5	0,0830	0,0640	0,0560	0,0010	-0,0130	-0,0140	0,0030	0,0000	-0,0450	-0,0250
Mm30r5	-0,0810	0,0310	0,0760	0,0240	0,0110	0,0180	0,0080	0,0360	0,0090	0,0570
Mm25r5	-0,1390	-0,0020	0,0060	0,0330	0,0480	0,0570	-0,0310	0,0330	0,0120	0,0670
Mm20r5	-0,1550	0,0110	0,0400	0,0160	0,0380	0,0000	-0,0030	-0,0090	0,0230	0,0690
Mb30r5	-0,0870	0,0290	0,0190	0,0210	0,0280	0,0070	0,0020	-0,0270	0,0100	0,0220
Mb25r5	-0,1220	0,0350	0,0540	0,0140	-0,0110	0,0150	-0,0080	0,0190	0,0160	0,0560
Mb20r5	-0,1460	0,0220	0,0500	0,0130	-0,0330	0,0090	0,0040	0,0030	0,0030	0,0360
Wm30r5	-0,0260	0,0740	0,0430	0,0110	0,0330	-0,0040	-0,0300	0,0020	-0,0500	-0,0930
Wm25r5	-0,0790	0,0370	0,0290	-0,0070	0,0320	0,0170	-0,0370	-0,0110	-0,0210	0,0040
Wm20r5	-0,1220	0,0420	0,0190	0,0500	0,0390	-0,0160	-0,0020	0,0110	0,0120	0,0380
Wb30r5	0,0050	0,0920	0,0990	0,0310	-0,0430	-0,0100	-0,0120	-0,0420	-0,0310	-0,0500
Wb25r5	-0,0580	0,0490	0,0530	-0,0330	-0,0210	-0,0220	-0,0060	-0,0080	-0,0090	-0,0050
Wb20r5	-0,1140	0,0610	0,0660	0,0030	0,0060	0,0000	0,0150	-0,0130	-0,0040	0,0160
Em30r5	0,1190	0,0980	0,0670	-0,0140	-0,0380	-0,0790	-0,0570	-0,0540	-0,0530	-0,0050
Em25r5	0,0810	0,0940	0,0280	-0,0290	-0,0150	-0,0880	-0,0540	-0,0270	-0,0230	0,0180
Em20r5	0,0490	0,0740	0,0200	-0,0160	0,0190	-0,0250	-0,0350	-0,0330	-0,0040	0,0210
Eb30r5	0,1230	0,0920	0,0570	-0,0420	-0,0790	-0,1040	-0,0500	-0,0980	-0,0770	-0,0860
Eb25r5	0,0810	0,0780	0,0360	0,0040	-0,0430	-0,0410	-0,0340	-0,0020	-0,0490	-0,0250
Eb20r5	0,0560	0,0570	0,0250	-0,0340	0,0050	-0,0020	0,0020	-0,0320	-0,0510	-0,0120

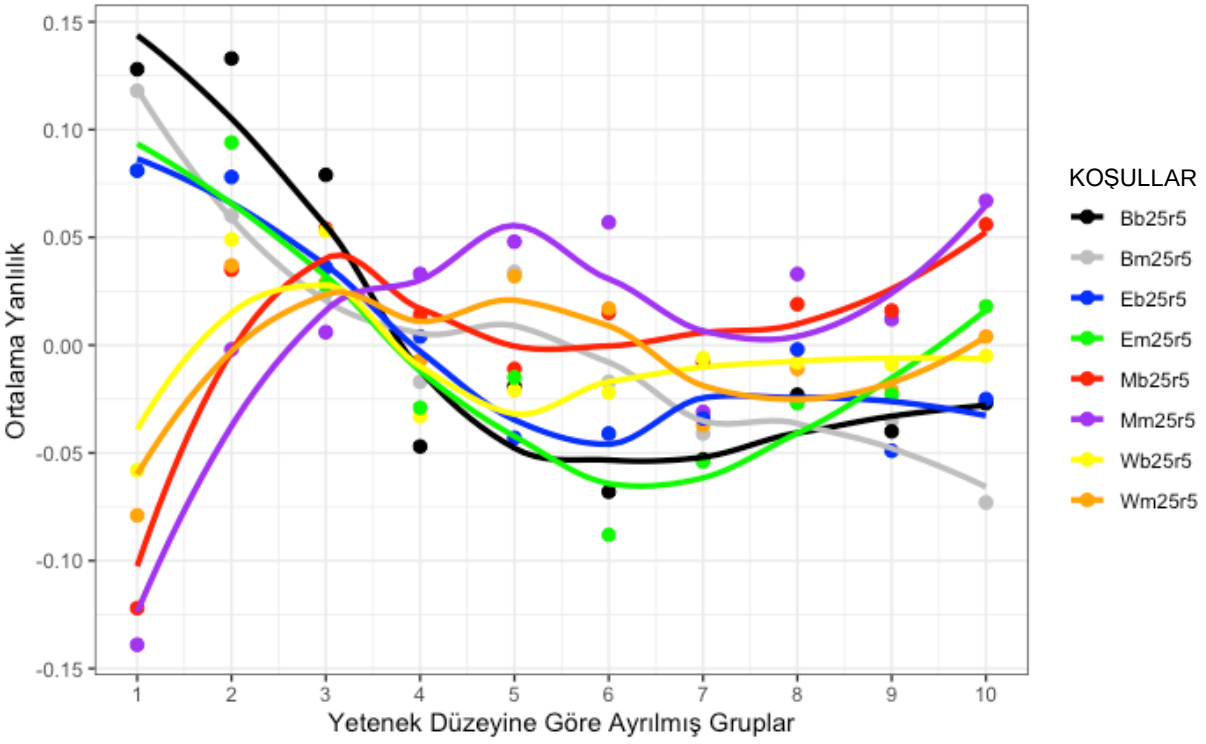
EK-K: Yetenek Düzeyine Göre Oluşturulan Gruplardaki Ortalama Test Uzunluğu

KOŞUL	OTL_D1	OTL_D2	OTL_D3	OTL_D4	OTL_D5	OTL_D6	OTL_D7	OTL_D8	OTL_D9	OTL_D10
Bm30	21,07	7,04	6,68	6,09	5,61	5,16	4,78	4,09	4,65	8,58
Bm25	28,05	11,37	10,75	9,30	8,44	7,40	6,77	5,84	6,68	12,93
Bm20	34,65	22,56	20,91	17,01	14,39	12,15	10,92	9,84	11,40	20,22
Bb30	21,09	10,74	11,66	10,92	10,10	9,34	8,78	8,31	8,50	11,47
Bb25	28,74	15,46	17,27	15,47	14,20	13,12	12,49	11,66	12,28	17,44
Bb20	34,66	25,83	27,15	24,18	21,70	19,97	18,56	17,80	19,55	25,93
Mm30	25,98	9,62	9,20	8,65	8,10	7,67	7,28	6,41	7,16	12,37
Mm25	31,26	13,86	13,27	12,05	10,91	9,92	9,27	7,83	8,83	16,57
Mm20	34,95	25,79	24,00	19,81	16,88	14,78	13,47	11,97	13,93	23,53
Mb30	26,20	14,00	15,06	14,42	13,76	13,17	12,36	11,77	12,21	16,45
Mb25	31,37	18,47	21,05	19,43	18,34	17,20	16,09	15,46	16,24	21,91
Mb20	34,95	29,30	30,73	28,18	25,84	24,30	22,23	21,65	23,05	29,51
Wm30	25,60	8,85	7,57	6,41	5,76	5,23	4,83	4,45	5,17	10,57
Wm25	30,81	12,92	11,63	9,92	8,59	7,62	6,89	6,29	7,33	15,13
Wm20	34,90	24,91	22,03	17,57	14,48	12,30	11,01	10,07	12,05	22,33
Wb30	25,85	11,40	12,52	11,25	11,06	10,21	9,88	9,69	10,61	14,95
Wb25	30,76	16,18	17,97	16,11	15,71	14,30	13,72	13,56	14,52	20,22
Wb20	34,94	27,54	28,14	24,50	23,17	20,64	19,77	19,38	21,35	28,09
Em30	19,92	7,83	7,35	7,08	6,52	6,33	6,16	5,79	6,30	9,76
Em25	28,58	12,09	11,91	10,78	9,62	9,16	8,58	8,11	8,61	14,77
Em20	34,59	23,83	21,98	18,64	15,57	14,19	12,49	11,79	13,21	22,09
Eb30	20,94	11,54	12,58	12,05	11,98	11,03	10,75	10,46	10,94	13,64
Eb25	29,02	16,30	18,58	17,01	16,48	15,14	14,71	14,27	15,12	19,77
Eb20	34,60	27,40	28,12	25,85	24,02	22,06	20,93	20,41	22,16	28,15
Bm30r3	21,21	7,56	7,20	6,36	5,91	5,66	5,25	4,72	5,25	8,54
Bm25r3	27,95	11,53	11,32	9,80	8,72	8,12	7,18	6,61	7,34	13,56
Bm20r3	34,67	22,81	21,16	17,51	14,66	12,71	11,41	10,28	11,74	20,76
Bb30r3	22,04	11,11	11,88	10,90	10,81	9,96	9,38	9,13	9,84	12,54
Bb25r3	29,32	16,14	17,65	15,48	15,53	13,89	13,11	13,02	13,75	18,35
Bb20r3	34,81	26,95	27,29	24,60	22,23	20,47	18,88	18,64	20,64	26,85
Mm30r3	26,23	10,18	9,59	8,95	8,55	8,30	7,57	6,77	7,77	13,11
Mm25r3	31,14	14,50	14,05	12,43	11,56	10,59	9,44	8,68	9,96	17,27
Mm20r3	34,92	26,52	24,01	20,44	17,06	15,38	13,62	12,46	14,27	23,68
Mb30r3	27,39	14,71	15,41	14,44	13,92	12,61	12,30	12,09	12,57	15,95
Mb25r3	31,88	19,09	20,58	19,57	18,49	16,80	16,36	15,36	16,40	21,34
Mb20r3	35,00	30,28	30,98	28,13	25,45	23,75	22,57	21,93	23,64	28,92
Wm30r3	25,24	8,72	8,02	6,90	6,35	5,75	5,37	5,16	5,76	10,61
Wm25r3	30,43	12,92	12,20	10,45	8,99	7,99	7,44	6,87	7,84	15,59
Wm20r3	34,91	25,01	22,59	18,07	14,76	13,33	11,43	10,59	12,54	22,25
Wb30r3	26,23	12,11	13,01	11,72	11,07	10,23	10,45	9,66	10,80	14,63
Wb25r3	31,46	17,38	18,63	17,21	15,91	14,43	14,15	13,71	14,44	20,73
Wb20r3	34,94	28,36	28,64	25,37	23,27	21,57	20,20	19,62	21,92	28,22

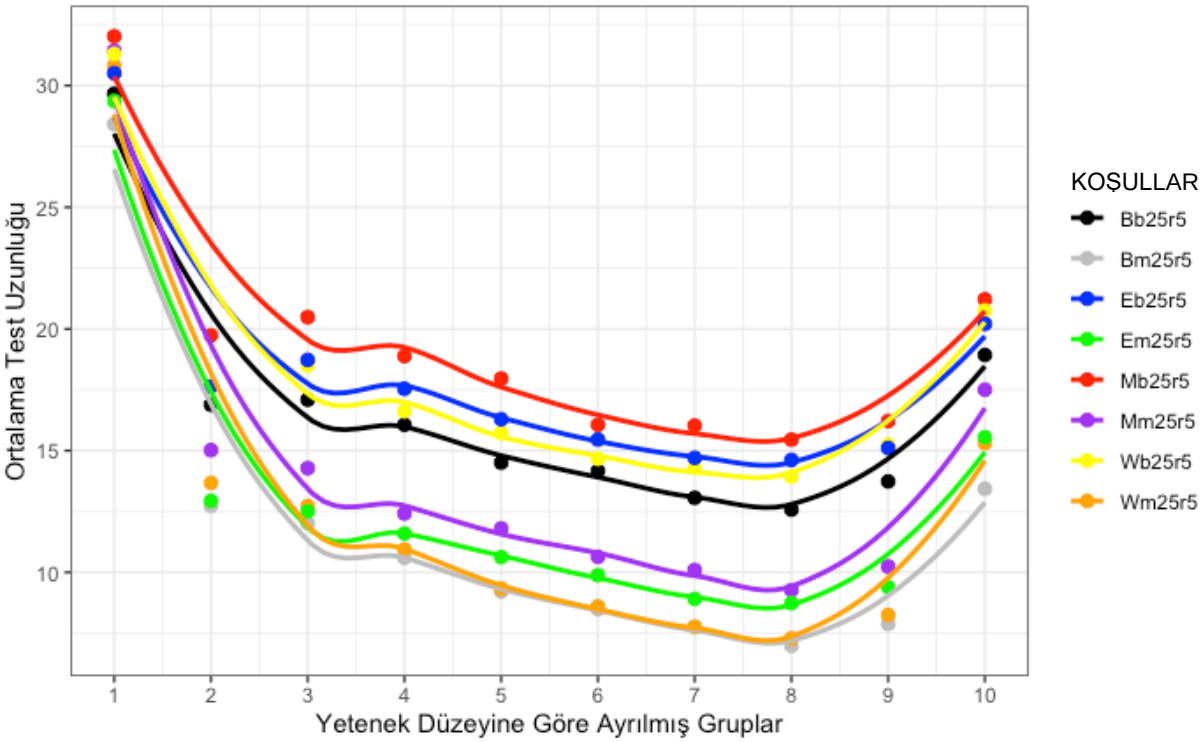
Em30r3	20,07	8,30	7,98	7,36	7,16	6,85	6,36	6,30	6,88	10,50
Em25r3	28,85	12,52	12,03	11,05	10,11	9,61	8,78	8,49	9,02	15,19
Em20r3	34,61	24,05	21,57	18,57	16,33	14,61	12,74	12,37	13,62	22,53
Eb30r3	21,70	12,40	13,46	12,68	12,65	11,78	11,22	10,78	11,48	14,00
Eb25r3	29,63	16,91	18,54	17,30	16,78	15,57	15,07	14,51	15,72	20,29
Eb20r3	34,74	28,19	28,34	26,64	24,13	21,85	20,97	21,14	22,32	28,21
Bm30r5	21,58	8,36	7,71	7,04	6,63	5,98	5,53	5,17	5,54	8,88
Bm25r5	28,42	12,73	12,02	10,60	9,21	8,49	7,75	6,98	7,88	13,43
Bm20r5	34,72	23,47	21,40	17,80	15,25	13,62	11,74	10,86	12,25	21,28
Bb30r5	22,98	11,80	12,21	11,40	10,99	10,12	9,72	9,13	9,99	12,67
Bb25r5	29,66	16,89	17,09	16,06	14,51	14,17	13,06	12,58	13,74	18,93
Bb20r5	34,85	28,00	27,51	24,29	22,13	20,23	19,17	18,77	20,43	27,00
Mm30r5	26,30	10,77	10,16	9,30	8,90	8,19	7,78	7,46	8,02	13,52
Mm25r5	31,43	15,02	14,28	12,42	11,81	10,64	10,09	9,27	10,24	17,50
Mm20r5	34,94	27,01	24,16	20,71	17,66	16,00	14,28	13,03	14,98	24,12
Mb30r5	27,58	14,85	15,04	14,26	13,51	13,25	12,18	12,23	12,36	16,06
Mb25r5	32,02	19,74	20,48	18,88	17,96	16,06	16,03	15,45	16,22	21,22
Mb20r5	34,95	31,05	30,68	27,31	25,89	23,63	22,68	21,70	23,51	28,98
Wm30r5	25,75	9,51	8,62	7,20	6,52	6,02	5,83	5,32	5,85	10,48
Wm25r5	30,81	13,67	12,73	10,95	9,34	8,60	7,77	7,29	8,25	15,31
Wm20r5	34,91	26,01	22,88	18,23	15,55	13,73	12,11	11,11	12,97	22,47
Wb30r5	26,66	12,57	13,04	12,32	11,59	10,62	10,59	10,59	11,07	14,86
Wb25r5	31,30	17,66	18,50	16,62	15,73	14,67	14,31	13,93	15,25	20,76
Wb20r5	34,96	29,26	28,78	25,63	22,94	21,34	19,80	19,92	22,24	28,36
Em30r5	20,88	9,01	8,41	7,90	7,67	7,23	6,74	6,65	7,03	11,26
Em25r5	29,36	12,94	12,50	11,60	10,63	9,88	8,91	8,73	9,40	15,56
Em20r5	34,71	24,59	21,91	19,41	16,30	14,75	13,10	12,49	13,88	22,37
Eb30r5	22,37	12,36	13,19	12,99	12,57	11,75	11,02	11,08	11,67	14,09
Eb25r5	30,50	17,64	18,73	17,54	16,29	15,46	14,71	14,62	15,11	20,21
Eb20r5	34,75	28,72	28,33	26,34	23,95	21,69	20,23	20,74	22,36	28,43

EK-L: “randomesque = 5” ve Hata Değeri .25 Olan Sekiz Koşulun Yetenek Düzeylerine Göre 10 Gruba Ayrılmış Bireylerdeki Ortalama RMSE, Yanlılık ve Test Uzunluğu Performansları

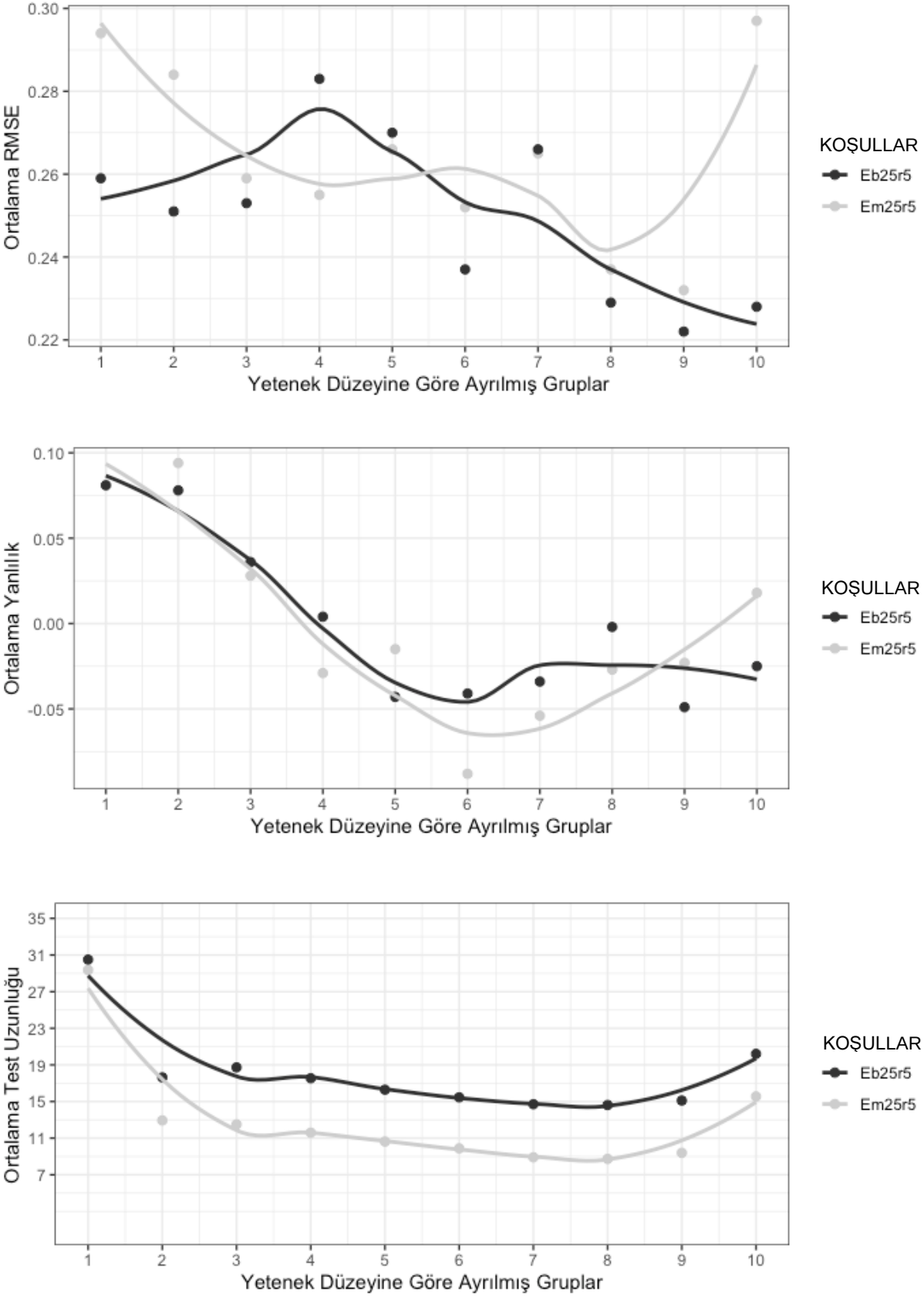




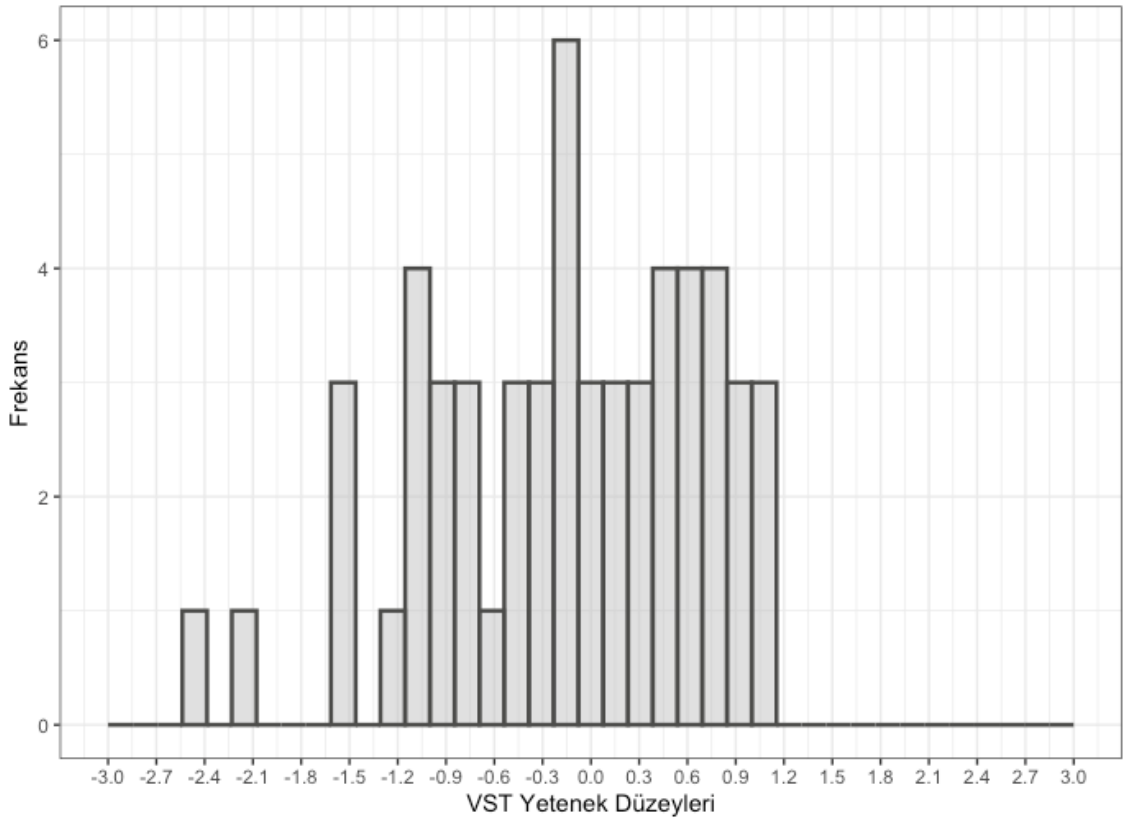
EK-L: Devamı

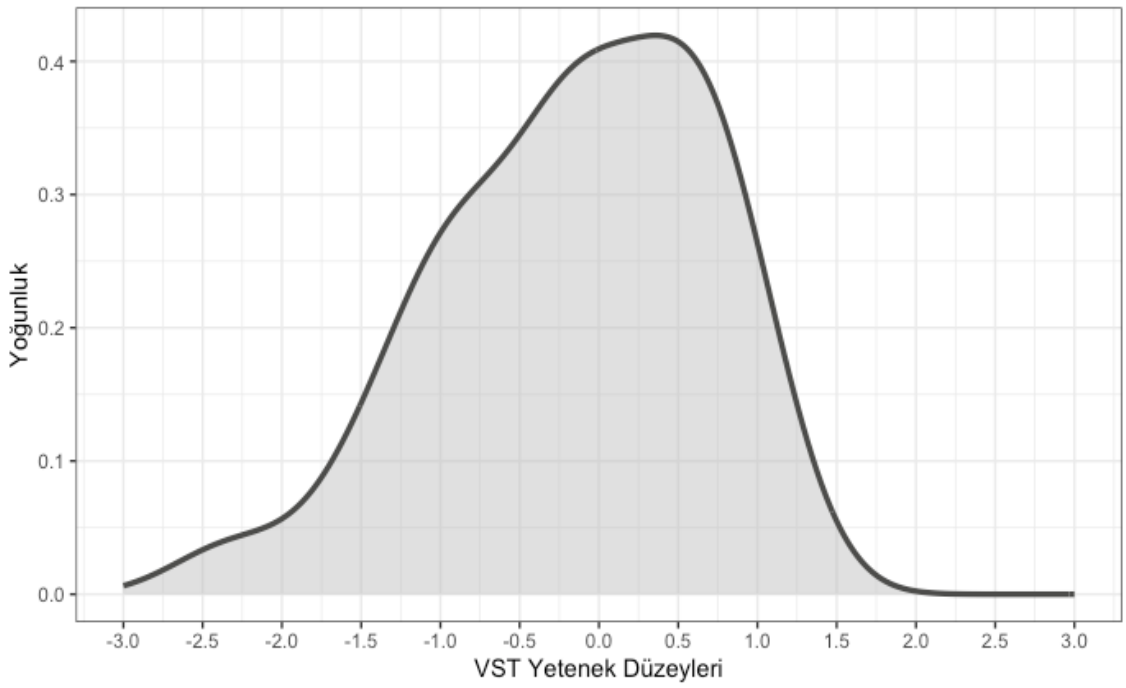

Ortalama Test Uzunluğu
30
25
20
15
10
1
2
3
4
5
6
7
8
9
10
Yetenek Düzeyine Göre Ayrılmış Gruplar
KOŞULLAR
Bb25r5
Bm25r5
Eb25r5
Em25r5
Mb25r5
Mm25r5
Wb25r5
Wm25r5

EK-M: Eb25r5 ve Em25r5 Koşullarının Yetenek Düzeylerine Göre 10 Gruba Ayrılmış Bireylerdeki Ortalama RMSE, Yanlılık ve Test Uzunluğu Performansları



EK-N: BBT Aşamasında VST'nin Kâğıt-Kalem Versiyonuna Verilen Cevaplardan Kestirilen Yetenek Düzeylerine Ait Histogram ve Yoğunluk Grafikleri





EK-O: Araştırma Etik Komisyonu Onay Bildirimi

Tarih: 16/06/2020
Sayı: 35853172-300-E.00001113493

0001113493

T.C.
HACETTEPE ÜNİVERSİTESİ
Rektörlük

Sayı : 35853172-300
Konu : Arş. Gör. Mustafa GÖKCAN (Etik Komisyon İzni)

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Enstitünüz Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı Doktora programı öğrencisi **Arş. Gör. Mustafa GÖKCAN'**ın **Dr. Öğr. Üyesi Derya ÇOBANOĞLU AKTAN** danışmanlığında yürüttüğü **"İngilizce Kelime Bilgisinin Bireyselleştirilmiş Bilgisayarlı Test Uygulaması ile Ölçülmesi"** başlıklı tez çalışması Üniversitemiz Senatosu Etik Komisyonunun **09 Haziran 2020** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

Bilgilerinizi ve gereğini saygılarımla rica ederim.

e-imzalıdır
Prof. Dr. Rahime Meral NOHUTCU
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 İnternet Adresi: www.hacettepe.edu.tr

Sevda TOPAT

EK-Ö: 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.

…..…/……/…….

(İmza)

Mustafa GÖKCAN

EK-P: Doktora Tez Çalışması Orijinallik Raporu

14/08/2023

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

Tez Başlığı İngilizce Kelime Bilgisi Testinin (VST) Bireyselleştirilmiş Bilgisayarlı Test Olarak Uygulanabilirliğinin İncelenmesi

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
14/08/2023	154	213,147	19/06/2023	%6	2145710040

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 Esaslarını 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: Mustafa Gökcan

Öğrenci No.: N16149922

Ana Bilim Dalı: Eğitim Bilimleri

Programı: Eğitimde Ölçme ve Değerlendirme

Statüsü: ☐ Y.Lisans ☒ Doktora ☐ Bütünleşik Dr.

İmza

DANIŞMAN ONAYI

UYGUNDUR.

(Doç. Dr. Derya ÇOBANOĞLU AKTAN)

EK-P: Dissertation Originality Report

14/08/2023

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

Thesis Title: Investigation of the Applicability of the Vocabulary Size Test (VST) as a Computerized Adaptive Testing

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
14/08/2023	154	213,147	19/06/2023	6%	2145710040

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: Mustafa Gökcan

Student No.: N16149922 Signature

Department: Educational Sciences

Program: Educational Measurement and Evaluation

Status: ☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED
(Assoc. Prof. Derya ÇOBANOĞLU AKTAN)

EK-R: Yayımlama ve Fikrî 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 hakları 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)

..... /..... /.....

(imza)

Mustafa GÖKCAN

"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.*