

Uzaktan Eğitimde Bilgisayar Uyarlamalı Test ve Kağıt-Kalem Testi ile Yetenek Düzeylerinin Belirlenmesi: Çok Merkezli Gerçek Veri Uygulaması

Determining Ability Levels by Using Computerized Adaptive Testing and Paper-And-Pencil Testing in the Distance Education: A Multi-Centred Real Data Study

Yusuf Kemal ARSLAN* (ORCID: 0000-0003-1308-8569)

Mergül ÇOLAK** (ORCID: 0000-0002-4762-8298)

Ülviye BİLGİN*** (ORCID: 0000-0001-5871-0089)

*Erzincan Binalı Yıldırım Üniversitesi Tıp Fakültesi, Erzincan, TÜRKİYE

**Erzincan Binalı Yıldırım Üniversitesi Spor Bilimleri Fakültesi, Erzincan, TÜRKİYE

***Gazi Üniversitesi Spor Bilimleri Fakültesi, Ankara, TÜRKİYE

Sorumlu Yazar: Yusuf Kemal ARSLAN, E-Posta: ykarslan@gmail.com

Özet

Amaç: İçinde bulunduğumuz Covid-19 pandemi sürecinde uzaktan eğitimin popüler hale gelmesi ile yapılan çevrimiçi sınavlarda kişilerin yetenek düzeyini daha doğru şekilde belirlemek ve sınav stratejilerini çözümlmek için bilgisayar uyarlamalı test kullanımı giderek önem kazanmaktadır. Bu çalışmada, uzaktan eğitim sürecinde öğrencilerin yetenek düzeylerinin bilgisayar uyarlamalı test ve kağıt-kalem testi ile kestirimlerinin karşılaştırılması amaçlanmıştır.

Anahtar sözcükler:
Bilgisayar Uyarlamalı Test, Kağıt-Kalem Testi, Uzaktan Eğitim, Spor Bilimleri, Alan Sınavı

Keywords:
Computerized Adaptive Testing, Paper-And-Pencil Test, Distance Learning, Sports Sciences, Field Exam

Gönderilme Tarihi
Submitted: 03.10.2021
Kabul Tarihi
Accepted: 22.10.2021

Yöntem: Araştırma, Mart 2020 – Aralık 2020 tarihleri arasında gerçekleştirilen kesitsel ve metodolojik bir araştırmadır. Araştırmanın çalışma evrenini Erzincan Binalı Yıldırım Üniversitesi, Gazi Üniversitesi, Erciyes Üniversitesi ve Kastamonu Üniversitesi Spor Bilimleri Fakültelerinde öğrenim gören beden eğitimi ve spor öğretmenliği 4. sınıf öğrencileri oluşturmaktadır. Bu çalışmada R programlama dili (ver. 3.6.2) ve RStudio (ver. 1.2.5033) yazılımı kullanılarak analizler gerçekleştirilmiştir. Kişi sayısı 176; madde bankasındaki madde sayısı 80 (4 farklı alanın (egzersiz fizyolojisi, psikomotor gelişim, antrenman bilgisi, egzersiz ve beslenme) her birinden uzmanlar tarafından hazırlanan 20'şer soru ile oluşturulmuştur); bilgisayar uyarlamalı test durdurma kriteri için standart hata $<0,5$ olarak alınmıştır. Kağıt-kalem testi ve bilgisayar uyarlamalı test θ değerlerinin uyumları sınıf içi korelasyon katsayısı ile

değerlendirilmiştir. Kişilere ait bilgisayar uyarlamalı test kestirimleri "randomCAT" fonksiyonu ile gerçekleştirilmiştir. Bilgisayar uyarlamalı testten elde edilen kişinin yetenek düzeyi, θ 'ya ait standart hata ve kullanılan madde sayısı analiz sonucunda sunulmuştur. Bilgisayar uyarlamalı testte madde seçimi için maksimum Fisher bilgi kriteri kullanılmıştır. Yetenek düzeyi kestirimlerinde iteratif yaklaşımlar için en çok olasılık kestirimi kullanılmıştır.

Künye: Arslan YK, Çolak M, Bilgin Ü. Uzaktan Eğitimde Bilgisayar Uyarlamalı Test ve Kağıt-Kalem Testi ile Yetenek Düzeylerinin Belirlenmesi: Çok Merkezli Gerçek Veri Uygulaması. Tıp Eğitimi Dnyası. 2022;21(63):95-103

Bulgular: Klasik test skoru (0-100 skalasında) ve bilgisayar uyarlamalı test arasında orta düzeyde uyum saptanmıştır (sınıf içi korelasyon katsayısı:0,682, güven aralığı %95GA=0,594-0,765, F(175,175)=17,1; p<0.001). Çalışmada, kağıt kalem testinden elde edilen cevaplardan madde yanıt teorisi doğrultusunda elde edilen yetenek düzeyleri ve bilgisayar uyarlamalı test ile yetenek düzeyi kestirimleri incelendiğinde uyumun mükemmel düzeyde (sınıf içi korelasyon katsayısı:0,90; %95GA:0,866-0,924) olduğu görülmüştür. Ayrıca bilgisayar uyarlamalı testin kağıt-kalem testine göre soru sayısını %80 oranında azalttığı bulunmuştur.

Sonuç: Bu çalışmadan elde edilen sonuçlar, bilgisayar uyarlamalı test uygulaması ile daha kısa sürede, daha az madde sayısı ile güvenli sonuçlar elde edilebileceğini ve testin tamamlanmasından kısa süre sonra sonuçların izlenebileceğini göstermektedir.

Abstract

Aim: *With the popularity of distance education in the current Covid-19 pandemic, the use of computerized adaptive testing is becoming increasingly important in order to determine the ability level of individuals more accurately and to analyze the exam strategies in online exams. In this study, it is aimed to compare the estimations of students' ability levels in the distance education with computerized adaptive testing and paper-pencil test.*

Methods: *This study is a cross-sectional and methodological study which was conducted between March 2020 and December 2020. The study population consists of 4th class students which were studying physical education and sports teaching at Erzincan Binali Yıldırım University, Gazi University, Erciyes University and Kastamonu University Sports Sciences Faculties. In this study, analyzes were performed by using R programming language (ver. 3.6.2) and RStudio (ver. 1.2.5033) software. Number of people was 176; the number of items in the question bank was 80 (formed with 20 questions each prepared by experts from each of 4 different fields (exercise physiology, psychomotor development, training information, exercise and nutrition)); the standard error for the computerized adaptive test stopping criterion was taken as <0.5. The agreement of Paper-And-Pencil Test and computerized adaptive test θ values were evaluated with the intraclass correlation coefficient. CAT estimates of individuals were performed with the "randomCAT" function in the catR package. The ability level of a person was obtained with computerized adaptive test application, the standard error of θ and the number of items used were presented as a result of the analysis. Maximum Fisher information criterion was used for question selection in computerized adaptive test. Maximum likelihood estimation was used for iterative approaches in the ability level estimations.*

Results: *There was moderate agreement between the classical test score (on a 0-100 scale) and the computer-adapted test (intraclass correlation coefficient: 0.682, confidence interval 95%CI=0.594-0.765, F(175.175)=17.1; p<0.001). In this study, the ability level estimations by using the item response theory from the answers obtained from the paper-and-pencil test and the ability level estimations with the computer adaptive test were examined. It was seen that the agreement was at an excellent level (intraclass correlation coefficient: 0.90; 95%CI: 0.866-0.924). Also it was also found that computerized adaptive test reduced the number of addressed questions by 80% when compared to Paper-And-Pencil Test.*

Conclusions: *The results obtained from this study show that with the application of computer-adapted testing, safe results can be obtained in a shorter time, with less number of items, and the results can be monitored shortly after the completion of the test.*

GİRİŞ

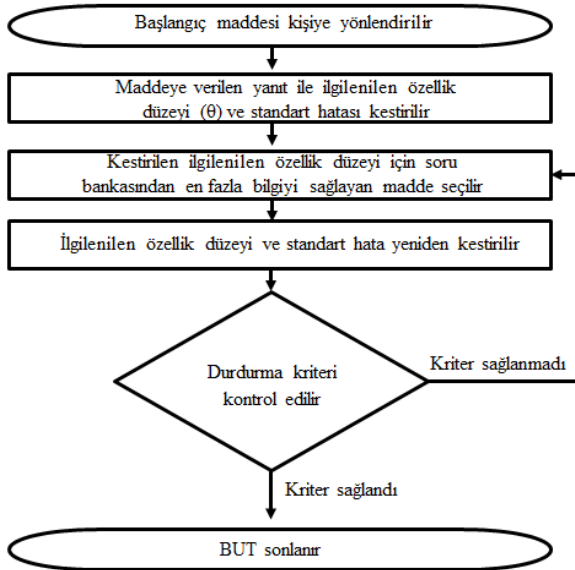
Günlük yaşamımızda ve bilimsel alanda ölçme önemli bir yer tutar. Ölçme, genel anlamıyla "herhangi bir niteliği gözlemek ve gözlem sonuçlarını sayılarla veya başka sembollerle"

ifade etmektir. Bir hastanın boyunun/kilosunun ölçülmesi fiziksel bir ölçme olup "doğrudan ölçme" kapsamına girerken, aynı hastanın engellilik düzeyinin ölçülmesi ise psikolojik

ölçme olup “dolaylı ölçme” kapsamına girmektedir (1). Ölçme araçlarından elde edilen sonuçlar, kişilerin incelenen özellikleri hakkında karar vermede kullanılır.

Ölçme alanında sıklıkla kullanılan iki teoriden biri olan Madde Yanıt Teorisi (MYT-Item Response Theory), klasik test teorisi ile karşılaştırıldığında ilgili alanda elde edilen ölçümlerin belirlenmesi için birçok avantaja sahiptir. Klasik test teorisinde elde edilen istatistikler, madde zorluk parametresi (doğru yanıt oranı), madde ayırt edicilik parametresi (düzeltilmiş madde-toplam test korelasyonu) ve güvenilirlik, ölçme aracının uygulandığı gruba bağımlıdır. MYT’de ise parametre tahminleri kullanılan örnekleme bağımlı değildir ve araştırma evrenindeki farklı gruplar için değişmediği varsayılır. Ayrıca, klasik test teorisi güvenilirlik için tek bir tahmin ve standart hata ölçümü verirken MYT, ölçme aracının kesinliğini, incelenen özellik düzeyi boyunca tahmin eder. Klasik test teorisinin diğer bir dezavantajı, kişinin ölçüm aracındaki sonucunun analizde kullanılan maddelere bağımlı olmasıdır. MYT’de ise, kişinin

incelenen özelliği, kullanılan maddelerden bağımsızdır. Kişinin incelenen özelliği, her bir maddeye verdiği yanıtlardan hesaplandığı için kestirim skoru, klasik test teorisinden elde edilen toplam skordan daha doğru bir kestirimdir (2). Bilgisayar uyarlamalı test (BUT, Computerized Adaptive Test-CAT), ilgilenilen özelliğin ölçümünde sıklıkla kullanılan, MYT temelli ve etkili bir yöntemdir. “Uyarlamalı test” yöntemi, testi oluşturan maddelerin, uygulama sırasında seçildiği bir süreçtir ve uyarlamalı testlerde kişinin yetenek düzeyine uygun maddeler yönlendirildiği için kişiye uyarlanmış bir test yapısı oluşur. Şekil 1’de görüldüğü gibi, BUT iteratif bir yaklaşım olup; maddelerin kişilere uyarlanma işi ardışık olarak yapılır. Maddelerin kişilere yönlendirilme sırası kişinin önceki maddedeki performansına göre düzenlenir. Bu süreç boyunca testi alan kişilerin yetenek düzeyi sürekli olarak değerlendirilir. BUT uygulamasında başlangıçta kişiye rastgele bir madde yönlendirilebilir, belirli maddeler sabit tutularak uygulama başlatılabilir veya orta zorluk düzeyinde bir madde sorulabilir.



Şekil 1. BUT Akış Şeması

BUT uyarlanabilir bir test olması sebebiyle kağıt-kalem testine (KKT) göre uygulanan madde sayılarını %50 azalabilir ve benzer yetenek düzeyi kestirimleri yapılabilir (3,4). Literatürde araştırmacılar, BUT kullanımıyla ilgili birtakım avantajlardan bahsetmişlerdir. Bunlardan ilki test yönetiminin esnek olmasıdır. Sınav sonuçları sınav bitiminden hemen sonra elde edilebilir ve bu durum sınava girenleri motive edebilir (4). Testteki maddeler katılımcının seviyesine uygun olduğu için kaygı düzeyi azaltılabilir (5). BUT'un uygulanması kâğıt ve kalem (paper-and-pencil, klasik test) testlerine göre zaman tasarrufu sağlar. Çünkü BUT'ta kullanılan madde sayısı ve bu soruları cevaplamak için gereken süre, geleneksel kâğıt ve kalem testlerine göre daha azdır (6-8). Her test, sınava giren her bir kişinin seviyesine göre uyarlandığından, geleneksel testlere kıyasla BUT'tan daha fazla bilgi toplanabilir (9). Diğer test türleri gibi BUT'un da kendine özgü dezavantajları vardır. Bu nedenle, testi uygulayan araştırmacılar bu kısıtlamaların farkında olmalıdır.

BUT uygulaması, büyük hedef kitleler ile yürütülen bilimsel ve araştırmaya dayalı alanlarda tercih edilmektedir. BUT, engellilik veya hastalıkların erken başlangıcını yakalamak için de kullanılabilir. Dijital hastanelerin ve büyük verinin yaygınlaştığı günümüzde sağlık alanında kullanılan ölçekler için de BUT uygulamaları kullanılabilir.

Literatürde araştırmacılar, BUT kullanımıyla ilgili birtakım dezavantajlardan bahsetmişlerdir. BUT'lar tüm madde türleri için kullanılmayan MYT modeline dayanmaktadır. Bu nedenle, açık uçlu sorulara ve kolayca kalibre edilemeyen maddelere uygulanamazlar (4). Buna ek olarak diğer önemli bir dezavantaj, sınava giren kişinin cevapladığı sorulara geri dönmesine ve cevapları değiştirmesine izin verilmemesidir, çünkü sonraki maddeler önceki maddelere göre seçilir (4). Ayrıca, madde kalibrasyonu, bir BUT'un başarısını etkileyen önemli faktördür. Maddeler, zorluk/yetenek ölçeğine uygun şekilde kalibre edilmemişse,

uygulanan test ne geçerli ne de güvenilir olacaktır. Bazı araştırmacılara göre tek boyutluluk BUT'un bir dezavantajıdır (7). Tek boyutluluk, tüm test maddelerinin tek bir özelliği başka özelliklere karıştırmadan ölçmesi anlamına gelir (5). Bu varsayım testler için her zaman sağlanmalıdır. BUT'ta, bazı maddeler diğerlerine göre daha sık seçilebilir ve bu maddeler ezberlenip diğer sınava giren kişilere aktarılabilir (8). Dolayısıyla, güvenlik BUT'ların kısıtlılıklarından biri olabilir. Bu nedenle hem güvenlik kısıtını engellemek hem de sınava girenlerin bilgilerini test etmek için geniş bir madde havuzu kullanılmalıdır. Bu çalışmada Covid-19 pandemisinde uzaktan eğitim sürecinde BUT ve KKT ile beden eğitimi öğrencilerinde yetenek düzeylerinin belirlenmesi amaçlandı. Güncellenen öğretmenlik alan bilgisi sınavı sebebiyle beden eğitimi ve spor öğretmen adaylarının yetenek/bilgi düzeylerinin uygun yöntemlerle ölçülmesinin, adayların sınava hazırlık aşamasında sergileyecekleri performanslarda, anlık bilgi düzeylerinin belirlenmesinde ve sınav başarısında etkili olduğu düşünülmektedir.

GEREÇ VE YÖNTEM

Araştırma, Mart 2020 – Aralık 2020 tarihleri arasında gerçekleştirilen kesitsel ve metodolojik bir araştırmadır.

Katılımcılar

Araştırmanın çalışma evrenini Erzincan Binali Yıldırım Üniversitesi, Gazi Üniversitesi, Erciyes Üniversitesi ve Kastamonu Üniversitesi Spor Bilimleri Fakültelerinde öğrenim gören beden eğitimi ve spor öğretmenliği 4. sınıf öğrencileri oluşturmaktadır. Araştırmada herhangi bir örneklem seçim yöntemi kullanılmamış, araştırmaya katılmayı gönüllü olarak kabul eden 176 öğrencinin tamamı araştırma kapsamına alınmıştır. Gönüllü olarak çalışmaya katılmak isteyen öğrencilerden aydınlatılmış onam formu alınmıştır Erzincan Binali Yıldırım Üniversitesi insan araştırmaları

etik kurulundan 27/02/2020 tarihli ve 02/01 protokol no'lu etik kurul izni alınmıştır.

Veri Toplama Araçları

Araştırmada veri toplama aracı olarak 80 adet çoktan seçmeli teorik sorudan oluşturulmuş madde havuzu kullanıldı. Alan uzmanı olan bir araştırmacı tarafından 4 farklı konuda hazırlanan (antrenman bilgisi, egzersiz ve beslenme, egzersiz fizyolojisi ve psikomotor gelişim) 20 şer soru eşit olarak ağırlıklandırıldı ve alan uzmanı olan bir diğer farklı kişi soruların kapsam geçerliğini belirlemede mantıksal yaklaşım kullanarak hem bilimsel hem de biçimsel olarak soruları değerlendirdi.

Veri Analizi

Bu çalışmada R programlama dili (ver. 3.6.2) ve RStudio (ver. 1.2.5033) yazılımı kullanılarak analizler gerçekleştirilmiştir (10). Kişi sayısı 176; madde havuzundaki madde sayısı 80; BUT durdurma kriteri için standart hata <0,5 olarak alınmıştır. KKT ve BUT θ değerleri ile uyumları sınıfıçı korelasyon katsayısı (SKK) ile (ICC(C,1)) değerlendirilmiştir. SKK yorumlanırken Cicchetti'nin (1994) kategorizasyonu kullanılmıştır (11). SKK *irr* paketindeki "icc" fonksiyonu kullanılarak hesaplanmıştır (12). Yöntemler arasındaki uyum değerlendirilirken Bland-Altman grafiği "BlandAltmanLeh" paketi yardımıyla çizdirilmiştir (13). Klasik test teorisinde kişiye sorulan 80 soruda her soru 1,25 puan olacak şekilde kişi en düşük 0 puan, en yüksek 100 puan alabilir.

İki Sonuçlu Rasch Modeli İle Madde Zorluk Parametre Kestirimi

İncelenen özellik düzeyi (θ) kestirimi, klasik istatistik ve Bayesci istatistik yaklaşımları kullanılarak yapılabilir. En sık kullanılan kestirim yöntemleri, klasik istatistik çerçevesinde geliştirilen En Çok Olabilirlik (maximum likelihood, ML) ve Bayesci istatistik çerçevesinde geliştirilen Sonsal Beklenti Kestirimidir (SBK) (expectation a posteriori,

EAP) (14). Çalışmanın asıl amacı olan KKT ve BUT kestirimleri Rasch modeli ile yapılmıştır. MYT altında yer alan Rasch modeli tek parametrelidir bir model olup, bu modelde madde zorluk parametresinin (b_j) yanı sıra kişinin θ düzeyi de bulunmaktadır. Georg Rasch tarafından iki kategorili maddeler (0,1) için geliştirilen Rasch modelinde (Eşitlik 1), kişinin bir maddeye belirli bir yanıtı verme olasılığı, kişinin θ düzeyi ile madde zorluk parametresi arasındaki farkın lojistik fonksiyonu olarak hesaplanır (15).

$$P_j(\theta_i) = P(Y_{ij} = 1 | \theta_i) = \left(\frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \right) \quad (\text{Eşitlik 1})$$

Havuzdaki maddelere ait zorluk parametresi En çok olabilirlik yöntemi ile kestirilirken *irtplay* paketindeki "est_irt" fonksiyonu kullanılmıştır ve a parametresi (discrimination parameter) 1 olarak kabul edilmiştir. Kişi yetenek düzeyleri MYT kapsamında Rasch modeli ile kestirilmiş ve bu kestirim işleminde *irtplay* paketindeki "est_score" fonksiyonu kullanılmıştır.

Prosedür/Veri Toplama Aracının Uygulanması

Veri toplama aracı olan sorular öğrencilere bilgisayarda oluşturulmuş arayüz yardımı ile uygulanmıştır. Uygulama arayüzü oluşturulurken Microsoft Visual Studio 2012'den faydalanılmıştır. Tüm öğrenciler (n=176) tüm soruları (n=80) yanıtlamış ve bu durum KKT olarak kabul edilmiştir. Uygulama tek seferde çevrimiçi olarak gerçekleştirilmiştir. Uygulama sonucunda 176 kişinin 80 maddeye verdikleri tüm yanıtlara ulaşılmıştır. Bu çalışmada kişilere BUT uygulanması durumunda yetenek düzeyi kestirimlerinin nasıl olacağını belirlemek adına *catR* paketindeki "randomCAT" fonksiyonu ile BUT uygulaması gerçekleştirilmiştir. BUT için veri seti olarak kişilerin tüm maddelere verdikleri yanıtlar kullanılmıştır. BUT uygulamasından elde edilen belirli bir kişinin yetenek düzeyi, θ 'ya ait standart hata ve kullanılan madde sayısı analiz sonucunda sunulmuştur. BUT'ta soru seçimi

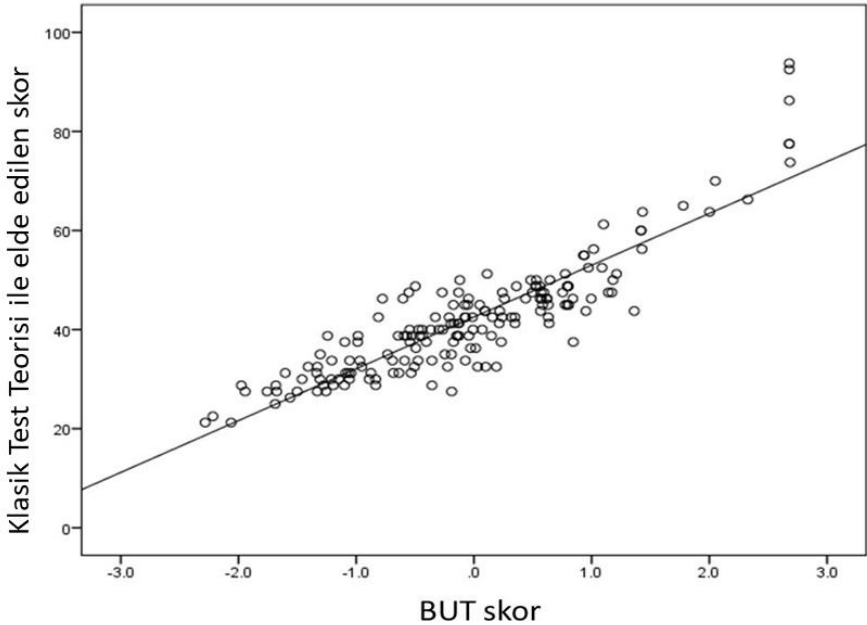
için maksimum Fisher bilgi kriteri kullanılmıştır. Yetenek düzeyi kestirimlerinde iteratif yaklaşımlar için en çok olabilirlik kestirimi kullanılmıştır. Bu adımsal yapıda bir kişinin başlangıç θ 'sından, log-olabilirliğin 1. türevinin 2. türevine oranı çıkartılır ve son-teta kestirilir. Bu iteratif süreç teta kestirimleri arasındaki fark önemsenmeyecek boyuta ($<0,0001$) gelene kadar devam eder ve θ 'ya ait ML kestirimi yapılmış olur. Bilgisayar uyarlamalı testlerde süreç katılımcıya ilk maddeyi yönlendirerek başlar. Başlangıçta sorulan orta zorluk düzeyine sahip maddeyi doğru yanıtladıysa, takip eden madde daha zor, yanlış yanıtladıysa takip eden madde daha kolay olur. BUT yaklaşımıyla, katılımcının ölçülen özelliğini (θ) kestirmede çok etkili olmayan uç noktadaki zorluk düzeyine sahip maddeler (çok zor veya çok kolay) yerine, katılımcının yetenek düzeyine uygun olan maddelerin seçilmesi sağlanır. Kişinin uygulanan orta zorluk düzeyli maddeye verdiği yanıt

sonrasında ilgilenilen özellik kestirilir. Bu MYT'de yetenek düzeyi (θ) olarak tanımlanır ve diğer maddelerden elde edilebilecek bilgi düzeyleri hesaplanır. En yüksek bilgiyi sağlayan madde katılımcıya uygulanır ve aynı işlem sürecin durması için belirlenen kriter sağlanana kadar devam eder. Bu durdurma kriterleri; süre, madde sayısı, $\hat{\theta}$ 'daki değişim veya $\hat{\theta}$ 'ya ait standart hatanın belirli bir eşik değerinin altına inmesi olabilir.

BULGULAR

Klasik test teorisi ve BUT

Klasik test teorisi ile elde edilen 0-100 puan arasındaki ham puanlar ve BUT ile kestirilen yetenek düzeyleri arasında pozitif yönde ilişki olduğu izlenmiştir (Şekil 2). Klasik test teorisi ile MYT temelli BUT arasında orta düzeyde uyum olduğu (sınıf içi korelasyon katsayısı:0,682 ve %95GA:0,594-0,765, F (175,175)=17,1; $p<0.001$) saptanmıştır.

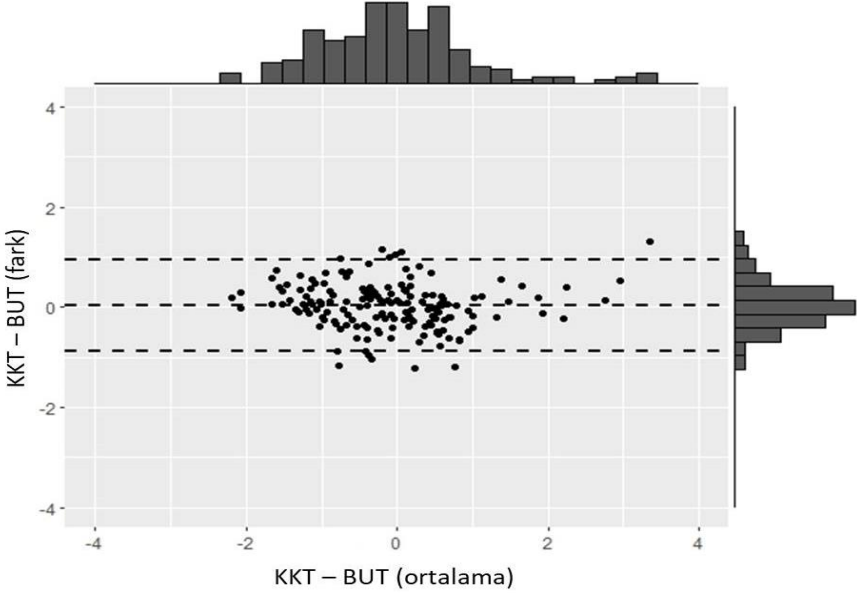


Şekil 2. Klasik Test Teorisi ve BUT ile Elde Edilen Skor Arasındaki İlişki

KKT ve BUT

Kişilerin KKT sonucunda 80 sorunun tümüne verdiği yanıtlardan Rasch modeli ile gerçek teta'lar kestirilmiştir. Gerçek teta'lar ile BUT yöntemiyle elde edilen teta'ların farkından hesaplanan yanlılığa ait (bias) ortalama değer $0,036 \pm 0,470$ $(-1,225; 1,321)$ olarak bulunmuştur. Bland-Altman grafiği incelendiğinde KKT ve BUT yöntemleri

arasında uyum olduğu ve grafikte orantısal bir yanlılık olmadığı izlenmektedir. BUT uygulaması ile yapılan kestirimlere ait ortalama yanlılık sıfıra çok yakın olarak saptanmıştır (Şekil 3). KKT ve BUT yetenek düzeyi kestirimleri için uyumun mükemmel düzeyde (SKK:0,90; %95GA:0,866-0,924, $p < 0,001$) olduğu saptanmıştır.



Şekil 3. KKT ve BUT ile Elde Edilen Skorlara Ait Bland-Altman Grafiği

BUT'ta yetenek düzeyi kestirilirken yapılan ortalama standart hata $0,491 \pm 0,007$ (0,468-0,500) olarak hesaplanırken, kullanılan ortalama madde sayısı $16,4 \pm 4,3$ (14-40) olarak hesaplanmıştır (BUT'ta bir kişi için kullanılan madde sayısı en fazla 40 idi). BUT uygulamasında kullanılan madde sayısı KKT'nin %50'si idi, buna göre BUT'ta uygulanan madde sayısı KKT'ye göre yaklaşık olarak %80 oranında (16/80) azalmaktadır.

TARTIŞMA

Teknoloji ve bilgi çağındaki gelişmelere paralel olarak kâğıt-kalem testleri yerini BUT'a

bırakmaktadır. Gelişen bilgisayar teknolojisi yardımıyla bu testlerde yalnızca kişinin yanıt deseni ile ilgili bilgi elde edilmez, bu duruma ek olarak yanıt süreleri ve kişinin maddeyi yanıtladığı andaki davranış şekli ile ilgili bilgi toplanabilir (16). BUT uygulaması sınava giren kişi için bir zaman tasarrufu sağlar; kişiler çok zor veya önemsiz derecede kolay olan maddeleri yanıtlamak için zamanı boşa harcamazlar. BUT kullanımı ile test hazırlama ve işaretleme için gereken süre azaldığından sonuçların tutarlılığı artabilir (6). BUT'ta kullanılan madde sayısı ve bu soruları cevaplamak için gereken süre geleneksel kâğıt

ve kalem testlerine göre daha azdır (7,8). Bu çalışmada beklendiği gibi BUT uygulamasında madde sayısının klasik teste göre çok daha az olduğu görülmüştür. BUT ve kağıt kalem yöntemlerinde madde sayılarının değişmesinin dışında, test performansı ve madde kestirimleri ile ilgili yapılan çalışmalar mevcuttur. Bu çalışmalarda farklı çıkarımlar söz konusudur. Bennet ve ark., BUT ve kâğıt kalem testleri arasında anlamlı farklılıklar bulurken (17), Threlfall ve ark., bu farkın minimal olduğunu (18), Wang ve ark. ise test modunun hiçbir etkisinin olmadığını bulmuştur (19). Şenel (2018) yaptığı kantitatif analiz sonucunda BUT ile yetenek düzeyi kestirimlerinin kâğıt-kalem testine göre daha düşük olduğunu bildirmiştir (20). Bu çalışmada da BUT kestirim değerleri KKT değerlerine göre benzer şekilde daha düşük bulunmuştur, fakat bu farklılığa ait ortalama değer sifıra çok yakındır (0,036 logit). Ayrıca KKT ve BUT için uyum grafiklerinde orantısal bir yanlılığa rastlanmamıştır.

Çalışmanın kısıtları: Bu çalışmada BUT ve KKT performanslarını karşılaştırmak için Rasch modeli kullanılmıştır. Rasch modelinden farklı olarak alternatif MYT modelleri ile (2PLM, 3PLM gibi) yeni çalışmalar planlanabilir fakat özellikle 3PLM için çok yüksek örneklem ölçümüne ihtiyaç duyulmaktadır (21). Sağlık alanında çevrimiçi yapılan sınavlarda henüz yeterince büyük sayıda madde havuzlarının oluşturulamamış olması yüksek örneklemli uygulamaya engel oluşturmuştur. Bu çalışmada literatürde sıklıkla kullanılan maksimum Fisher bilgi kriteri dışındaki farklı madde seçim kriterleri (geleneksel yaklaşım; olabilirlikle ağırlıklandırılmış maksimum bilgi, bayesci yaklaşım; Sonsal Dağılımlı Ağırlıklandırılmış Maksimum Bilgi, Beklenen Maksimum Bilgi, Beklenen Minimum Sonsal Varyans, Sonsal Dağılımlı Ağırlıklandırılmış Beklenen Maksimum Bilgi) ve daha yüksek örneklem ölçümlü gerçek veri setleri kullanılarak gelecekte farklı çalışmalar gerçekleştirilebilir.

Madde bankasında dört alt alana ait sorular eşit ağırlıklı olmasına rağmen BUT'ta 4 alt alanı temsil edecek şekilde içerik dengeleyerek (content balancing) öğrencilere soru yönlendirilememiş olması çalışmanın bir diğer sınırlılığıdır.

Bu çalışmada, gerçek bir veri seti üzerinde sınanan teorinin güncellenen alan bilgisi sınavı sebebiyle beden eğitimi ve spor öğretmen adaylarının yetenek/bilgi düzeylerini belirlemede geçerli olduğu görülmüştür. BUT uygulamasının bu çalışmada ortaya koyduğu en önemli sonuç sorulan soruları %80 oranında (ortalama uygulanan madde/tüm test-16/80) azaltmasıdır.

SONUÇ

Bu çalışmadan elde edilen sonuçlar, bilgisayar uyarlamalı test uygulaması ile daha kısa sürede, daha az madde sayısı ile güvenli sonuçlar elde edilebileceğini göstermektedir. Bunun yanı sıra BUT kullanımı ile test hazırlama süresinin ve madde işaretleme için gereken sürenin azalması, sınav sonuçlarının testin tamamlanmasından çok kısa bir süre sonra ilan edilebilmesi gibi kazanımlar sağlayabileceğine işaret etmektedir. İçinde bulunduğumuz pandemi krizinde uzaktan eğitimde kaçınılmaz hale gelen çevrimiçi sınavlarda kişi yetenek düzeyini daha hızlı ve tutarlı şekilde belirlemek ve kişilerin sınav stratejilerini çözümlmek için BUT kullanımının geliştirilmesi uygun olacaktır.

Teşekkür

Bu çalışmada başlıca yazarın tez merkezi 666606 nolu doktora tezindeki teorilerden yararlanılarak gerçek veri üzerinde uygulama gerçekleştirilmiştir. Yazarlar desteklerinden ötürü tez danışmanı Prof Dr. Atilla Halil ELHAN'a teşekkürlerini sunmaktadır.

KAYNAKLAR

1. Turgut MF. Eğitimde ölçme ve değerlendirme metotları (10. Baskı). Ankara: Gül Yayınevi. 1997

2. Reeve BB. An introduction to modern measurement theory. National Cancer Institute, 2002;1-67.
3. Weiss DJ, Kingsbury GG. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 1984;21(4):361-75.
4. Rudner LM. An online, interactive, computer adaptive testing tutorial. 11/98. Erişim Adresi: <http://edres.org/scripts/cat/catdemo.htm>
5. Rezaie M, Golshan M. Computer adaptive test (CAT): Advantages and limitations. *International Journal of Educational Investigations*. 2015;2(5):128-37.
6. Callear D, King T. Using computer-based tests for information science. *ALT-J*, 1997;5(1):27-32.
7. Madsen HS. Computer-adaptive testing of listening and reading comprehension. (In P. Dunkel Ed.), *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House. 1991;237-57.
8. Wainer H. *Computerized Adaptive Testing: A primer*. Hillsdale, N J: LEA. 1990.
9. Young R, Shermis MD, Brutton SR, Perkins K. From conventional to computer-adaptive testing of ESL reading comprehension. *System*. 1996;24(1):23-40.
10. R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>. 2020.
11. Cicchetti DV. "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology". *Psychological Assessment*. 1994;6(4): 284–90.
12. Gamer M, Lemon J, Gamer MM, Robinson A, Kendall's W. Package 'irr'. Various coefficients of interrater reliability and agreement. CRAN R Project. 2012.
13. Lehnert B. BlandAltmanLeh: plots (slightly extended) Bland-Altman plots. R package version 0.3.1. 2015.
14. Si CF, Schumacker RE. Ability estimation under different item parameterization and scoring models. *International Journal of Testing*. 2004;4(2):137-81.
15. Rasch G, *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press. 1960.
16. Lee YH, Chen H. A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*. 2011;53(3):359.
17. Bennett RE, Braswell J, Oranje A, Sandene B, Kaplan B, Yan F. Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*. 2008;6(9):1–39.
18. Threlfall J, Pool P, Homer M, Swinnerton B. Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*. 2007;66(3):335–48.
19. Wang S, Jiao H, Young M, Brooks T, Olson J. Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*. 2008;68:5–24.
20. Şenel S, Kutlu Ö. Comparison of two test methods for VIS: paper-pencil test and CAT. *European Journal of Special Needs Education*. 2018;33(5):631-45.
21. De Ayala RJ. *The theory and practice of item response theory*. Guilford Publications. 2013.