1. BİREYE UYARLANMIŞ TEST ARAŞTIRMALARI ULUSAL SEMPOZYUMU

14-15 EYLÜL 2023 | BOĞAZİÇİ ÜNİVERSİTESİ, İSTANBUL

DESTEKLEYENLER

ETS | GLOBAL A subsidiary of ETS

KANGURU MATEMATİK

DÜZENLEME KURULU

Doç. Dr. Serkan Arıkan	Boğaziçi Üniversitesi
Doç. Dr. Eren Can Aybek	Pamukkale Üniversitesi
Dr. Güneş Ertaş	Boğaziçi Üniversitesi
Prof. Dr. Emine Erktin	Boğaziçi Üniversitesi
Prof. Dr. R. Nükhet Çıkrıkçı	İstanbul Aydın Üniversitesi
Doç. Dr. Murat Doğan Şahin	Anadolu Üniversitesi
Doç. Dr. İlker Kalender	Bilkent Üniversitesi
Belgin Eriz	Boğaziçi Üniversitesi
Gamze Uçak Ersizer	Boğaziçi Üniversitesi
Ceren Berfin Özdemir	Boğaziçi Üniversitesi

GÖNÜLLÜ SEMPOZYUM GÖREVLİLERİ

Aybüke Doğaç	Hacettepe Üniversitesi
Sinem Coşkun	Hacettepe Üniversitesi
Leyla Burcu Dinçsoy	Hacettepe Üniversitesi
Yegâne Toksoy	Gazi Üniversitesi
Ahmet Güveli	Çankırı Karatekin Üniversitesi
Gizem Kılıç	Boğaziçi Üniversitesi
Ceren Çem	Boğaziçi Üniversitesi

BİLİM KURULU

HAKEM KURULU

Bu sempozyumda yer alması için kabul edilen bildirilere iki farklı hakem değerlendirmesi sonucunda karar verilmiştir

Prof. Dr. Ali Baykal	Bahçeşehir Üniversitesi
Prof. Dr. Hülya Kelecioğlu	Hacettepe Üniversitesi
Prof. Dr. Neşe Güler	İzmir Demokrasi Üniversitesi
Doç. Dr. Beyza Aksu Dünya	Bartın Üniversitesi
Doç. Dr. Burak Aydın	Ege Üniversitesi
Doç. Dr. Celal Deha Doğan	Ankara Üniversitesi
Doç. Dr. Eren Can Aybek	Pamukkale Üniversitesi
Doç. Dr. İlker Kalender	Bilkent Üniversitesi
Doç. Dr. Kübra Atalay Kabasakal	Hacettepe Üniversitesi
Doç. Dr. Murat Doğan Şahin	Anadolu Üniversitesi
Doç. Dr. Özge Altıntaş	Ankara Üniversitesi
Doç. Dr. Sedat Şen	Harran Üniversitesi
Doç. Dr. Seher Yalçın	Ankara Üniversitesi
Doç. Dr. Selma Şenel	Balıkesir Üniversitesi
Doç. Dr. Semirhan Gökçe	Niğde Ömer Halisdemir Üniversitesi
Doç. Dr. Serkan Arıkan	Boğaziçi Üniversitesi
Dr. Öğr. Üyesi Arzu Uçar	Hakkari Üniversitesi
Dr. Öğr. Üyesi Ebru Balta	Ağrı İbrahim Çeçen Üniversitesi
Dr. Güneş Ertaş	Boğaziçi Üniversitesi

ÇAĞRILI KONUŞMACILAR

Prof. Dr. Giray Berberoğlu — Başkent Üniversitesi

Bilgisayar Ortamında Bireye Uyarlanmış Ölçme Uygulamalarının Psikometrik ve Lojistik Sorunları

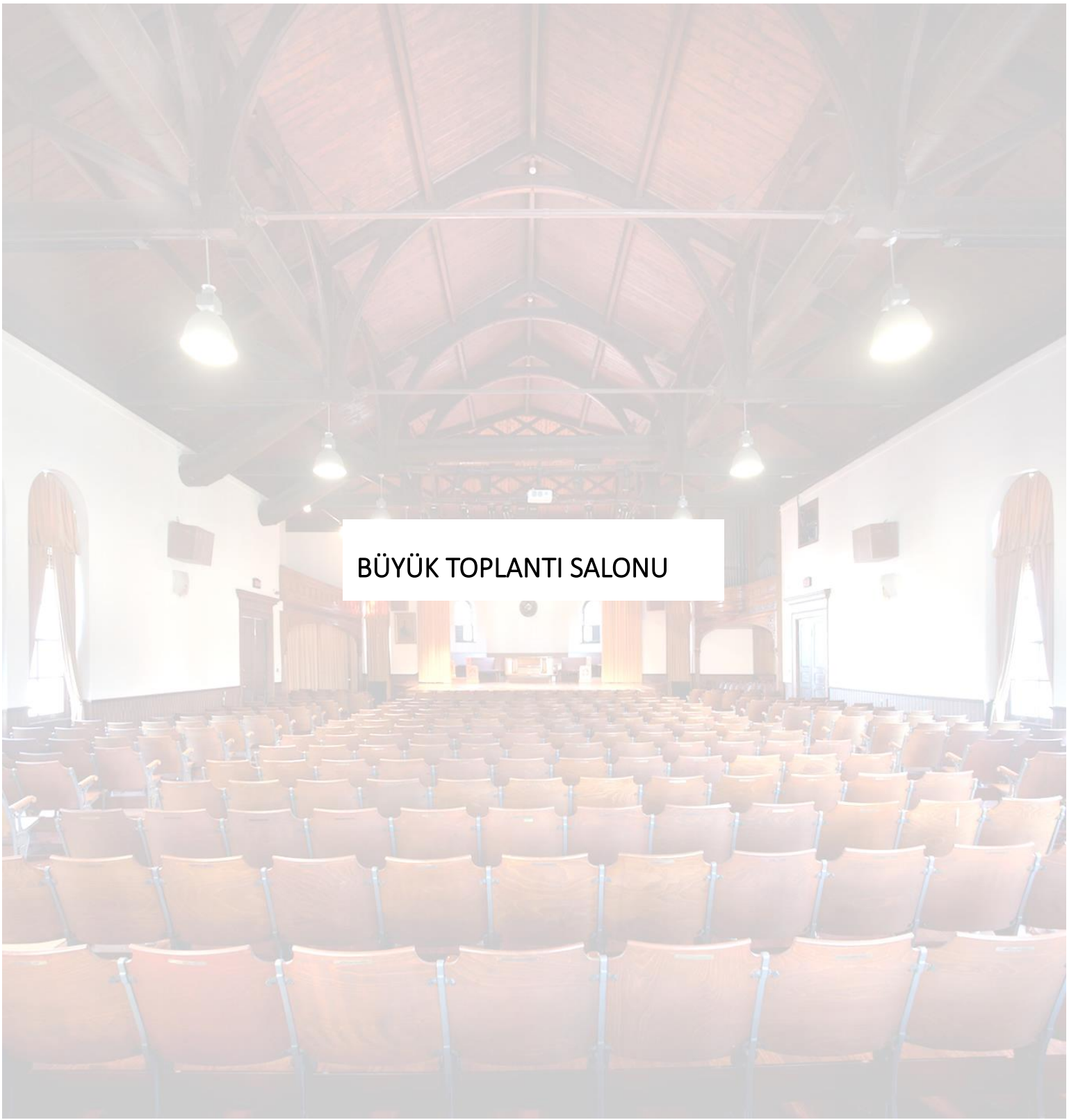
Prof. Dr. Kadriye Ercikan — ETS

Opportunities and Challenges in Adaptivity in Innovative Digital Assessments

Prof. Dr. R. Nükhet Çıkrıkçı — İstanbul Aydın Üniversitesi

Ulusal Eğitim Sorunlarımızın Çözümünde Bireye Uyarlanmış Testler Bir Seçenek Olabilir mi?

İçindekiler


BÜYÜK TOPLANTI SALONU

Bilgisayar Ortamında Bireye Uyarlanmış Testler ile İlgili Araştırmaların Eğilimi: 1980-2023

Kübra Ceren PINARLI[1] Cemile ŞAHİN[1] Semirhan GÖKÇE[1]

ÖZET

GİRİŞ

Özellikle son 40 yılda bilgisayar teknolojisinde yaşanan gelişmeler eğitimdeki ölçme-değerlendirme uygulamalarında bir çeşitliliğe ve zenginliğe imkân sağlamıştır. Buna durum eğitimde ölçme-değerlendirme uygulamalarında farklı arayışlara neden olmuş ve geleneksel kâğıt-kalem testlerinden bilgisayar ortamında uygulanan testlere yönelik değişim ve dönüşüm ile sonuçlanmıştır. Bu dönüşüm sadece ortamın farklılığıyla sınırlı kalmamıştır. Test uygulamalarının kâğıt-kalem formundan bilgisayar ortamında uygulanabilir bir yapıya kavuşturulmasında kuramsal ve uygulamalı araştırmaların önemli bir rolü bulunmaktadır (Sireci vd., 2008). Yürütülen araştırmalar bilgisayar ortamında lineer test uygulamalarından bireye uyarlanmış testlere ve sonrasında tasarlanan çok aşamalı bireye uyarlanmış testlere kadar birçok farklı spektrumda yer alan testlerin geliştirilmesinde başrol oynamıştır.

Kâğıt-kalem testleri ile karşılaştırıldığında bilgisayar ortamında bireye uyarlanmış testlerin birçok avantajı bulunmaktadır. Bilgisayar ortamında bireye uyarlanmış test sürecinde katılımcıların yanıtlayacağı sorular, diğer bir ifadeyle test formu, katılımcıların önceki yanıtlarına göre belirlenmektedir (Luecht ve Sireci, 2012; van der Linden, 2010). Madde tepki kuramının kullanıldığı bu süreçte yetenek kestiriminde ve madde parametrelerinin belirlenmesinde farklı yöntemler kullanılmaktadır. Buna bağlı olarak her katılımcı için yetenek düzeyine optimize edilmiş bir test formu uygulama sürecinde şekillenmektedir. Bilgisayar ortamında bireye uyarlanmış test uygulamaları daha az soru ile daha güvenilir sonuçlar elde edilmesini sağlamasının yanısıra testlerin uygulama süresini kısaltmaktadır (Eggen, 2007; Hambleton vd., 1991; Meijer ve Nering, 1999; Mills ve Stocking, 1996; Verschoor ve Straetmans, 2010). Bununla birlikte, testten elde edilen puanların uygulama bitiminde öğrenilmesi (Wainer, 2000) ve her soruda harcanan sürenin kaydedilmesi gibi faktörler de bilgisayar ortamında testlerin önemli katkıları olarak karşımıza çıkmaktadır. Bu nedenle, özellikle eğitim ve psikoloji alanlarında kullanılan testlerde, bilgisayar ortamında bireye uyarlanmış testler giderek daha yaygın hale gelmektedir(Özbaşı,2016).

Bu çalışmanın amacı, 1980-2023 yılları arasında Web of Science veritabanında yer alan bilgisayar ortamında bireye uyarlanmış testler ile ilişkili kavramları belirlemek, bu kavramların oluşturduğu yapıları incelemek ve yapılan araştırmaların eğilimini ortaya koymaktır. Bu doğrultuda analizlerde ortaya çıkan yapılar tanımlanmış ve ardından makalelerin içerikleri değerlendirilerek kavramlar arasındaki ilişkiler derinlemesine incelenmiştir.

YÖNTEM

Bu çalışmada, 1980-2023 yılları arasında bilgisayar ortamında bireye uygulanmış testleri ile ilgili yayınlanan makalelere odaklanılmıştır. Web of Science veritabanında yer alan bu makalelerin incelenmesinde bibliyometrik analiz yöntemi kullanılmıştır. Bibliyometrik analiz, araştırmaya geriye dönük bir bakış sağlamak ve belirli bir dönemdeki araştırma eğilimini ortaya çıkarmak için kullanılmaktadır (Krauskopf, 2018). Bu bağlamda büyük bilimsel verileri araştırmak ve analiz etmek, belirli bir kavramın gelişim süreçlerini ortaya koyarken diğer yandan o kavramla ilişkili diğer kavramları

[1] Niğde Ömer Halisdemir Üniversitesi

tanımlamak ve haritalamak mümkün hale gelmektedir (Donthu vd., 2021). Bibliyometrik analiz, eş-oluşum analizi, anahtar kelime analizi, kümeleme analizi ve bibliyometrik haritalar gibi teknikleri kullanarak alanyazındaki çalışmaları sistematik olarak incelemek için önemli bir yaklaşım olarak görülmektedir (Song ve Wang. 2020). Geriye dönük bir inceleme ile araştırma çalışmalarının zaman içindeki eğilimler hakkında fikir edinmeye olanak tanımaktadır.

Çalışmanın (1) arama ve filtreleme, (2) analiz etme ve görselleştirme (3) kümeleme ve tanımlama (4) kanıt arama ve doğrulama olmak üzere dört aşaması bulunmaktadır. Web of Science veritabanında yürütülen makale arama ve filtreleme aşamasında PRISMA akış şeması ile süreç özetlenmiştir. Analiz etme ve görselleştirme süreci VOSviewer yazılımı ile gerçekleştirilmiştir. VOSviewer bibliyometrik ağların oluşturulması ve görselleştirilmesi için geliştirilen bir yazılım aracıdır (Van Eck & Waltman, 2020). Kümeleme ve tanımlama aşaması VOSviewer tarafından oluşturulan yapılarda yer alan kavramlar arası ilişkiler dikkate alınarak alan uzmanları tarafından yürütülmüştür. Tanımlanan yapılar bilgisayar ortamında bireye uyarlanmış testlerin boyutları konusunda bir fikir vermektedir. Son olarak kanıt arama ve doğrulama aşamasında ise çalışmada yer alan makaleler incelenmiş ve tanımlanan yapıların geçerliği konusunda bir doğrulama süreci yürütülmüştür.

BULGULAR/BEKLENEN BULGULAR

Çalışma kapsamında bilgisayar ortamında bireye uyarlanmış testler ile ilgili yayınlanan makale sayılarının yıllar içerisindeki değişimi, makalelerin ülke bazındaki dağılımı, makalelerde yer alan anahtar sözcüklerin analizi, anahtar sözcüklerin onar yıllık periyotlar halindeki eğilimleri paylaşılacaktır. Bunun yanı sıra bilgisayar ortamında bireye uyarlanmış testler ile ilgili kavramlar ve kavramlararası ilişkilerin yer aldığı ağ ve yoğunluk haritaları ile ilgili yorumlara yer verilecektir. Çalışmada ayrıca tanımlanan yapılarda sık tekrar eden kavramlar ile ayrıntılı bilgiler raporlanacaktır.

Çalışmanın ön bulguları bilgisayar ortamında bireye uyarlanmış test uygulamaları ile ilgili makalelerin temel anlamda teorik ve uygulamaya dönük araştırmalar olmak üzere iki boyutlu yapısına dikkat çekmektedir. Söz konusu çalışmanın gelecekte yürütülecek bilgisayar ortamında bireye uyarlanmış testler ile ilgili çalışmamaların şekillendirilmesine katkı sağlayacağı düşünülmektedir.

KAYNAKÇA

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Gökçe, S., & Güner, P. (2021). Forty Years of Mathematics Education: 1980-2019. International Journal of Education in Mathematics, Science and Technology, 9(3), 514-539.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory (Vol. 2). Sage.

Krauskopf, E. (2018). A bibiliometric analysis of the Journal of Infection and Public Health: 2008–2016. Journal of Infection and Public Health, 11(2), 224-229. https://doi.org/10.1016/j.jiph.2017.12.011

Luecht, R. M. & Sireci, S. G. (2012). A review of models for computer-based testing. Research Report RR-2011-12. New York: The College Board.

Meijer, R. R. & Nering M. L. (1999). Computerized adaptive testing: overview and introduction. Applied Psychological Measurement, 23, 187-194.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. Applied Measurement in Education, 9 (4), 287-304.

Özbaşı, D. (2016). Bilgisayar ortamında bireye uyarlanmış test uygulamasına ve kâğıt kalem testine katılan öğrenci görüşlerinin incelenmesi. Uluslararası Sosyal Araştırmalar Dergisi, 9(42), 1270-1274.

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., & Hambleton, R. K. (2008). Massachusetts Adult Proficiency Tests Technical Manual, Version 2. Center for Educational Assessment Research Report No, 677.

Song, P., & Wang, X. (2020). A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years.Asia Pacific Education Review, 21(3), 473–486. https://doi.org/10.1007/s12564-020-09640-2

Van Eck, N. J., & Waltman, L. (2020). Manuscript for VOSviewer version 1.6.15. Leiden: Univeristeit Leiden, 1(1), 1-52.

Van der Linden, W. J. (2010). Item selection and ability estimation in adaptive testing. Elements of Adaptive Testing, 3-30. Springer.

Verschoor, A. J., & Straetmans, G. J. J. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of Adaptive Testing, (pp. 137-149). Statistics for Social and Behavioral Sciences. Springer.

Wainer, H. (2000). Computerized Adaptive Testing: A Primer. Mahvah, NJ: Erlbaum.

Anahtar Kelimeler: bilgisayar ortamında bireye uyarlanmış test, bibliyometrik analiz

TIMSS 2019 Sekizinci Sınıf Matematik Başarı Testlerinin Bireye Uyarlanmış Test Olarak Uygulanabilirliğinin İncelenmesi

Mehmet Furkan GİTMİŞ[1] Semirhan GÖKÇE[1]

ÖZET

GİRİŞ

1980'li yıllardan itibaren teknolojinin gelişimi ve bilgisayar kullanımının yaygınlaşması eğitimde ölçme ve değerlendirme uygulamalarını da derinden etkilemiştir. Günümüzde her ne kadar geleneksel kâğıt kalem testleri uygulama kolaylığından ötürü sınıf içi ölçme değerlendirme süreçlerinin hala vazgeçilmez aracı olsa da geniş ölçekli uygulamalarda bilgisayar ortamındaki testlere ilgi artmış ve kullanımı yaygınlaşmaya başlamıştır. Geçmişte sadece istatistiksel analizler için faydalanılan bilgisayarlar 1990'lı yıllardan itibaren bir "ortam" olarak görülmüş buna bağlı olarak da testlerin yapılması, uygulanması, puanlanması ve raporlanmasında kullanılmaya başlanmıştır (Zenisky & Sireci, 2002).

Uluslararası geniş ölçekli ölçme ve değerlendirme uygulamalarında kâğıt-kalem ortamından bilgisayar ortamında uygulanan testlere doğru bir geçiş göze çarpmaktadır. Örneğin, TIMSS 2019 uygulamaları katılımcıların tercihine bağlı olarak kâğıt-kalem ya da bilgisayar ortamında uygulanmaya başlamıştır. Bilgisayar ortamında gerçekleşen test uygulamalarında lineer testler kullanılmıştır. Uygulamalarda ortak maddeler içeren 14 farklı kitapçıkta yöneltilen sorularda eksik test tasarımı kullanılmıştır.

Bu çalışmanın amacı, TIMSS 2019 sekizinci sınıf matematik başarı testlerinin bireye uyarlanmış test olarak uygulanabilirliğini incelemektir. Testlerde kullanılan çoktan seçmeli maddelerin yer aldığı simülasyon çalışmalarında en uygun bireye uyarlanmış test algoritmasının belirlenmesi amaçlanmaktadır.

YÖNTEM

Bu çalışmada 2019 TIMSS 8. Sınıf verilerinden çoktan seçmeli olanlarına odaklanılmıştır. TIMSS 2019 International Database veri tabanında SPSS dosyası indirilerek verilere ulaşılmıştır. 2019 TIMSS sınavlarına elektronik ortamda giren ülkeler ve kağıt kalem ortamında giren ülkeler ayrı ayrı ele alınmıştır. Elde edilen verilen SPSS programı ile düzenlenmiş ve analiz yapmaya hazır hale getirilmiştir. Verilerin aritmetik ortalaması ve standart sapmaları hesaplanmıştır. Elde edilen sonuçlar kullanılarak SimulCAT yazılımı üzerinden simülasyon çalışması yapılmıştır.

Simülasyon çalışmaları "… olsa ne olurdu?" sorusuna cevap arayan çalışmalardır. Simülasyon çalışmaları araştırmacıların örgütsel sistemlerin doğasında var olan karmaşıklığı verili olarak varsaymasına izin verir. Simülasyon, geleceğe "ileriye" giderek gözlemler yarattığı için daha karmaşık sistemlerin çalışılmasına olanak sağlarken, diğer araştırma yöntemleri ne olduğunu ve nasıl olduğunu belirlemek için tarihte geriye doğru bakmaya çalışır. (Dooley, 2002).

Çalışmada farklı Test başlama kuralları, Soru seçim yöntemleri, Yetenek Kestirim Yöntemleri ve Test Sonlandırma Kuralları kullanılarak simülasyonlar oluşturulmuştur.

Madde seçim kriteri (Item Selection Criterion) olarak; Maximum Fisher Information (MFI) ve Likelihood Weightted Information (LWI) olmak üzere 2 kriter belirlenmiştir. Item Exposure Control olarak; No Exposure Control ve Fade Away Method (FAM) olmak üzere 2 kriter belirlenmiştir. Test

[1] Niğde Ömer Halisdemir Üniversitesi

uzunluğu (Test Length) kriteri olarak; Fixed Length 10 madde, Fixed Length 15 madde, Veriable Length 0.3 ve Veriable Length 0.4 olmak üzere 4 kriter belirlenmiştir. Yetenek kestirim kriteri (Score Estimation) olarak ise Maximum Likelihood Estimation (MLE) ve Bayes Expected a Posteriori (EAP) olmak üzere 2 kriter belirlenmiştir. Toplamda 2*2*4*2= 32 farklı kombinasyonun ayrı ayrı elektronik ortamda yapılan TIMSS ve Kağıt Kalem ortamında yapılan TIMSS sınavı verileri kullanılarak simülasyonları yapılacaktır.

Simülasyon çalışmalarından elde edilen bulgular tek tek 2019 TIMSS 8. Sınıf elektronik ortamda yapılan ve 2019 TIMSS 8. Sınıf kağıt kalem ortamında yapılan sınav verilerinden elde edilen bulgularla SPSS programında korelasyon analizi yapılmıştır. Yapılan analizler sonucunda hangi yöntemin 2019 TIMSS 8. Verileriyle daha benzer sonuçlar elde edilebileceği ayrı ayrı belirlenmiştir.

BULGULAR/BEKLENEN BULGULAR

Çalışma kapsamında TIMSS sınavlarının Bireye Uyarlanmış Test (Adaptive Test) Formatında uyarlanabilirliği ve hangi test başlama kuralının, hangi soru seçim yönteminin, hangi yetenek kestirim yönteminin ve hangi test bitirme kuralının daha uygun olacağı paylaşılacaktır. Bunun yanı sıra simülasyon çalışmasıyla farklı yöntemlerle elde edilen sonuçlar ile TIMSS 2019 verilerinden elde edilen sonuçlar yorumlarıyla birlikte çalışma kapsamında sunulacaktır.

Çalışmanın ön bulguları 2019 TIMSS 8. Sınıf elektronik ortamda yapılan ve kâğıt kalem ortamında yapılan sınav analizleri ile simülasyonda elde edilen sonuçlar arasında %90 üzerinde korelasyon ilişkisi bulunmaktadır. En yüksek korelasyon ilişkisi bulunan yöntem yorumlarıyla birlikte yorumlarıyla birlikte çalışma kapsamında sunulacaktır.

Elde edilen bulgular; TIMSS sınavlarının sadece eksik test tasarımı yöntemiyle değil bireye uyarlanmış test formunda da yapılabileceğini göstermektedir.

KAYNAKÇA

Dooley, K. (2002), “Simulation research methods,” Companion to Organizations, Joel Baum (ed.), London: Blackwell, pp. 829-848.

Zenisky A. L., & Sireci, S. G. (2002) Technological innovations in large-scale assessment, Applied Measurement in Education, 15:4, 337-362.

Anahtar Kelimeler: bireye uyarlanmış test, matematik başarı testi, simülasyon çalışması, TIMSS

TIMSS 2019 Dördüncü Sınıf Matematik Uygulaması İçin En Uygun Bilgisayar Ortamında Bireye Uyarlanmış Test Algoritmasının Belirlenmesi

Ece Elif CEVİZ[1] Zehra Nur APAYDIN[1] Semirhan GÖKÇE[1]

ÖZET

GİRİŞ

Ölçme-değerlendirme alanında, geçerli ve güvenilir test geliştirme süreci için Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) kullanılmaktadır. KTK, madde ve test istatistiklerinden elde edilen örnekleme bağlı kalınarak bireylerin yeteneklerini madde güçlük düzeyi ve madde ayırt ediciliği ile kestirmeye çalışmaktır (Hambleton ve Swaminathan, 1985). MTK'de, KTK'den farklı olarak test başarı puanı yerine her bir maddeye verilen cevabın doğru ya da yanlışlığı ile ilgilenilmektedir. MTK, bireyin yetenek düzeyinin belirlenmesinde ve matematiksel olarak ifade edilmesinde test maddelerinin verimliliği bakımından Bilgisayar Ortamında Bireye Uyarlanmış Test (BBT)'in geliştirilmesine imkân sunmaktadır (Hambleton vd., 1991). BBT, geleneksel ölçme-değerlendirme uygulamalarından (kâğıt-kalem ortamından) farklı olarak bilgisayar ortamında, bireyin karşısına yetenek düzeyine en uygun soruları çıkararak daha kısa sürede sonuç almayı sağlamaktadır (Segall, 1996).

İlk uygulaması 1995 yılında yapılan ve sonrasında da kâğıt-kalem ortamında gerçekleştirilmeye devam eden Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS), katılımcı ülkelere 2019 yılında ilk kez hem kâğıt-kalem hem de bilgisayar ortamında uygulama fırsatı sunmuştur (Milli Eğitim Bakanlığı [MEB], 2020). TIMSS 2019 uygulamasına 64 ülke, dördüncü sınıf, sekizinci sınıf veya her iki sınıf düzeyinde katılım sağlamıştır (Mullis vd., 2020). TIMSS uygulamalarında ortak maddeler içeren 14 farklı kitapçık yer almakta ve yöneltilen sorularda eksik test tasarımı kullanılmaktadır. Bu çalışmanın amacı, TIMSS 2019 verilerine göre dördüncü sınıf matematik başarı testlerinin simülasyon çalışmaları ile en uygun bilgisayar ortamında bireye uyarlanmış test algoritmasının belirlenmesidir.

YÖNTEM

Bu bölümde TIMSS 2019 uygulamasının 4. sınıf matematik başarı testlerinde yer alan çoktan seçmeli sorular kullanılarak yürütülen madde tepki kuramı analizleri ve en uygun bireye uyarlanmış test uygulama algoritmasının belirlenmesine yönelik gerçekleştirilen simülasyon çalışmalarına yönelik bilgiler yer almaktadır. Öncelikle eksik test tasarımıyla hazırlanmış ve 14 farklı kitapçıktan oluşan TIMSS 2019 uygulamasının matematik veri setinde sadece elektronik ortamda uygulanan soruların yanıtları dikkate alınmış ve bunlar üzerinde gerekli düzenlemeler yapılmıştır. Çalışmada sadece çoktan seçmeli matematik sorularına verilen yanıtlar kullanılmıştır. Kitapçıklarda yer alan ortak sorular yardımıyla 2 parametreli lojistik modele uygun madde analizleri gerçekleştirilmiştir. Elde edilen madde güçlük (b) ve madde ayırtedicilik (a) parametreleri kullanılarak SimulCAT programında farklı simülasyonlar test edilmiştir. Bireye uyarlanmış testler 4 temel adımdan oluşmaktadır: (1) test başlama kuralının belirlenmesi, (2) madde seçim yönteminin belirlenmesi, (3) yetenek kestirim yönteminin belirlenmesi ve (4) test sonlandırma kriterinin belirlenmesi. Gerçekleştirilen simülasyonlarda 2 farklı yetenek kestirim yöntemi (maksimum olabilirlik ve Bayes beklenen sonsal dağılım), 2 farklı soru seçim yöntemi (Fisher'in maksimum bilgi ve olabilirlik kapsamında ağırlıklandırılmış bilgi) ve 4 farklı test sonlandırma kriteri (iki farklı sabit test uzunluğu ve iki farklı değişken test uzunluğu) yer almıştır. Bunlarla birlikte madde kullanım sıklığının etkilerini belirlemeye yönelik Sympson-Hetter ve Fade Away yöntemlerinin etkililiği araştırılmıştır (Gökçe and Glas, 2018).

[1] Niğde Ömer Halisdemir Üniversitesi

BULGULAR/BEKLENEN BULGULAR

Çalışmanın ön bulguları yetenek kestirim yöntemlerinin birbirlerine yakın sonuçlar verdiğini, madde seçim yöntemi olarak Fisher'in maksimum bilgi yönteminin etkili olduğunu ve madde kullanım sıklığı kontrolünün en uygun algoritmada yer alması gerektiğini belirtmektedir. Araştırma bulgularının elektronik ortamda gerçekleştirilen TIMSS uygulamalarının bilgisayar ortamında bireye uyarlanmış olarak uygulanabilirliği konusunda katkı sağlayacağı düşünülmektedir.

KAYNAKLAR

Gökçe, S., and Glas, C. A. W. (2018). Can TIMSS mathematics assessments be implemented as computerized adaptive test?. Journal of Measurement and Evaluation in Education and Psychology, 9(4), 422-435. DOI: 10.21031/epod.487351

Hambleton, R. K. and Swaminathan, H. (1985). Some background to item response theory. Item response theory in (s. 1-14). Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). Fundamentals if item response theory. Journal of Educational Measurement, 30(1), 84-87.

Milli Eğitim Bakanlığı (2020). TIMSS 2019 Türkiye ön raporu. Ankara: Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D., and Fishbein, B. (2020). TIMSS 2019 international results in mathematics and science. Boston College, TIMSS & PIRLS International Study Center.

Segall, D. O. (1996). Multidimensional adaptive testing. Psychometrika, 61(2), 331-354.

Anahtar Kelimeler: Bilgisayarla bireye uyarlanmış test, simülasyon çalışması, matematik değerlendirme, TIMSS

E-TIMSS 2019 Matematik Uygulaması Çok Aşamalı Bireye Uyarlanmış Test Olarak Uygulanabilir mi?

Rumeysa Uslu[1] Semirhan Gökçe[2]

ÖZET

GİRİŞ

Günümüzde önemli kararlar alınmasında kullanılan geniş ölçekli test uygulamaları kâğıt-kalem olarak uygulanmaktadır. Kâğıt-kalem ortamında uygulanan testlerde katılımcıların yanıtlaması beklenen sorular uygulama öncesinde testi geliştiren psikometri uzmanları tarafından bilinmektedir. Uzmanlar, farklı yetenek düzeyinden katılımcıların yetenek düzeylerini daha güvenilir ölçebilmek adına çok kolaydan çok zora farklı güçlük düzeylerinde çok sayıda soruya testte yer vermektedir. Gelişen bilgisayar teknolojisi ve Madde Tepki Kuramı'nın (MTK) ölçme ve değerlendirme sürecine katkıları sonucunda her katılımcının yetenek düzeyine uygun soru sorulması ve katılımcının doğru ya da yanlış yanıtına göre sonraki sorunun belirlenmesi mümkün hale gelmiştir. Kısacası, her katılımcının testinde yer alan soruları katılımcının kendisi belirlemektedir. Test uzunluğunu azaltması ve buna bağlı olarak da uygulama süresini kısaltması gibi temel avantaja sahip uyarlanabilir testler sayesinde daha tutarlı ve hassas ölçümlerin yapılabilmesi sağlanmıştır (Wainer, 1990). Alanyazında uyarlanabilir testler için iki durumun söz konusu olduğu belirtilmektedir: (1) madde düzeyinde uyarlanabilir testler ve (2) çok aşamalı uyarlanabilir testler. Madde düzeyinde uyarlanabilir testlerde yetenek kestirimi her maddeye verilen yanıt sonrasında gerçekleşirken, çok aşamalı uyarlanabilir testlerde yetenek kestirimi bir grup maddeye verilen yanıt sonrasında gerçekleşmektedir (Yamamoto, Shin & Khorramdel, 2018). çok aşamalı uyarlanabilir testlerin pratiklik, ölçme doğruluğu ve test formları üzerinde kontrol fırsatı yarattığı ifade edilmektedir (Zenisky, Hambleton ve Luecht, 2010). Çok aşamalı uyarlanabilir testlerin katılımcılar için test süresini artırmadan bireysel ve grup düzeylerinde ölçme hatasını azalttığı ifade edilmektedir (Oranje vd., 2014).

Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) 1995 yılından itibaren dörder yıllık periyotlarda gerçekleştirilen ve matematik ve fen bilimleri alanlarında kazandıkları bilgi ve becerilerin değerlendirilmesini amaçlayan bir tarama araştırmasıdır (MEB, 2020). 2019 yılından itibaren tercihe bağlı olarak bilgisayar ortamında da uygulanmaya başlayan ve eTIMSS olarak yeniden isimlendirilen uygulamalarda 14 farklı kitapçıkta sorular yöneltilmektedir. Bu çalışmanın amacı bilgisayar ortamında uygulanan eTIMSS 2019 uygulamasında yer alan matematik başarı testlerinin çok aşamalı bireye uyarlanmış test olarak uygulanabilirliğini araştırmaktır.

YÖNTEM

Çalışma kapsamında eTIMSS 2019 sekizinci sınıf matematik başarı testlerinde kullanılan ve ikili puanlanan (0 ve 1 olarak) tüm çoktan seçmeli ve açık uçlu sorulara verilen yanıtlar analiz edilecektir. TIMSS uygulamaları kapsamında yürütülen madde kalibrasyon süreci sonunda madde güçlüğü (b parametresi) ve madde ayırtediciliği (a parametresi) olmak üzere iki farklı değer rapor edilmektedir (Fishbein, Foy & Yin, 2021). Bu nedenle 14 farklı kitapçıkta yer alan ve eksik test tasarımının yer aldığı eTIMSS sekizinci sınıf matematik verilerinin madde kalibrasyonunda 2 parametreli lojistik model kullanılacaktır. Madde kalibrasyonları sonucunda her madde için a ve b parametrelerinin yanı sıra her katılımcı için yetenek kestirimleri belirlenecektir.

[1] Kayseri Erkilet General Emir Ortaokulu
[2] Niğde Ömer Halisdemir Üniversitesi

En uygun çok aşamalı uyarlanabilir test algoritmasının belirlenmesine yönelik gerçekleştirilecek simülasyonlarda Han (2013) tarafından geliştirilen MSTGen programı kullanılacaktır. Program kapsamında katılımcıların önceki yanıtlarına göre önceden birleştirilmiş birkaç test modülünden birine yönlendirildiği tipik, geleneksel çok aşamalı uyarlanabilir testlerin yanı sıra test bilgisi fonksiyon hedeflerine dayalı olarak her aşama için anında bir öğe modülü şekillendiren yeni çok aşamalı uyarlanabilir test yaklaşımı desteklenmektedir.

Çalışma kapsamında üç farklı modül seçim yönteminin (Fisher'in maksimum bilgi, minimum ortalama güçlük farkı ve rastgele seçim) ve üç farklı yetenek kestirim yönteminin (MLE, EAP ve MAP) etkililiği araştırılacaktır. Simülasyonlardan elde edilen yetenek kestirimleri eTIMSS madde kalibrasyonları sonucunda elde edilen yetenek kestirimleri ile karşılaştırılacaktır. Bu kapsamda korelasyon değerleri rapor edilecek ve her iki yetenek kestirimleri grafik olarak paylaşılacaktır.

BULGULAR/BEKLENEN BULGULAR

Araştırma sonucunda bilgisayar ortamında uygulanan eTIMSS 2019 uygulamasında yer alan matematik başarı testlerinin çok aşamalı bireye uyarlanmış test olarak uygulanması durumunda hangi modül seçim yöntemi ile yetenek kestirim yönteminin daha uygun sonuçlar vereceği belirlenebilecektir. Çalışma kapsamında tasarlanan farklı modüller ile içerik kontrolünün de sağlanabileceği ifade edilebilir. 2019 yılına kadar sadece kâğıt-kalem ortamında uygulanan TIMSS uygulamalarının bu yıldan sonra tercihe bağlı olarak elektronik ortamda uygulanmaya başlaması gelecek yıllarda bu uygulamaların çok aşamalı bireye uyarlanabilir test olarak uygulanabileceği fikrini akıllara getirmektedir. Bu çalışmadan elde edilen bulguların gelecekte yürütülecek geniş ölçekli bu tarz bir uygulamanın en uygun test algoritmasının oluşturulmasında bir katkı sağlayabileceği düşünülmektedir.

KAYNAKLAR

Fishbein, B., Foy, P., & Yin, L. (2021). TIMSS 2019 User Guide for the International Database (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-database/

Han, K. T. (2013). MSTGen: simulated data generator for multistage testing. Applied Psychological Measurement, 37(8), 666-668.

Milli Eğitim Bakanlığı [MEB] (2020). TIMSS 2019 Türkiye Raporu, Eğitim Analiz ve Değerlendirme Raporları Serisi No: 15.

Oranje, A., Mazzeo, J., Xu, X., & Kulick, E. (2014). A multistage testing approach to group-score assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized multistage testing: Theory and applications (pp. 371-390). Boca Raton, FL: Chapman and Hall/CRC.

Wainer, H. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. Educational Measurement: Issues and Practice, 37(4), 16-27.

Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of adaptive testing (pp. 355–372). New York, NY: Springer.

Anahtar Kelimeler: çok aşamalı bireye uyarlanmış test, eTIMSS, madde tepki kuramı, simülasyon, yetenek kestirim yöntemi

Bilgisayar Ortamında Bireyselleştirilmiş Testlerde Sınıflama Geçerliği

İlker Kalender[1]

ÖZET

Giriş

Bilgisayar Ortamında Bireyselleştirilmiş Testler (Computerized Adaptive Testing – CAT) fikirsel olarak 1900'lerin başında bilgisayar olmaksızın "bireyselleştirilmiş test" olarak ortaya çıkmıştır. "Bilgisayar ortamında" teriminin bu test yaklaşımına entegre edilmesi 1900'lü yılların ortasını bulmuştur. Bilgisayar teknolojisinin gelişmesi, ucuzlaması ve bunlara paralel olarak yaygınlaşması sonunda CAT uygulamaları yaygınlaşmıştır. Genel olarak, CAT uygulamalarının temel işlevi daha az soru ile geçerlik ve güvenirlik kaybı olmadan bireylere bir puan atanmasını içerecek şekilde yapılandırılmıştır.

Ölçme değerlendirmenin belki de en temel amacının sınıflama olduğu düşünüldüğünde (Rudner ve Guo, 2011), bireylere puan vermenin dışında bireyleri ikili ya da çoklu kategorilere göre sınıflana konusunda da CAT araştırmalarının önemi büyüktür.

CAT kullanarak sınıflama konusunda Computerized Adaptive Classification Test (CACT) başlığı altında özellikle sınıflanma amacı ile tasarlanmış CAT uygulamaları ve buna paralel olarak bir literatür hali hazırda mevcuttur ve genişlemeye devam etmektedir. Thompson (2007) bu CACT formatı hakkında önemli bir kaynak olarak görülmektedir. Fakat bu uygulamalar genellikle sabit olmayan soru sayısı ile çalışacak şekilde tasarlanmışlardır. Buna karşın, uygulamada sabit soru sayısının kullanılmasını gerektiren sınıflama amaçlı CAT uygulamalarına ihtiyaç duyulabilir. Özellikle, test sonlandırma kuralının sabit ya da sabit olmayan soru sayısı olarak tanımlanmasının CAT tasarlanması esnasında kurumlara bir özgürlük sağlayabilecek olması bakımından önemi büyüktür.

Bu çalışmada özellikle sınıflama amacı ile tasarlanmamış CAT uygulamalarının sınıflama performanslarının incelenmesi konusu ele alınmıştır. Kullanılabilecek sınıflama performans göstergeleri, farklı CAT yaklaşımları, vs. boyutlar sunulacaktır. Gerçek verilere dayalı bir CAT uygulaması ile sınıflama performansının çeşitli göstergelerden nasıl etkilendiği gösterilecektir.

Yöntem

Bu çalışmanın veri kümesi bir üniversitenin İngilizce hazırlık okulunda kullanılan dil yeterliliği testinden elde edilmiştir. CAT uygulamalarının sınıflama geçerliğini bu dil test bağlamında ele alınmıştır. Bu yeterlik testinin sonucuna göre öğrencilerin bölümlerine başlayıp başlamayacakları ve bölüme başladıkları zaman ek İngilizce dersleri alıp almayacakları konusunda karar verilmektedir. Bu bakımdan bu testin öğrencileri farklı dil sınıflarına yerleştirme amacı ile de kullanıldığı düşünüldüğünde, sınıflama analizi için uygun bir veri olduğu düşünülmektedir.

Bireylerin bu dil testinin kağıt ve kalem formatına verdikleri yanıtlar kullanılarak farklı CAT kombinasyonları oluşturulmuştur. Farklı başlangıç kuralları, farklı sınıflama noktaları ve yaklaşımları, farklı test sonlandırma kuralları gerçek veriye dayalı simülasyonlar ile CAT formatında ele alınmıştır.

Veriler IRT modelleri ile kalibre edilmiş ve en iyi uyumun 2 parametreli model ile olduğu görülmüştür. Bu modele göre elde edilen madde ve yetenek kestirimleri kullanılarak farklı kombinasyonlar ile CAT simülasyonları tanımlanmıştır. Her bir simülasyon sonucunda elde edilen bireylerin yetenek kestirimleri sınıflama analizlerinde kullanılmıştır.

[1] İhsan Doğramacı Bilkent Üniversitesi

Sınıflanma performansı için farklı göstergeler ile incelenmiş ve sonuçlar kağıt ve kalem testinden gelen sonuçları ile karşılaştırılmıştır. Sınıflama göstergesi olarak CA (Classification Accuracy) ve CC (Classification Consistency) kullanılmıştır (Lee, 2010). CA indeksi bireylerin sınıflandırılması gereken "gerçek" kategori ile test soncuna göre yerleştirildiği kategoriler arasındaki uyumun bir göstergesi iken CC indeksi ise farklı test uygulamaları tarafından bireylerin ne kadarının hep aynı kategoriye yerleştikleri konusunda bilgi sağlayan bir göstergedir.

Bulgular/Beklenen Bulgular

Çalışma sonuçları sabit olmayan soru sayı CAT uygulamalarının da sabit sorulu CAT'ler kadar iyi sınıflama yaptığını ortaya koymuştur. Bu bakımdan CAT tasarımları yapılırken sınıflama amacı için sabit olmayan soru sayılı test sonlandırma kuralının da kullanılabileceği söylenebilir. Ayrıca, CAT kombinasyonları hiçbiri kağıt kalem testinin sınıflama geçerliği düzeyine ulaşamamıştır. Elde edilen sonuçlar CAT ve kağıt kalem testleri üzerinden karşılaştırmalı olarak tartışılacaktır.

Özellikle ülkemizdeki CAT araştırmalarının bireylere puan atanmasının ötesine geçirecek çalışmalara doğrıu kayması hem yerli literatürün zenginleşmesi hem de araştırma alanı olarak Dünya ile entegresyon açısından büyük önem taşımaktadır. Başarı testleri dışında, duyuşsal becerilerin ölçülmesi gibi konularda da CAT araştırmalarına ihtiyaç duyulmaktadır. Bu çalışmanın CAT literatürüne yapacağı katkı yanında yukarıda anılan boyutlarda da katkı yapması umulmaktadır.

Kaynakça

Lee, W. (2010). Classification consitency and accuracy for complex assessments using item response theory. Journal of Educatinoal Measurement, 47, 1–17.

Rudner, L. M., & Guo, F. (2011). Computer Adaptive Testing for Small Scale Programs and Instructional Systems. GMAC Research Reports, RR-11-01. Retrievable at http://www.gmac.com/NR/rdonlyres/91495C8D-0276-41FA-93BE-6595979B8D24/0/RR1101 CATforSmallScalePrograms.pdf

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. Practical Assessment Research and Evaluation, 12. Available online: http://pareonline.net/getvn.asp?v=12&n=1

Anahtar Kelimeler: CAT, sınıflama geçerliği, test sonlandırma

Değişen Madde Fonksiyonunun Farklı Koşullarda Bilgisayar Ortamında Bireye Uyarlanmış Testler Üzerindeki Etkisi

Merve ŞAHİN KÜRŞAD[1] Seher YALÇIN[2]

ÖZET

Giriş

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) bağlamında değişen madde fonksiyonu (DMF) tespiti ile ilgili yapılan araştırmalarda dikkate alınması gereken konulardan biri, ön uygulama ve madde bankasındaki maddelerin geliştirilmesi sırasında DMF'nin nasıl tespit edileceğiyle ilgilidir. Ayrıca gerçek BOBUT uygulaması sırasında, meydana gelebilecek DMF'nin yetenek kestirimi üzerine etkisi diğer önemli bir konudur. DMF'li maddenin test uygulama süresi boyunca hangi aşamada ortaya çıktığı da önemli diğer bir faktördür. DMF'li madde ile test uygulama sürecinin başlangıcında karşılaşan bir bireyin yetenek kestirimi sonraki sorularla dengelenebilir ancak bireyin aldığı son madde DMF'li ise bu durumun yetenek kestirimi üzerindeki olumsuz etkisi düzeltilememektedir. Ayrıca DMF, BOBUT'un erken aşamalarında meydana gelse bile, BOBUT'un DMF'in etkisine etkili bir şekilde uyum sağlaması garanti edilememektedir (Piromsombat, 2014). Ek olarak kağıt-kalem uygulamaları ile kıyaslandığında, BOBUT'da daha az madde uygulandığından DMF'li maddeler, daha ciddi sonuçlar doğurabilir (Lei, Chen, & Yu, 2006; Zwick, 2000). Bu bağlamda, bu çalışmada, DMF gösteren maddelerin bulunduğu testlerde bireyselleştirilmiş bilgisayarlı test yöntemlerinin performansının farklı koşullar altında incelenmesi amaçlanmıştır.

Yöntem

Bu çalışmada, DMF gösteren maddelerin bulunduğu testlerde bireyselleştirilmiş bilgisayarlı test yöntemlerinin performansının farklı koşullar altında incelenmesi amaçlanmıştır. Bu amaçla, simülasyon yöntemi ile yetenek parametreleri dağılımı -3 ile +3 aralığında değişen normal dağılıma sahip 1500 referans, 1500 odak grup üyesi olmak üzere 3000 kişi ve 500 maddelik bir havuz oluşturulmuştur. Madde parametreleri Rasch modele göre b parametresi -3 ile +3 aralığında normal dağılım olacak şekilde ayarlanmıştır. DMF'li madde yaratmak için maddelerin %25 ve %50'si rastgele seçilerek b parametresinde +1.00 logit değişim olacak şekilde odak ve referans grup arasında fark oluşturulmuştur. Sadece tek biçimli DMF koşuluna bakılmıştır. Test başlama kuralı beş madde uygulandıktan sonra ve yetenek kestirim yöntemi olarak beklenen sonsal dağılım (BSD) yöntemi seçilmiştir. Çalışmada ele alınan koşullar, i) DMF'li madde oranları (%25 ve %50), ii) madde seçim yöntemi [Maksimum Fisher Bilgisi (Maximum Fisher Information-MFB) ve Kullback-Leibler Bilgisi (Kullback-Leibler Information-KLB)], iii) madde kullanım sıklığı kontrol yöntemi [Azalarak kaybolma (fade-away) yöntemine göre 0.20 değişim olduğu ve madde kullanım sıklığının kontrol edilmediği] ve iv) test durdurma kuralıdır [standart hata<0.40, 0.20 ve sabit uzunluk (15, 30 madde)]. Çalışmada toplam 32 koşul, 100 replikasyon ile incelenmiştir. Rstudio catIRT paketi ile veriler üretilmiş, CAT analizleri SimulCAT programı ile yapılmıştır. Üretilen verilerde DMF olup olmadığı difR paket programı ile test edilmiştir. Simülasyonlar sonucunda, Test Bilgi Fonksiyonu ve korelasyon değerleri hesaplanmış ve bu değerler aracılığıyla ele alınan koşulların performansları değerlendirilmiştir.

[1] TED Üniversitesi
[2] Ankara Üniversitesi

Beklenen Bulgular

Analizler henüz tamamlanmamıştır. DMF'li madde oranının %25 olduğu ve madde kullanım sıklığı kontrol yöntemi olarak azalarak kaybolma (fade-away) yöntemine göre 0.20 değişim olduğu koşulda madde bilgi fonksiyonu değerlerinin DMF'li madde oranının %50 ve madde kullanım sıklığının kontrol edilmediği duruma göre daha yüksek olması beklenmektedir. Madde seçim yöntemi olarak bazı çalışmalarda (örn. Deng, Ansley ve Chang, 2010) Maksimum Fisher Bilgisi (MFB) daha iyi sonuçlar ürettiği görülürken bazı çalışmalarda (örn. Balta & Uçar, 2022) Kullback-Leibler Bilgisi (KLB) daha iyi sonuçlar ürettiği, bazı çalışmalarda (Gündeğer ve Doğan, 2018) ise her iki seçim yönteminin de benzer sonuçlar ürettiği görülmüştür. Alanyazında farklı koşullar altında yapılan çalışmalarda, bu iki madde seçme yönteminin performansları hakkında tutarlı bulgular olmadığı görülmüştür.

Kaynakça

Balta, E., & Uçar, A. (2022). Investigation of measurement precision and test length in computerized adaptive testing under different conditions. E-International Journal of Educational Research, 13(1), 51-68. DOI: https://doi.org/10.19160/e-ijer.1023098

Deng, H., Ansley, T., & Chang, H. H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. Journal of Educational Measurement, 47(2), 202-226. doi: 10.1111/j.17453984.2010.00109.x

Gündeğer, C. & Doğan, N. (2018). Bireyselleştirilmiş bilgisayarlı sınıflama testi kriterlerinin test etkililiği ve ölçme kesinliği açısından karşılaştırılması. Journal of Measurement and Evaluation in Education and Psychology, 9 (2), 161-177. DOI: 10.21031/epod.401077

Lei, P. W., Chen, S. Y. & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. Journal of Educational Measurement, 43(3), 245-264. Retrieved from http://dx.doi.org/10.1111/j.1745-3984.2006.00015.x

Piromsombat, C. (2014). Differential item functioning in computerized adaptive testing: Can cat self-adjust enough? (Doctoral Dissertation). University of Minnesota.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. In W. J. van der Linden & C. A.W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 221–244). Boston: Kluwer Academic Publishers.

Anahtar Kelimeler: değişen madde fonksiyonu, bilgisayar ortamında bireye uyarlanmış test, simülasyon çalışması

Bilgisayar Ortamında Bireye Uyarlanmış Test Simülasyonlarında Replikasyon Sayısının Farklılaşmasının Ölçme Kesinliğine Etkisi

Serap Büyükkıdık[1]

ÖZET

Giriş

Son yıllarda teknoloji kullanımının da yaygınlaşması ile birlikte Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) (Computerized Adaptive Testing (CAT)) uygulamaları yaygınlaşmaktadır. BOBUT uygulamalarında gerçek veri seti üzerinde yapılan araştırmaların yanı sıra çeşitli simülasyon koşullarında da araştırmalar gerçekleştirilmektedir. Bu simülasyon araştırmalarında araştırmacıların karar vermesi gereken hususlardan biri ise replikasyon sayısıdır. Harwell, Stone, Hsu ve Kirisci (1996) replikasyon sayısının MTK temelli simülasyon araştırmalarında minimum 25 olması gerektiğini önermişlerdir. Alanyazına bakıldığında ise 5 (Bolt ve Lall, 2003), 10 (Martelli, Matteucci ve Mignani, 2016), 25 (Boztunc Ozturk & Sahin, 2019), 100 (Tat, 2020) replikasyonla yapılan birçok simülasyon çalışması bulunmaktadır. Simülasyon çalışmalarında araştırmacıların çalışılan koşulların tespit etmesinin yanı sıra kaç replikasyon yapılacağını da belirlemesi önem arz eder. Bu araştırmada BOBUT simülasyonunda 1, 10, 25, 100, 1000 ve 10000 replikasyon yapıldığında Yanlılık (Bias), Hata Kareler Ortalaması Karekökü (Root Mean Squared Error (RMSE) ve Ortalama Mutlak Hata (Mean Absolute Error (MAE)) değerlerinin nasıl değiştiği incelenmiştir. Böylelikle BOBUT simülasyonlarında replikasyon sayısının farklılaşmasının ölçmenin kesinliğine etkisi ele alınmıştır.

Yöntem

Bu araştırma BOBUT simülasyon çalışmasıdır. Simülasyon koşullarının belirlenmesinde alanyazın taramasına başvurulmuştur. Gu, Ling, ve Qu (2019) tarafından yapılan BOBUT araştırmasında ele alınan parametre değerleri ve dağılımları bu araştırmadaki simülasyon çalışmasında da kullanılmıştır. BOBUT örneklemindeki adayların yetenek dağılımları, ortalaması 0 ve standart sapması 1 olan normal dağılımdan (N(0,1)) türetilmiştir. BOBUT simülasyonunda "maddelerinin a-parametreleri, ortalama 0.8 ve standart sapma 0.3 olan bir log-normal dağılımdan; b-parametreleri, ortalama -0,4 ve standart sapma 1,05 olan bir normal dağılımdan ve c-parametreleri, ortalama 0,18 ve standart sapma 0,09 olan bir beta dağılımından türetilmiştir (Gu, Ling ve Qu, 2019: s.2)". Bu araştırmada Gu, Ling ve Qu'nun (2019) araştırmasında olduğu gibi 3000 örneklemle, içerik dengeleme ile ve maruz kalma kontrolü yapılmadan (no content balancing ve no exposure control) simülasyon gerçekleştirilmiştir. Madde seçim yöntemi olarak Maksimum Fisher Bilgisi, sonlandırma kuralı olarak 0.35 standart hatanın altı ele alınmıştır. Madde seçiminde Maksimum Fisher Bilgisi ve standart hata miktarının sonlandırma kuralının ele alınmasında Aybek (2016)'nın bu yöntemlerin en etkili yöntemler olduğunu bulması etkili olmuştur. Replikasyon sayılarının belirlenmesinde alanyazın taramasına başvurulmuştur. Ölçme kesinliğinin belirlenmesinde Yanlılık (Bias), Hata Kareler Ortalaması Karekökü (Root Mean Squared Error (RMSE), ve Ortalama Mutlak Hata (Mean Absolute Error (MAE)) kullanılmıştır. Yetenek parametresinin kestirilmesinde ise Maksimum Olabilirlik Kestirimi (Maksimum Likelihood Estimation (MLE)), Bayesçi Maksimum Sonsal Dağılım Kestirimi (MAP), ve Bayesçi Beklenen Sonsal Dağılımı Kestirimi (EAP) kullanılmıştır.

[1] Sinop Üniversitesi

Bulgular/Beklenen bulgular

Araştırma sonucunda özellikle 100 replikasyondan sonra replikasyon sayısı arttırmanın simülasyon zamanını oldukça arttırdığı ortaya çıkmıştır. Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz, 64 bit işletim sistemli bilgisayarda MLE kestiriminde bahsedilen koşullarda 10000 replikasyon yapmak yaklaşık 49 dakika, MAP kestiriminde aynı replikasyonu yapmak yaklaşık 36 dakika, EAP kestiriminde ise aynı replikasyonu yapmak 52 dakika sürmüştür. BOBUT performansını belirlemede ölçme kesinliği kriterleri incelendiğinde, Yanlılık, RMSE ve MAE değerlerinin her replikasyon koşulunda benzer sonuçları verdiği bulunmuştur. Araştırmacılara 25 replikasyon yaparak BOBUT simülasyonunu yapmaları önerilebilir. Bu araştırmada 3000 örneklem için replikasyon sayısının farklılaşmasının BOBUT uygulamalarında ölçme kesinliğine etkisine bakılmıştır. Farklı örneklem büyüklüğü, test koşulları, madde ve birey parametreleri, sonlandırma ve madde seçim kuralları dikkate alınarak daha kapsamlı araştırmalar tasarlanabilir.

Kaynakça

Aybek, E. C. (2016). *Kendini Değerlendirme Envanteri'nin bilgisayar ortamında bireye uyarlanmış test (BOBUT) olarak uygulanabilirliğinin araştırılması*. Unpublished dissertation. Ankara University, Ankara.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414. https://doi.org/10.1177/0146621603258350

Boztunc Ozturk, N., & Sahin, M. G. (2019). Effects Of Item Pool Characteristics On Ability Estimate And Item Pool Utilization: A Simulation Study. *Hacettepe University Journal of Education, 34*(2): 473-486. https://doi.org/10.16986/HUJE.2018042418

Gu, L., Ling, G., & Qu, Y. (2019). A Modified a-Stratified Method for Computerized Adaptive Testing. *ETS Research Report Series*, *2019*(1), 1-27.

Harwell, M., Stone, C.A., Hsu, T.-C., & Kirisci, L.(1996). Monte carlo studies in item response theory. *Applied Psychological Measurement. 20*, 101–125. https://doi.org/10.1177/014662169602000201

Martelli, I., Matteucci, M., & Mignani, S. (2016). Bayesian estimation of a multidimensional additive graded response model for correlated traits. *Communications in Statistics – Simulation and Computation, 45*(5), 1636-1654. https://doi.org/10.1080/03610918.2014.932804

Tat, O. (2020). *Açıklayıcı madde tepki modellerinin bilgisayar ortamında bireye uyarlanmış testlerde kullanımı*. Unpublished dissertation. Hacettepe University, Ankara.

Ölçme Değerlendirme Okuryazarlığına Yönelik Bir Bilgisayar Ortamında Bireye Uyarlanmış Test Geliştirme Süreci

Beyza Aksu Dünya[1] Mehmet Can Demir[1]

ÖZET

Giriş

Yükseköğretimde akademik personel ölçme değerlendirme bağlamında çok az formel hazırlığa sahiptir (Knapper, 2010). Dolayısıyla, akademisyenlerin biçimlendirici geri bildirim ve öğrenci merkezli ölçme değerlendirme uygulamaları minimal düzeyde kalmıştır (Massey, DeLuca & LaPointe-McEwan, 2020). Bu nedenle, yükseköğretimdeki ölçme uygulamalarının özetleyici değerlendirme ağırlıklı olması eleştirilmiştir (Yorke, 2003). Bu eleştirilere cevap olarak, yükseköğretim kurumları tarafından çeşitli eğiticilerin eğitimi programları geliştirmiştir (Taylor & Colet, 2010). Bu eğitimlerin hedefe yönelik olabilmesi, akademisyenlerin mevcut ölçme değerlendirme okuryazarlıklarının (assessment literacy) tespitiyle doğrudan ilişkilidir.

Bu çalışmanın genel amacı, bilgisayar ortamında bireye uyarlanmış test (BOBUT) uygulamasının avantajlarını kullanarak, yükseköğretimdeki akademik personellerin ölçme değerlendirme okuryazarlık düzeylerini ölçecek bir ölçme aracı geliştirilmesidir. Araştırma bu amaca bağlı olarak, üç aşamada gerçekleştirilecektir. Araştırmanın birinci aşamasında, akademik personelin ölçme ve değerlendirme okuryazarlığı düzeyini belirlemede kullanılacak madde havuzu hazırlanacaktır. Araştırmanın ikinci aşamasında ön uygulama ile madde parametre kestirimleri yapılacak, madde uyum istatistikleri ve değişen madde fonksiyonu incelenecek ve post-hoc simülasyonlara göre BOBUT'un psikometrik özellikleri belirlenecektir. Araştırmanın üçüncü aşamasında ise, esas BOBUT uygulaması yapılacak, test uzunlukları, test süreleri ve özellik kestirimlerinin hassasiyetleri belirlenecektir.

Yöntem

Madde havuzu oluşturulurken ilk olarak sistematik literatür taraması yapılarak ölçme değerlendirme okuryazarlığı kavramının kuramsal çerçevesi çizilmiş ve baskın kuramlar gözetilerek belirtke tablosu hazırlanmıştır. Sistematik tarama yapılırken makaleler sıralanan ölçütlere göre seçilmiştir: (a) Kör hakemlik sürecinden geçerek yayınlanmış makaleler, (b) tam metin erişimi olan makaleler, (c) doğrudan yükseköğretim bağlamını ele alan makaleler. Çalışmaların hariç tutulma kriterleri ise şu şekildedir: (a) Hakemlik süreci olmaksızın yayınlanmış makaleler, teknik raporlar ve tezler, (b) doğrudan yükseköğretim bağlamından olmayan makaleler (örneğin, öğretmenler).

Daha sonra, ölçülen yapının (ölçme değerlendirme okuryazarlığı) davranışsal göstergeleri belirlenmiş ve madde yazımına başlanmıştır. Yeni maddelerin yanı sıra, halihazırda psikometrik özellikleri belirlenmiş ölçme araçlarındaki uygun ve kullanımına izin verilen maddeler de yükseköğretim bağlamına göre revize edilerek madde havuza dahil edilecektir.

Madde havuzuna dahil edilen maddelere nihai hali verildikten sonra kalibrasyon çalışmasına başlanacaktır. Kalibrasyon aşamasında, parametre kestiriminde kullanılacak ölçme modelinin (Rasch model) varsayımları ile model uyumu incelenecek ve madde parametrelerinin kestirimi için deneme uygulaması yapılarak model uyumu incelenecek ve bilgisayar ortamında doğrusal (yani bireye uyarlanmamış) bir test uygulaması ile toplanan verilerle madde parametreleri kestirilecektir.

Bulgular/Beklenen Bulgular

[1] Bartın Üniversitesi

Sistematik tarama sonucunda, yükseköğretimde ölçme değerlendirme okuryazarlığı için 5 farklı bilgi alanından oluşan bir test kapsamı oluşturulmuştur. Bu alanlar aşağıdaki gibidir:

• Kazanımlara uygun, çok sayıda yüksek kaliteli ölçme araçları kullanma,

• Öğrenci performansını dışsal hata kaynaklarını gözeterek değerlendirme,

• Değerlendirmelerin uygulamasını ve puanlamasını uygun şekilde gerçekleştirme,

• Ölçme sonuçlarını farklı düzeyden paydaşlara doğru şekilde aktarma ve

• Ölçme sürecini etik ve yasal olarak gerçekleştirme (Brookhart, 2011; Popham, 2009; Stiggins, 1991).

Belirlenen bu bilgi alanlarına yönelik olarak, BOBUT uygulaması için madde havuzu oluşturulması amacıyla madde yazımı halen devam etmektedir. Hedeflenen madde havuzu çeşitli okuryazarlık düzeylerindeki cevaplayıcılara hitap edecek farklı madde güçlük parametresi ranjına sahip olacaktır.

Kaynaklar

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. Educational Measurement: Issues and Practice, 30(1), 3-12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Knapper, C. (2010). Changing teaching practice: Barriers and strategies. In J. C. Hughes and J. Mighty (Eds.), Taking stock: Research on Teaching and Learning in Higher Education (pp. 229–242). School of Policy Studies Queens University.

Massey, K. D., DeLuca, C., & LaPointe-McEwan, D. (2020). Assessment literacy in college teaching: Empirical evidence on the role and effectiveness of a faculty training course. To Improve the Academy: A Journal of Educational Development, 39(1).
http://dx.doi.org/10.3998/tia.17063888.0039.109

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? Theory into Practice, 48(1), 4-11. https://doi.org/10.1080/00405840802577536

Stiggins, R. J. (1991). Assessment literacy. Phi Delta Kappan, 72(7), 534-539.

Taylor, K. L., & Colet, N. R. (2010). Making the shift from faculty development to educational development. In A. Saroyan & M. Frenay (Eds.), Building teaching capacities in higher education: A comprehensive international model (pp. 139-167). Stylus Publishing.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. Higher Education, 45(4), 477-501.

Anahtar Kelimeler: bilgisayar ortamında bireye uyarlanmış test, ölçme değerlendirme okuryazarlığı, yükseköğretim.

Bilgisayar Ortamında Bireye Uyarlanmış Testlere Karşı Öğrenci Tutumları: Teknoloji Kabul Modeli ile Ölçek Geliştirme Çalışması

Beyza Arpacıoğlu[1] Serkan Arıkan[1]

ÖZET

1. GİRİŞ

Son yıllarda, bilgi ve iletişim teknolojisi alanında birçok ilerleme kaydedilmiştir. Eğitim sisteminin temel öğelerinden biri olan ölçme ve değerlendirme uygulamaları da bu değişikliklerden en çok etkilenen ve bu değişimlere hızlıca adapte olan alanlar arasındadır (Bennett, 2002; Linn, 1993). Bilgisayar kullanımındaki önemli artış, özellikle ölçme ve değerlendirme alanında bilgisayar tabanlı uygulamalar ile etkisini göstermiştir (McDonald, 2002). Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) uygulamaları ve bu testleri alan katılımcı sayısı da bilgisayar kullanımındaki artış ile birlikte önemli düzeyde yükselmiştir. Bu katılımcıların BOBUT uygulamalarına yönelik tutumlarını anlamak, gelecekteki uygulamaların daha etkili olması için önemli bir gereksinim haline gelmiştir (Lilley, Pyper & Wernick, 2011; Özbaşı, 2016). Ancak, literatürde katılımcıların BOBUT uygulamalarına yönelik tutumlarını araştıran çalışma sayısı sınırlıdır (Daphine, Sivakumar & Selvakumar, 2020; Lilley, Pyper & Wernick, 2011; Nikolaus, Bode, Taal, Vonkeman, Glas, van de Laar, 2014; Özbaşı, 2016; Schmidt, Urry & Gugel, 1978; Tonidanel & Quiñones, 2000). Ayrıca, katılımcıların BOBUT uygulamalarına yönelik tutumlarını ölçen bir ölçek geliştirmeyi amaçlayan çalışmalar da sayıca yetersizdir. Bu çalışmanın amacı, öğrencilerin BOBUT uygulamalarına yönelik tutumlarını ölçümlemek için bir ölçek geliştirmektir.

Bu çalışmanın alan yazına özellikle iki noktada katkı sağlaması beklenmektedir: Birincisi, katılımcıların BOBUT uygulamalarına yönelik tutumlarını ölçen ölçekler çoğunlukla sağlık sektörü, işe alım süreçleri ve çalışan değerlendirme prosedürleri sonrasında kullanılmaktadır. Eğitim alanında öğrencilerin BOBUT uygulamalarına yönelik tutumlarını ölçen ölçekler sınırlıdır. Bu nedenle, eğitim alanında BOBUT uygulamalarına yönelik tutumları araştırmak, gelecekteki eğitim testleri için BOBUT uygulamalarının iyileştirilmesine katkı sağlayacaktır. İkincisi, bu çalışmada geliştirilen ölçek, gerçek bir BOBUT deneyiminden sonra katılımcılardan veri toplayarak oluşturulacak olması nedeniyle önemlidir. Önceki araştırmalarda, verilerin, katılımcılardan BOBUT deneyimi yaşamadan teorik bilgilerine göre toplandığını göstermektedir. Katılımcılardan gerçek bir BOBUT deneyiminden hemen sonra veri toplayarak BOBUT uygulamasına yönelik tutumları ölçen bir ölçek geliştirmenin literatüre büyük katkı sağlaması beklenmektedir.

Bu çalışmanın araştırma soruları şöyledir:

1-Öğrencilerin BOBUT uygulamalarına yönelik tutumunu ölçmek için kullanılan ölçek puanları iç tutarlı mıdır?

2- Öğrencilerin BOBUT uygulamalarına yönelik tutumunu ölçmek için kullanılan ölçek puanları geçerli midir?

3- Cinsiyete göre ölçme değişmezliği sağlanmakta mıdır?

2. YÖNTEM

2.1 Örneklem

[1] Boğaziçi Üniversitesi

15 maddeden oluşan ölçek, amacı doğrultusunda özel okul-devlet okulu ve kız erkek, yaklaşık 300 4. Sınıf öğrencisine uygulanacaktır.

2.2 Araştırmanın Modeli

Bu çalışma, bir ölçek geliştirme çalışmasıdır. Bu amaç doğrultusunda, Crocker & Algina (2008) ve DeVellis (2017) tarafından ölçek geliştirmek için izlenmesi gereken adımlar takip edilmiştir.

Davis (1993) tarafından önerilen TAM, bu çalışmanın kuramsal çerçevesini oluşturmaktadır. Geliştirilen ölçek 3 boyut içermektedir: Algılanan Kullanışlılık (ALK), Algılanan Kullanım Kolaylığı (ALKK) ve Algılanan Eğlence (ALE). ALK ve ALKK, Teknoloji Kabul Modeli (TKM) içerisinde yer almakla birlikte bilgi teknolojileri kabul ve kullanımı konusunda etkili oldukları savunulmaktadır (Davis, 1993). ALE boyutunun ise bilgi teknolojileri kabul ve kullanımı konusunda etkili olduğunu gösteren çalışmalar yer almaktadır ve bu boyut literatürde önemli bir yere sahiptir. Bu ölçeğin maddeleri, ALK ve ALKK için TKM'de, ALE için ise literatürde yer alan tanımlara göre oluşturulmuştur. Bu tanımlar şöyledir:

- "ALK, katılımcıların, özellikle geleneksel kağıt-kalem testine kıyasla, BOBUT uygulamasını daha avantajlı bir şekilde kullanabileceğine inanma derecesidir."
- "ALKK, katılımcıların BOBUT kullanmanın zorluk veya çaba gerektirmeyen bir iş olduğuna inanma derecesidir".
- "ALE, katılımcıların BOBUT uygulaması için genel memnuniyet durumudur".

Bu ölçek, kesinlikle katılmıyorum (1)'dan, kesinlikle katılıyorum (4)'a kadar dört cevap seçeneği ile bağlantılı 15 maddeden oluşan Likert Tipi bir ölçme aracı olarak planlanmıştır.

2.3 Verilerin Analizi

Test maddelerinin iç tutarlılığı hesaplanacak ve geçerlik kanıtları toplanacaktır. Önerilen modelin doğrulanmasını test etmek amacıyla Mplus ile doğrulayıcı faktör analizi kullanılacak ve uyum iyiliği indeksleri değerlendirilecektir. Ölçeğin madde analizi Madde Tepki Kuramı kullanılarak yapılacaktır. Güçlük ve ayırt edicilik parametreleri ile birlikte madde bilgi fonksiyonları da analiz edilecektir. Ölçme değişmezliği cinsiyet bağlamında test edilecektir.

3. BEKLENEN BULGULAR

Öğrencilerin BOBUT uygulamalarına yönelik tutumlarını ölçmek için kullanılan ölçek puanlarının, iç tutarlılık değerlendirmesi açısından, Cronbach Alfa katsayısı değerinin kabul edilebilir olması beklenmektedir. Cichetti (1994) kriterlerine göre, bu değer 0.70'den büyük olduğu durumlarda kabul edilebilirdir. Aynı zamanda, geçerlik değerlendirmesi için, uyum indekslerinin, TLI,CFI ve RMSEA değerlerinin, kabul edilebilir olması beklenmektedir. Ullman (2001) kriterlerine göre elde edilen değerler değerlendirilecek olup, RMSEA için 0.06'dan düşük değerler, CFI ve TLI için ise 0.95'ten büyük değerler beklenmektedir. Faktör yükleri 0.40'tan büyük olmalıdır (Salkind, 2010). Biçimsel değişmezlik, metrik değişmezlik ve skalar değişmezlik değerlendirmelerinin cinsiyet açısından iki grup arasında aynı anlama sahip olduğunu göstermesi de beklenen bulgular arasındadır.

KAYNAKÇA

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. Journal of Technology, Learning and Assessment, 1(1), 1-24.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284–290.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Mason, OH: Cengage Learning.

Daphine, R., Sivakumar, P., & Selvakumar, S. (2020). A study on student's attitude towards online computer adaptive test (CAT) in physics education through observation schedule. Journal of Xidian University, 14(5), 4703-4708.

Davis, F. D. (1993). User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. International Journal of Man-Machine Studies, 38(3), 475-487.

DeVellis, R.F. (2017). Scale development: Theory and applications (4th ed.). Los Angeles: SAGE Publications.

Lilley, M., Pyper, A., & Wernick, P. (2011). Attitudes to and usage of CAT in assessment in higher education. Innovation in Teaching and Learning in Information and Computer Sciences, 10(3), 28-37.

Linn, R. L. (ed.). (1993). Educational Measurement (3th ed). Phoenix, AZ: American Council on Education and the Oryx Press.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. Computers & Education, 39(3), 299-312.

Nikolaus, S., Bode, C., Taal, E., Vonkeman, H. E., Glas, C. A., & van de Laar, M. A. (2014). Acceptance of new technology: a usability test of a computerized adaptive test for fatigue in rheumatoid arthritis. JMIR Human Factors, 1(1), e3424.

Özbaşı, D. (2016). Bilgisayar ortamında bireye uyarlanmış test uygulamasına ve kağıt- kalem testine katılan öğrenci görüşlerinin incelenmesi. Journal of International Social Research, 9(42).

Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer-assisted tailored testing: Examinee reactions and evaluations. Educational and Psychological Measurement, 38(2), 265-273.

Tonidanel, S., & Quiñones, M. A. (2000). Psychological reactions to adaptive testing. International Journal of Selection and Assessment, 8(1), 7-15.

Ullman, J. B. (2001). Structural equation modeling. In B. Tabachnick & L. S. Fidell (Eds.), Using multivariate statistics (4th ed., pp.653-771). Boston: Allyn &Bacon.

Anahtar Kelimeler: BOBUT, tutum.TKM

Mesleki Alan İlgi Envanteri'nin Bilgisayar Ortamında Bireye Uyarlanmış Formunun Geliştirilmesi

Volkan Alkan, Kaan Zülfikar Deniz[1]

ÖZET

BOBUT uygulamasının, kâğıt-kalem uygulamalarına karşı öne çıkan birçok önemli özelliği olmasına rağmen, dünyada ve ülkemizde BOBUT uygulamaları yeterince kullanılmamaktadır. Günümüzün gelişmiş teknolojisine rağmen, hâlihazırda birçok ölçme aracı kâğıt-kalem formatında uygulanmaktadır. BOBUT uygulaması olarak geliştirildiğinde uygulama süresi kısalacak ve eklenen yeni madde ve alt ölçeklerle kapsam geçerliği daha yüksek olacak bu ölçme araçlarının, henüz BOBUT uygulamalarının geliştirilmemiş olması, bu ölçme araçlarının uygulama alanını oldukça kısıtlı tutmaktadır. Bu araştırmaya konu olan Mesleki Alan İlgi Envanteri (MAİ) de henüz BOBUT uygulaması geliştirilmemiş envanterlerden biridir. MAİ 14 alt ölçeği bulunan, 156 maddeden oluşan ve kâğıt-kalem uygulaması yaklaşık 15-20 dakika süren bir ölçme aracıdır. Bu envanterin BOBUT uygulaması geliştirildiğinde, açıktır ki hem envanterin uygulama süresi kısalarak kullanışlılığı artacak hem de yeni meslek alanlarının ortaya çıkması halinde envanterin uygulama süresinin uzamasına ilişkin herhangi bir endişe duyulmadan envantere yeni alt ölçekler eklenebilecektir. Tüm bu sebeplerden ötürü, bu araştırmanın temel amacını, öğrenci yönlendirme hizmetlerinde öğrencilere meslek seçimi konusunda yardımcı olmak için geliştirilmiş olan MAİ'nin BOBUT formunun en uygun MTK modeli, test sonlandırma kuralı ve madde seçim yöntemi belirlenerek geliştirilmesi oluşturmaktadır.

Araştırmada veri toplama aracı olarak Deniz (2009) tarafından geliştirilen MAİ'nin kağıt kalem formu (MAİ-KK) kullanılmıştır. Araştırmada iki farklı çalışma grubu bulunmaktadır. Birinci çalışma grubu, post-hoc simülasyon uygulamasında temel alınan verilerin elde edildiği MAİ-KK uygulamasının gerçekleştirildiği gruptur. Bu grupta 2018-2019 öğretim yılında Mersin ilinde farklı türdeki liselerde 10, 11 ve 12. sınıfta öğrenim görmekte olan 1425 öğrenci yer almaktadır. Birinci çalışma grubuna yapılan uygulama sonucunda, MAİ-KK formundan elde edilen veriler kullanılarak MAİ'nin KTM ve GKPM'ye göre madde parametreleri elde edilmiştir. Ardından, uygulama sonucunda elde edilen madde parametlereleri temel alınarak Firestar (Choi, 2009) yazılımı aracılığıyla BOBUT formu için post-hoc simülasyonu gerçekleştirilmiştir. BOBUT simülasyonu her bir alt ölçek için ayrı ayrı gerçekleştirilmiş ve her bir alt ölçeğin ortalama kaç maddede sona erdiği, ortalama standart hata değerleri, tüm maddeler ile kestirilen θ düzeyleri ile simülasyon sonucunda kestirilen θ düzeyleri arasındaki ilişkinin belirlenmesi amacıyla korelasyon katsayıları hesaplanmıştır. Simülasyonlarda, MTK modeli olarak KTM ve GKPM, test sonlandırma kuralı olarak .30, .40 ve .50 standart hata değerleri ve madde seçim yöntemi olarak Beklenen En Yüksek Bilgi (BEYB), Fisher En Yüksek Bilgi (FEYB), Beklenen En Yüksek Sonsal Ağırlıklandırılmış Bilgi (BEYSAB) ve Beklenen En Düşük Sonsal Varyans (BEDSV) yöntemleri kullanılmıştır. İkinci çalışma grubu ise hem MAİ kağıt-kalem formu çevrimiçi uygulaması (MAİ-KKÇU) hem de MAİ-BOBUT uygulamasının gerçekleştirildiği gruptur. Bu grupta Mersin İlinde farklı türdeki liselerin 10, 11 ve 12. sınıflarında öğrenim görmekte olan 150 öğrenci yer almaktadır. Bu 150 kişilik gruba ise, Concerto platformunda geliştirilmiş olan MAİ-B uygulaması çevrimiçi ortamda uygulanmıştır. Benzer şekilde, MAİ-KKÇU'da Google Anket Uygulaması kullanılarak aynı öğrencilere uygulanmıştır. Sıra etkisinin önüne geçmek amacıyla hem uygulamalar arasına 15 günlük süre konulmuş hem de ilk uygulamada MAİ-KKÇU alan 75 öğrencinin ikinci uygulamada MAİ-B alması sağlanmıştır.

[1] Ankara Üniversitesi

Çalışmanın ilk aşamasında toplanan veriler üzerinden gerçekleştirilen simülasyonlardan elde edilen sonuçlara göre BOBUT uygulaması için en ideal ölçütlerin, MTK modeli olarak GKPM, test sonlandırma kuralı olarak .40 standart hata değeri, madde seçim yöntemi olarak da FEYB yöntemi olduğuna karar verilmiştir. Ayrıca, uygun görülen ölçütlerle gerçekleştirilmiş olan BOBUT simülasyonu sonucunda kâğıt-kalem formu 156 madde olan MAİ ortalama 59 madde ile sonlanmış ve kâğıt-kalem formundan alınan puanlar ile simülasyonla kestirilen teta (θ) düzeyleri arasındaki korelasyonların .91-.97 aralığında olduğu görülmüştür. Post-hoc simülasyon sonuçlarına göre BOBUT uygulaması geliştirilmiş ve 150 öğrenciye uygulanmıştır. Öğrencilerin kâğıt-kalem formunun çevrimiçi uygulamasından aldıkları puanlar ile BOBUT formu ile kestirilen θ düzeyleri arasındaki korelasyonların .73 ile .91 arasında değiştiği görülmüştür. Ayrıca, BOBUT uygulamasının kâğıt-kalem formunun çevrimiçi uygulamasına göre, madde sayısı açısından yaklaşık %66 ve uygulama süresi açısından da yaklaşık %67 avantaj sağladığı tespit edilmiştir. Araştırma sonuçları, Mesleki Alan İlgi Envanteri'nin BOBUT olarak uygulanabilir olduğunu göstermektedir. Ayrıca madde havuzunun daha yüksek bilgi sağlayan maddeler ile genişletilmesinin ve envantere zamanın gereksinimlerine uygun olarak yeni mesleki ilgi alanlarının eklenmesinin envanteri BOBUT için daha elverişli hale getirebileceği düşünülmektedir.

KAYNAKÇA

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. Applied Psychological Measurement, 33(8), 644–645.

Deniz, K. Z. (2009). Occupational Interest Inventory (OFII) development study. Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 6(1), 289-310.

Anahtar Kelimeler: Bilgisayar Ortamında Bireye Uyarlanmış Test, Madde Tepki Kuramı, Mesleki Alan İlgi Envanteri, Mesleki İlgi

Test Anı Kaygı Düzeyine Bağlı Başlama Kuralının BOBUT Standart Ölçme Hatasına Etkisi

Serkan ARIKAN[1] Eren Can AYBEK[2] Güneş ERTAŞ[1]

ÖZET

GİRİŞ

Bilgisayar ortamında bireye uyarlanmış testlerin (BOBUT) kullanımı giderek yaygınlaşmaktadır. BOBUT algoritması, öğrencilerin düzeylerine en uygun soruları seçmektedir ve öğrencilerin düzeylerine uygun olmayan, yani öğrencinin düzeyine göre çok zor veya çok kolay sorular öğrencilerin karşısına gelmemektedir (Wainer, 2000). BOBUT ile ilgili en büyük sınırlılık sürdürülebilirliktir. Belirli bir alt yapı ve iş gücü istediği için kağıt kalem testlere göre daha maliyetli ve uzman iş gücü gerektirmektedir. Sürdürülebilirlik sağlandığında BOBUT çok daha verimli bir sistem sunmaktadır. BOBUT'un verimliliği sayesinde daha az soru ile daha az ölçme hatasına sahip puanlama yapılabilmektedir (Thompson & Weiss, 2011).

BOBUT'un verimliliğini artırmak için araştırmalar devam etmektedir. Öğrenciler hakkında demografik bilgiler, başarı durumları veya tutumları ile ilgili ön bilgiler daha etkili bir BOBUT sistemi için kullanılabilmektedir (Castro, Suarez ve Chirinos, 2010). Eğer öğrenciler hakkında ön bilgiye sahip olunursa sadece testteki sorular değil, testin başlama noktası da bireye uyarlanarak değişkenlik gösterebilir. Bu ön bilgi kişilerin akademik başarıları ile ilgili bilgiler olabileceği gibi, ölçme hatası ile ilişkili olabilecek ve ölçülen yapı hakkında istenmeyen bir varyansa neden olabilecek duyuşsal özellikler de olabilir. Örneğin, öğrencilerin test anı kaygı düzeylerinin test puanlarında istenmeyen bir varyans yarattıkları bilinmektedir (Orbach, Herzog ve Fritz, 2019). Eğer test anı kaygı durumunun ölçme hatasına etkisi kontrol edilebilirse ortaya çıkacak puanların daha az ölçme hatası içereceği düşünülmektedir.

Bu çalışmanın amacı, öğrencilerin test anı kaygı durumlarını tespit ederek, kaygı düzeylerini azaltabilecek çok basit sorularla teste başlanmasının ölçme hatasına etkisini incelemektir. Özellikle test anı kaygı düzeyi çok yüksek olan öğrencilerin çok basit sorular ile başlamaları durumunda daha az ölçme hatası oluşup oluşmayacağı incelenecektir.

Araştırma Sorusu:

1) Test uzunluğu sabit tutulduğunda, test anı kaygı düzeyi ve başlangıç sorularının zorluğu öğrencilerin ölçme hatalarını nasıl etkilemektedir?

2) Ölçme hatası sabit tutulduğunda, test anı kaygı düzeyi ve başlangıç sorularının zorluğu test uzunluğunu nasıl etkilemektedir?

YÖNTEM

Katılımcılar

Çalışmanın katılımcıları özel ve devlet okullarında okuyan 4. sınıf öğrencileridir. Yaklaşık 600 öğrenciye ulaşılması hedeflenmektedir. Çalışmaya katılmayı kabul eden okullardaki öğrenciler çalışmaya katılacağı için uygun örnekleme (convenience sampling) yöntemi kullanılacaktır.

[1] Boğaziçi Üniversitesi
[2] Pamukkale Üniversitesi

Ölçme Aracı

Bu çalışmanın ölçme araçları BOBUT formatında geliştirilmiş 4. sınıf matematik testi ve test anı kaygı ölçeğidir. BOBUT testinin geliştirilmesi için 540 madde hazırlanmış, 3108 öğrenciden eksik test deseni kullanılarak veri toplanmış ve madde kalibrasyonları Rasch modeli kullanılarak yapılmıştır. Madde kalibrasyonları sonucunda geliştirilecek BOBUT uygulaması için 513 sorunun yer aldığı soru bankası oluşturulmuştur. Sorular MEB kazanımlarına göre hazırlanmıştır ve sayılar, geometri, ölçme ve veri konu alanlarını içermektedir. TIMSS değerlendirme çerçevesine paralel olarak bilme, uygulama ve akıl yürütme düşünme süreçleri temel alınarak sorular hazırlanmıştır.

Test anı kaygı ölçeği ise araştırmacılar tarafından geliştirilmiş 12 maddelik, 4'lü Likert tipi bir ölçektir. Ölçeğin güvenirlik ve geçerlik analizleri Madde Tepki Kuramı (MTK) temelinde yapılmıştır. Test anı kaygı ölçeği için kesme puanları sırasıyla $\theta>1$, $-1<\theta<1$ ve $\theta<-1$ olarak belirlenecek ve öğrenciler test anı kaygı düzeyi yüksek, orta ve düşük olarak üç gruba ayrılabilecektir.

Veri Analizi

Çalışmada kullanılacak yöntemleri ve koşulları belirlemek amacıyla simulatif olarak 1000 kişiye ait yetenek düzeyi standart normal dağılımdan çekilmiştir. Buna göre madde seçim yönteminin MFI ve yetenek kestirim yönteminin ise MAP olarak belirlenmesine karar verilmiştir. Simülasyon sonuçlarına göre, araştırmanın koşulları standart hata için sabit tutulan değerin 0.50 ve sabit test uzunluğunun 10 olarak belirlenmesi uygun bulunmuştur.

Birinci araştırma sorusu için öğrencilere bilgisayar ortamında bir kaygı ölçeği uygulanacaktır. Öğrenciler kaygı düzeylerine göre üç grupta ele alınacaktır: kaygı düzeyi yüksek ($\theta>1$), kaygı düzeyi orta ($-1<\theta<1$) ve kaygı düzeyi düşük ($\theta<-1$). Ardından her bir gruba rastgele iki tip kitapçık atanacaktır. Birinci kitapçıkta ilk 4 soru çok basit soru olacak şekilde lineer formda olacak, ardından BOBUT sisteminde 10 soru ile test sonlanacaktır. İkinci kitapçıkta ise doğrudan BOBUT sistemindeki sorularla teste başlanacak ve 10 soru ile test sonlanacaktır. Bu durumda kaygı düzeyine göre ayrılan 3 gruba 2'şer kitapçık atandığından 6 grupta yer alan öğrencilerin yetenek kestiriminde standart hataları hesaplanacaktır. Gruplar arası ortalama standart hata farkları ve etkileşim faktöriyel ANOVA ile test edilecektir.

İkinci araştırma sorusu için yine öğrencilere bilgisayar ortamında bir kaygı ölçeği uygulanacaktır. Öğrenciler kaygı düzeylerine göre üç grupta ele alınacaktır: kaygı düzeyi yüksek ($\theta>1$), kaygı düzeyi orta ($-1<\theta<1$) ve kaygı düzeyi düşük ($\theta<-1$). Ardından yine her gruba rastgele kitapçık ataması yapılacaktır. Birinci kitapçıkta ilk 4 soru çok basit soru olacak şekilde lineer formda olacak, ardından BOBUT sistemindeki sorulara geçilecek ve standart hata 0.50'nin altına düşene kadar test devam edecektir. İkinci kitapçıkta ise BOBUT sistemindeki sorularla teste başlanacak ve standart hata 0.50'nin altına düşene kadar test devam edecektir. Bu durumda kaygı düzeyine göre ayrılan 3 gruba 2'şer kitapçık atandığından 6 grupta öğrencilerin test boyunca kaç madde ile karşılaştıkları bir başka deyişle test uzunlukları kaydedilecektir. Gruplar arası ortalama test uzunluğu farkları ve etkileşim faktöriyel ANOVA ile test edilecektir.

BULGULAR

Çalışmanın verileri Mayıs-Haziran 2023'te toplanacaktır. Beklenen bulgular, özellikle kaygı düzeyi yüksek olan öğrencilerin basit sorular ile başlamaları ile daha az standart hata ile ve daha az soru ile ölçümlerinin gerçekleşecek olmasıdır. Ayrıca bir etkileşim de gözlenebilir. Etkileşime göre, kaygı düzeyi arttıkça standart hata ve test uzunluğu artıyor olabilir ama bu artış basit sorular ile başlandıkça daha

az oranda olabilir. Tüm bu değerlendirmeler etki büyüklüğü de hesaplanarak ve karşılaştırılarak yapılacaktır.

Kaynakça

Castro, F., Suarez, J., & Chirinos, R. (2010). Competence´s initial estimation in computer adaptive testing. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Orbach, L., Herzog, M., & Fritz, A. (2019). Relation of state-and trait-math anxiety to intelligence, math achievement and learning motivation. Journal of Numerical Cognition, 5(3), 371-399.

Thompson, N., & Weiss, D. (2011). A Framework for the Development of Computerized Adaptive Tests. Practical Assessment, Research, and Evaluation, 16(1).

Wainer, H. (2000). Computerized Adaptive Testing: A Primer. Mahvah, NJ: Erlb

Anahtar Kelimeler: Canlı BOBUT, Başlama Kuralı, Test Anı Kaygı Düzeyi, Ölçme Hatası, Test Uzunluğu

BOBUT için Geniş Madde Havuzlarının Kalibrasyonunda Eksik Test Deseninin Kullanımı: Bir Matematik Testi Örneği

Eren Can Aybek[1] Serkan Arıkan[2] Güneş Ertaş[2]

ÖZET

GİRİŞ

Bilgisayar Ortamında Bireye Uyarlanmış Testler (BOBUT), maddelerin bir havuzdan bireyin o anki yetenek düzeyine göre seçilerek uygulandığı ve her bir yanıttan sonra bireyin yetenek düzeyinin yeniden kestirildiği bir döngüye sahiptir. Bu döngüde, bireyin büyük olasılıkla doğru ya da büyük olasılıkla yanlış yanıtlayacağı maddeler, havuzdan seçilmez ve böylece daha az sayıda madde ile bireyin yetenek düzeyi kestirilebilir. Bireyin vereceği yanıtın belli bir olasılıkla tahmin edilmesinin gerekliliği, Madde Tepki Kuramı'nı (MTK), BOBUT uygulamaları için oldukça kullanışlı hale getirmekte ve MTK, BOBUT uygulamalarının temelini oluşturmaktadır. BOBUT geliştirme sürecinde madde havuzunun tüm yetenek düzeyleri hakkında bilgi sağlayabilecek kadar çok maddeden oluşması; aynı zamanda bu havuzdaki maddelerin bir MTK modeline göre ölçeklenmesi gerekmektedir. Bu durum ise madde havuzundaki maddelerin çok sayıda kişiye uygulanmasını gerekli kılmaktadır.

Özellikle madde havuzunun çok sayıda maddeden oluşması durumunda, her bir katılımcının madde havuzunda yer alacak tüm maddeleri yanıtlaması ve ardından buna göre MTK kalibrasyonlarının yapılması uygulanabilirlikten çok uzaktır. Bunun yerine, maddelerin kitapçıklara ayrılarak, katılımcıların ortak sorular içeren farklı kitapçıkları alması ve ardından tek seferde madde kalibrasyonlarının yapılması çok daha kullanışlı bir yoldur (Gonzalez & Rutkowski, 2010). Bu araştırmada da BOBUT geliştirilmek için kullanılacak 540 maddelik bir madde havuzunun Rasch modeline göre ölçeklenmesi ve yapılacak olan bu ölçekleme işlemine ait adımların paylaşılması amaçlanmıştır.

YÖNTEM

Çalışmanın katılımcılarını Millî Eğitim Bakanlığı'na bağlı özel ve devlet ilk ve orta okullarına devam etmekte olan yaklaşık 3200 4. Sınıf ve 5. Sınıf öğrencisi oluşturmaktadır. Çalışmanın madde havuzunda yer alan 540 madde, 36 kitapçığa ayrılmıştır. Kitapçıklar oluşturulurken beşer madde çapa madde olarak atanmış ve 1. Kitapçığın son beş maddesi ile 2. Kitapçığın ilk beş maddesi; 2. Kitapçığın son beş maddesi ile 3. Kitapçığın ilk beş maddesi aynı olacak şekilde 36 kitapçık oluşturularak bir eksik test deseni oluşturulmuştur. Bu bağlamda 36 kitapçık doğrusal bir formda Concerto v5.1.0 üzerinde bilgisayara aktarılmış ve bazı katılımcılardan Concerto Platformu'nun yardımıyla, bazı katılımcılardan ise kağıt-kalem formundan veriler toplanmıştır. Toplanan veriler R 4.2.1 (R Core Team, 2022) üzerinde mirt 1.38.1 (Chalmers, 2012) paketi kullanılarak analiz edilmiştir. Analizler öncesinde yerel bağımsızlık varsayımı Yen'in Q3 istatistiği yardımıyla incelenmiş ve her bir kitapçığı alan öğrenci sayısının sınırlılığı nedeniyle Rasch modeli ile çalışılması tercih edilmiştir.

BULGULAR

Öncelikle maddelere verilen yanıtlar betimsel olarak incelenmiş ve sıfır varyansa sahip olan iki madde ile baskıda hata olan iki madde havuzdan çıkarılmıştır. Ardından Q3 istatistiği 0.20'nin üzerinde olan 23 madde yerel bağımsızlığı bozması nedeniyle madde havuzundan çıkarılmıştır. Buna göre

[1] Pamukkale Üniversitesi
[2] Boğaziçi Üniversitesi

kalibrasyonlar 517 madde ile yapılmıştır. Araştırmada madde kalibrasyonlarının ardından b parametrelerinin elde edilmesi, ayrıca madde havuzuna ait test bilgi fonksiyonu incelenecektir. Olası eksik test desenleri hakkında bilgiler sunulacaktır.

KAYNAKÇA

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. IEA-ETS Research Institute Monograph, 3, 125-156.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

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

Anahtar Kelimeler: eksik test deseni, madde kalibrasyonu, madde havuzu geliştirme, matematik testi

Bilişsel Basamaklara Göre Aşamalandırılmış Bireye Uyarlanmış Testlerle Okuduğunu Anlama Becerisinin Ölçülmesi

Selma ŞENEL[1] Ömer KUTLU[2]

ÖZET

Bilgisayar ortamında bireye uyarlanmış testler (BOBUT), yanıtlayıcıların yeterliklerine ilişkin sürekli bir ölçek üzerinde (Weiss, 2011) bilgi sunar. Yeterlik kestirimleri tipik olarak yaklaşık -5 ila +5 arasında değişir. Madde Tepki Kuramı'na (MTK) dayalı yeterlik kestirimleri yapıldığı için, farklı formları alan yanıtlayıcıların puanları bu metrik üzerinde karşılaştırabilir (Davey, 2011). Bireylerin yeterliklerine ilişkin sürekli bir ölçek üzerinde sunulan bir puanın yorumlanması, puan aralıklarına göre belirli yeterliklerin atfedilmesi, betimlenmesi ile mümkün olmaktadır.

Eğitimde, bilişsel özelliklerin aşamalı olarak sınıflandırıldığı çok sayıda taksonomi bulunmaktadır. Bu aşamalı sınıflamalar, bireylere kazandırılacak özelliklerin düzeyi, bu özelliklerin kazandırılması için gerekli eğitimsel etkinlikler ile ölçme ve değerlendirme yaklaşımlarının birbirleriyle tutarlı ve etkili biçimde yürütülebilmesi açısından yol gösterici olmaktadır. Başarının izlenmesinde, hedeflenen özellikler bağlamında, öğrencilerin bilişsel düzeyin hangi basamağında olduğunun belirlenmesi, eğitim hedef ve etkinliklerinin planlanması için oldukça önemlidir. Bu bakımdan ölçme sonuçlarının, öğrencilerin hangi bilişsel düzeyde olduğunu ortaya koyması eğitim uygulayıcıları açısından bir test puanına göre daha çok daha yorumlanabilir ve kullanılabilir olabilmektedir. Uluslararası sınavlarda da test sonuçlarının puanlarla birlikte, yine öğrencilerin eriştikleri yeterlik düzeyleri bağlamında raporlanması da bu durumun bir göstergesidir (Milli Eğitim Bakanlığı, 2019).

Alanyazında farklı uygulama, teknik özellik ve desenleri olan çok sayıda bireye uyarlanmış test yaklaşımı söz konusudur. BOBUT bunlardan, madde bazlı bireye uyarlamanın olduğu temel yaklaşımdır. Bireye uyarlamanın test bazında yapıldığı testler ise çok aşamalı testler (Magis et al., 2018) olarak anılmaktadır. Çok boyutlu MTK modellerine dayalı olarak, BOBUT'un birden fazla boyut üzerinde yapılandırılması sağlanarak, çok boyutlu bireyselleştirilmiş testler (Frey & Seitz, 2009; Seo & Weiss, 2015) ortaya koyulmuştur. Ek olarak çok aşamalı testlerin ve bireye uyarlanmış testlerin çeşitli şekillerde harmanlandığı hibrit modellere ilişkin örnekler (Raborn & Sarı, 2021; Wang et al., 2016) de bulunmaktadır. Bireye uyarlamanın farklı yaklaşım ve desenlerini gördüğümüz bu örneklerde, çok aşamalı testler ile klasik BOBUT uygulamalarını, bilişsel düzeylere göre aşamalar oluşturarak birleştiren bir yaklaşıma rastlanmamıştır. Bu tür bir yaklaşımın, özellikle her bir bilişsel düzey bağlamında yanıtlayıcının yeterliğini resmetmesi, her bir bilişsel basamak içerisinde madde bazında bireye uyarlama ile BOBUT'un güçlü yönlerini kullanması hedeflenebilir. Bilişsel basamaklar temelinde oluşturulmuş alt testlerle test bazında da uyarlama yapan bu yaklaşımla, bireyin yeterlik puanı ve aşamalı sınıflamadaki yerine ilişkin kullanışlı verilerin elde edilebileceği değerlendirilmektedir.

Bu araştırmada bilişsel sınıflamalara göre aşamalandırılmış alt testlerden oluşan çok aşamalı test ile klasik BOBUT uygulamalarının psikometrik gücü birleştirerek; başarının izlenmesinde yeterlik düzeyi noktasında sınıflamalı bilgi ve aynı zamanda geçerliği yüksek ölçme sonuçları sunan bir yaklaşım oluşturularak yeterliğinin incelenmesi amaçlanmıştır. Bu yaklaşımın incelenmesinde, çok sayıda becerinin gelişimi için önkoşul olan okuduğunu anlama becerisi odağa alınmıştır. Bu temel amaç çerçevesinde aşağıdaki sorulara yanıt aranmıştır:

[1] Balıkesir Üniversitesi
[2] Ankara Üniversitesi

• Havuz maddelerinin klasik bir BOBUT oturumunda uygulanması sonucunda elde edilen yetenek parametreleri ile sınıflama düzeylerine göre oluşturulmuş alt testler temelinde oluşturulan BOBUT sonucunda edilen yetenek parametreleri arasındaki korelasyon ne düzeydedir?

• Sınıflama düzeylerine göre oluşturulmuş alt testler temelinde oluşturulan BOBUT'un performans göstergeleri (ortalama test uzunluğu, RMSE, yanlılık değerleri) nasıldır?

Yöntem

Araştırma, korelasyonel bir araştırma olarak tasarlanmıştır.

Araştırmada BOBUT uygulaması madde havuzunun deneme uygulamasını yürütmek üzere ortaokul 7. ve 8. sınıflardan yaklaşık 20 öğrenci ile ön deneme ve 500 öğrenci ile deneme uygulamaları yürütülecektir. Bunun yanında, geliştirilen BOBUT uygulamalarının araştırma amaçları doğrultusunda uygulanması için aynı sınıf düzeylerinden yaklaşık 50 öğrenci ile BOBUT uygulaması gerçekleştirilecektir.

BOBUT uygulamaları için geniş bir madde havuzu gerekliliği bulunmaktadır. Okuduğunu anlama becerisini ölçmek üzere araştırma kapsamında kullanılacak madde havuzu, daha önce araştırmacının tez çalışmasında geliştirilmiş madde havuzunun (166 madde) genişletilmesi ile edilecektir. Genişletme sürecinde yaklaşık 70 yeni madde geliştirerek, uzman görüşü, ön deneme ve deneme uygulamaları sonrasında, maddelerin Madde Tepki Kuramı'na göre kalibrasyonu sağlanacaktır. Araştırmada, okuduğunu anlama testi, Uluslararası Okuma Becerilerinde Gelişim Projesi'nde (PIRLS) kullanılan dört aşamalı kavrama süreçleri sınıflamasına göre hazırlanacak ve alt testlere ayrılacaktır (Milli Eğitim Bakanlığı, 2003):

1. Düzey: Metinde açıkça ifade edilenlerden uygun çıkarımlar yapma.

2. Düzey: Metinde açıkça ifade edilmemiş bilgi ve fikirlerden çıkarımlar yapma.

3. Düzey: Metinde geçen olayları kişisel bilgi ve deneyimlerle ilişkilendirme.

4. Düzey: Metnin ögelerini, içeriğini, dilini inceleme ve değerlendirme

BOBUT uygulama yazılımı araştırmacı tarafından geliştirilecektir. Araştırmada, aynı madde havuzunu farklı şekilde kullanan; iki ayrı BOBUT tasarımı uygulanarak, yetenek parametreleri arasındaki ilişkiler inceleneceği için, iki ayrı BOBUT tasarımı da aynı öğrencilere tek oturumda uygulanacaktır.

Verilerin analizinde, madde kalibrasyonunda R ltm paketi kullanılacaktır. BOBUT uygulaması sonucunda elde edilen yetenek parametreleri BOBUT uygulaması sonucunda, yazılım veri tabanına kaydedilecek, BOBUT performansı göstergeleri (orta¬lama test uzunluğu, RMSE, yanlılık değerleri) uygulama sonucunda kaydedilen değerlerden hesaplanacaktır.

Beklenen Bulgular

Alanyazın incelendiğinde, farklı BOBUT desenlerinde ortalama test uzunluğu, RMSE değeri gibi açılardan performans farklılıkları olsa da bireye uyarlanmış testlerin genel olarak iyi performans gösterdikleri gözlenmektedir. Bu çalışmada uygulanan desende, her bir bilişsel düzey temelinde ayrı BOBUT uygulamaları olacağı için, ortalama test uzunluğunun yüksek olması beklenebilir. Ancak, ölçme sonuçlarının, aşamalı sınıflama temelinde verilmesi gereksinimi olduğu durumda bu tür bir yaklaşımın kullanılabilirliği, geleneksel bir BOBUT uygulaması ile benzer kestirimler üretmesi ve performans göstergelerine göre değerlendirilecektir.

Kaynakça

Davey, T. (2011). A Guide to Computer Adaptive Testing Systems Written by Tim Davey, Educational Testing Service for Technical Issues in Large-Scale Assessment (TILSA) State Collaborative on Assessment and Student Standards (SCASS). November. http://files.eric.ed.gov/fulltext/ED543317.pdf

Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. Studies in Educational Evaluation, 35(2-3), 89-94. https://doi.org/10.1016/j.stueduc.2009.10.007

Magis, D., Yan, D., & von Davier, A. A. (2018). Computerized adaptive and multistage testing with R: Using packages catR and mstR. In Measurement: Interdisciplinary Research and Perspectives, 16 (4). https://doi.org/10.1080/15366367.2018.1520560

Milli Eğitim Bakanlığı. (2003). Uluslararası Okuma Becerilerinde Gelişim Projesi (PIRLS) 2001 Ulusal Raporu. Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

Milli Eğitim Bakanlığı. (2019). PISA 2018 Türkiye Ön Raporu. PISA 2019 Türkiye Ön Raporu, Eğitim Analiz ve Değerlendirme Raporları Serisi, 10, 1-97.

Raborn, A., & Sarı, H. (2021). Mixed adaptive multistage testing: A new approach. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 12(4), 358-373. https://doi.org/10.21031/epod.871014

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. Educational and Psychological Measurement, 75(6), 954-978. https://doi.org/10.1177/0013164415575147

Wang, S., Lin, H., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: from group sequential design to fully sequential design. Journal of Educational Measurement, 53(1), 45-62. http://www.jstor.org/stable/43940603

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. Journal of Methods and Measurement in the Social Sciences, 2(1), 1. https://doi.org/10.2458/jmm.v2i1.12351

Anahtar Kelimeler: bilgisayar ortamında bireye uyarlanmış test, çok aşamalı test, bilişsel düzey, okuduğunu anlama becerisi

Bilgisayar Ortamında Bireye Uyarlanmış Müziksel İşitme Testinin Uygulanabilirliğinin İncelenmesi

Emre Kucam[1] Hamide Deniz Gülleroğlu[2]

ÖZET

Özet

Bu araştırmanın amacı, güzel sanatlar liselerine/fakültelerine öğrenci, TRT, Kültür ve Turizm Bakanlığı gibi kurumlara ise sanatçı alımı amacıyla bireysel olarak uygulanabilen Müziksel İşitme Testi'nin Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) olarak uygulanabilirliğinin incelenmesidir Çalışmada, R programında bulunan gm paketiyle bir ses, iki ses, üç ses ve melodiler üretilmiştir. Ardından, indirilen bu ses dosyaları google drive'a aktarılmış ve güzel sanatlar lisesi müzik öğretmenleri ile üniversitede müzik alanında görev yapan akademisyenlerden google formlar aracılığıyla görüşler alınmıştır. Bu görüşler alınırken, uzmanlardan, hem bir hem iki hem üç ses ile melodilere yönelik "kolay", "orta güçlükte" ve "zor" kategorilerine göre sınıflama yapılması istenmiş ve ek önerilerini de ifade etmeleri talep edilmiştir. Araştırma sonucunda; Müziksel İşitme Testi 'nin BOBUT olarak uygulanabilir olması, yetenek düzeyi yüksek olan katılımcıların süreçte daha hızlı belirlenebilmesi ve ÖSYM, MEB, TRT gibi kurumların ön eleme amacıyla bireye uyarlanmış müziksel işitme testini uygulamalarında kullanmaları beklenmektedir.

Giriş

Endüstride "kullanılabilirlik" kavramı, insanların, bir ürünün üretim sürecinde yer alan araçların nasıl kullanılacağına değil, o araçla yapacakları işi en kolay şekilde nasıl bitireceklerine odaklanmalarını ifade etmektedir (Çağıltay, 2016). Endüstride olduğu gibi eğitim teknolojilerinde de "kullanılabilirlik" kavramı oldukça önemlidir. Özellikle okullarda, tahtaya yazılan yazılı sorularından başlayıp, soruların fotokopiyle çoğaltılmasıyla devam eden süreç yerini önce daha kullanılabilir olan ve bilgisayarda hazırlanıp, çıktı alınan testlere, sonrasında da çok daha kullanılabilir olan bilgisayarla üretilen ve bilgisayarda uygulanan sorulara bırakmıştır. Türkiye'de Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM) ve Millî Eğitim Bakanlığı (MEB); e-ALES, e-YDS ve Motorlu Taşıt Sürücü Kursiyerleri e-Sınavı gibi bilgisayar ortamında testler uygulasa da, bu testlerin kâğıt–kalem testlerinden tek farkı, sonucun çok kısa sürede ve bilgisayar tarafından belirlenmesidir. Üstelik bu testlerde katılımcıların yalnızca matematik, fizik, kimya, tarih, edebiyat, yabancı dil v.b. alanlarda çeşitli bilişsel süreç basamaklarını yoklayan soruları belli bir ölçüde cevaplandırmaları beklenmektedir. Oysaki, belli ölçüde yanıtlanması gereken miktarda soru çözemeyen, ancak görsel/müzikal/bedensel yeteneklere sahip olan öğrencilerin ilgili alanlara yapacağı olası katkıların da çok önemli olduğu ve bu alanlarda yapılacak değerlendirmelerin, yetenek belirlemede hız ve maliyet bakımından büyük önem arz ettiği düşünülmektedir. Ayrıca, görsel/müzikal/bedensel yetenekler, psikomotor özelliklerinden dolayı bilişsel özelliklere göre çok daha keskin, gözlenebilir ve farkedilebilir olduğundan, herkese aynı lineer testin uygulanması durumunda bazı maddeler, yüksek yetenek düzeyindeki bireyler için bir egzersiz niteliğinde olacaktır. Buradan hareketle, bu araştırmada, güzel sanatlar liselerine/fakültelerine öğrenci, TRT ile Kültür ve Turizm Bakanlığı gibi kurumlara sanatçı alımı için yapılan bireysel test uygulamalarına katılımın çok yaygın ve tüm Türkiye çapında olduğu düşünüldüğünde, bu kadar ilgi gören sınavlarda yetenek belirlemenin bireye özgü yapılabilmesi, yetenek düzeyi yüksek olan

[1] TOBB ETÜ Koleji Özel Laboratuvar İlkokulu
[2] Ankara Üniversitesi

katılımcıların ise süreçte daha hızlı belirlenebilmesi için bireye uyarlanmış müziksel işitme testinin geliştirilmesi amaçlanmaktadır.

Yöntem

Bu çalışmada, müziksel işitme testinin BOBUT olarak uygulanabilirliğinin sağlanması, yani teorik bilgilerle yeni bir ürün ortaya çıkarılması hedeflenmiştir. Bu yönüyle araştırmanın uygulamalı araştırma türünde olduğu söylenebilir. Uygulamalı araştırma, var olan bir sorunu çözme, bir durumu daha iyi hâle getirme ve geliştirme amacı taşımakta, araştırmacı problem çözümünde belirli bir hedefe yönelmekte ve problemleri fiili olarak çözmesi gerekmektedir (Karasar, 2007). Araştırmanın evrenini Ankara ilinde bulunan ve müzik eğitimi veren devlet ve özel güzel sanatlar liseleri oluşturmaktadır. Ankara İl Millî Eğitim Müdürlüğü'nün internet sitesinde yayınladığı 2021 yılı istatistiklerine göre; Ankara'da bulunan 3 devlet ve 1 özel olmak üzere 4 güzel sanatlar lisesinin müzik ile ilgili bölümlerinde toplamda 862 öğrenci bulunmaktadır. MTK varsayımlarının incelenerek madde parametrelerinin belirlenmesi geniş örneklem gerektirmektedir (Hambleton, 1990). Bu çalışma özelinde ulaşılabilecek öğrenci sayısı bu gerekliliği karşılayabilecek niceliktedir. MTK'nın, madde parametrelerinin grup yetenek düzeyinden bağımsız olması özelliği de, uygulamanın yapıldığı örneklemden farklı bir grup üzerinde deneme uygulamasının yapılmasına olanak sağlamaktadır. Kalibrasyon amacıyla bu öğrencilerden bir bölümüne MEB'den alınan izinler doğrultusunda bir deneme uygulaması yapılacaktır. Araştırmalar kalibrasyonun, bilgisayar ortamında bireye uyarlanmış testlerin bankalarındaki maddelerin güncellenmesi, yeni maddelerin eklenmesi ve çıkarılmasının, yetenek puanları kestirimine doğrudan etkisi olduğunu göstermektedir (Glas, 2000; Levine & Williams, 1998; Samejima, 2000). Buradan hareketle, madde parametreleri elde edilecek ve kestirilen teta düzeyleri arasındaki korelasyonlar hesaplanacaktır. Verilerin hangi MTK modeline daha uygun olduğu tespit edilerek, BOBUT uygulaması yapılacaktır.

Beklenen Bulgular

COVID-19 salgını ile birlikte yüz yüze eğitime ara veren birçok ülke; uzaktan eğitim, uzaktan öğretim ya da uzaktan acil öğretim adlarıyla anılabilen web tabanlı bir sisteme geçmiştir. Tüm ders içerikleri öğrencilere ya video ya da online olarak verilmiş, ancak bu içeriklerin ölçme ve değerlendirme süreci sistematik olarak yapılandırılmamıştır. Dolayısıyla; öğretmenler, kendi sınırlılıklarında ölçme ve değerlendirme uygulamaları yapabilmişlerdir. Özellikle psikomotor becerilerin ön planda olduğu derslerde, sadece öğrencilerin derse katılımının bile tam puan almasına yeterli koşul olarak belirlendiği durumlara rastlanmıştır. Özel olarak güzel sanatlar müzik bölümlerinde verilen seslerin hava yoluyla bulaş durumuna yol açmaması için öğrenci alımları online olarak yapılandırılmaya çalışılmış ancak teknik aksaklıklar değerlendirme adına birçok sınırlılığı beraberinde getirmiştir. Böyle bir ortamda bu sınırlılıkları ortadan kaldırmak adına BOBUT uygulamalarının kullanılmasının büyük kolaylık sağlayacağı çok açıktır. Çalışmanın sonucunda Müziksel İşitme Testi'nin BOBUT olarak uygulanabilir olacağı, yetenek düzeyi yüksek olan katılımcıların süreçte daha hızlı belirlenebileceği ve ÖSYM, MEB, TRT gibi kurumların ön eleme sınavlarında, bireye uyarlanmış bu müziksel işitme testi uygulanmasının kullanılabileceği düşünülmektedir. Ayrıca, bu uygulama ile, bu tür sınavlarda oluşturulan jürilerin, yalnızca belli bir yetenek düzeyi üzerindeki bireylerle karşılaşarak, onlar arasından seçim yapmalarının sağlanabileceği ve jürinin iş yükünün azaltılabileceği düşünülmektedir.

Kaynakça

Çağıltay, K. ve Hedefleri, B. (2016). İnsan bilgisayar etkileşimi ve eğitim teknolojileri. Öğretim Teknolojilerinin Temelleri: Teoriler Araştırmalar Eğilimler (2 Baskı., ss. 297-314). Ankara: Pegem Akademi.

Glas, C. A. W. (2000). Item calibration and parameter drift. In W. J. van der Linden & C. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 183- 199). Norwell, MA: Kluwer Academic.

Karasar, N. (2007). Araştırma Yöntemi. Ankara: Nobel Yayın Dağıtım.

Levine, M. V. Williams B. A. (1998). Development and evaluation of online calibration procedures. (Algorithm Design and Measurement Services, Inc. TCN # 96-216). Champaign, IL.

Samejima F (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. Psychometrika, 65(3), 319-335. doi:10.1007/BF02296149

Anahtar Kelimeler: BOBUT, müziksel, işitme, uyarlanmış, test

Bireyselleştirilmiş Çok Aşamalı Test Uygulamalarında Madde Ön Bilgisi Kaynaklı Test Hilesini Belirlemede Kullback-Leibler ve Jensen-Shannon Uzaklık Ölçülerinin Performanslarının Karşılaştırılması

Celal Deha Doğan[1] Arzu Uçar[2] Ebru Balta[3]

ÖZET

Giriş

Eğitimsel ve psikolojik testler ve test sahtekarlığı ile ilgili anormal test davranışları, test puanlarının geçerliliğine zarar vermekte ve anormal test davranışları; cevap kopyalama,iş birliği, madde ön bilgisi, yaratıcı düşünme, şansla tahmin, rastgele yanıt verme ve test tahrifatı olarak sınıflandırılabilmektedir (Cizek ve Wollack, 2017; Haberman ve Lee, 2017; Karabatsos, 2003; Kingston ve Clark, 2014; Lee ve Haberman, 2016; Sinharay, 2017b; Sinharay, 2020; van der Linden ve Guo, 2008; van Krimpen-Stoop ve Meijer, 2001) .Küresel çapta görülen COVID 19 pandemi dönemi ile birlikte bireylerin hayatlarında önemli kararların alınmasında önemli yer tutan, geniş ölçekli ve yüksek riskli sınavların uygulanması ve uzaktan eğitim sürecinde, ölçme ve değerlendirme uygulamalarında kullanılan çevrim içi sınav uygulamalarının gerekliliği günümüzde giderek hız kazanmaktadır. Kullanımı yaygınlaşan çevrim içi bilgisayar temelli, bilgisayar ortamında bireye uyarlanmış test ve testin aşamalar şeklinde oluşturulduğu algoritmaya dayalı bir test yaklaşımı olan çok aşamalı test uygulamalarında, test puanlarının geçerlik ve güvenirliğine ilişkin daha fazla kanıt sunulmasının, test güvenliğinin sağlanmasına yönelik hile belirleme çalışmalarının gerçekleştirilmesi gerekliliği görülmektedir. Bu nedenle, mevcut çalışmada, BÇAT uygulamalarında, madde ön bilgisi kaynaklı test hilelerini tespit etmek için, Kullback-Leibler ve Jensen-Shannon Uzaklık ölçülerinin çeşitli koşullar altında performanslarının belirlenmesi amaçlanmaktadır.

Araştırma Yöntemi

Araştırmada kullanılan veriler, Monte-Carlo simülasyon yöntemi ile elde edilmiştir. BÇAT simülasyonunda; maksimum madde kullanım oranı, test birleştirme yöntemi, MTK modeli, modül uzunluğu, panel deseni, yetenek kestirim yöntemi, madde/modül seçim yöntemi ve örneklem büyüklüğü sabit tutulmuştur. Madde havuzu oluşturulurken, 3 parametreli lojistik model (3 PLM) temel alınarak 600 maddeye ilişkin parametre üretimi yapılmıştır. Bu aşamada, üç farklı güçlük düzeyi için farklı güçlük parametre dağılımı kullanılmış, ayırıcılık ve şans parametrelerinin dağılımları ise tüm güçlük düzeyleri boyunca aynı parametre dağılımı kullanılarak üretilmiştir. Her bir düzeyde 200'er madde bulunması sağlanmıştır. N(0, 1)'den alınan yetenek düzeylerinde 10,000 dürüst birey simüle edilmiştir. Maksimum madde kullanım oranı BÇAT için 0.25 olarak sabitlenmiş ve yetenek kestirimi için EAP yöntemi (önsel dağılım N(0,1)), madde/modül seçim yöntemi olarak Maksimum Fisher Bilgisi yöntemi seçilmiştir. Araştırmada, birinci aşamada orta güçlük düzeyindeki Modül-1 (O), ikinci aşamada kolay güçlük düzeyindeki Modül-2 (K), orta güçlük düzeyindeki Modül-2 (O) ve zor güçlük düzeyindeki Modül-2 (Z), üçüncü aşamada kolay güçlük düzeyindeki Modül-3 (K), orta güçlük düzeyindeki Modül-3 (O) ve zor güçlük düzeyindeki Modül-3 (Z) bulunan 1-3-3 panel deseni kullanılmıştır. Modül uzunluğunun 12 olduğu "1-3-3" panel deseninde birey, testin sonunda, toplam 36 maddeyi cevaplamıştır. Aşağıdan yukarıya BÇAT yapısı oluşturularak hedef Test Bilgi Fonksiyonu

[1] Ankara Üniversitesi
[2] Hakkari Üniversitesi
[3] Ağrı İbrahim Çeçen Üniversitesi

(TBF) değeri Ortalama Maksimum Bilgi (OMB) yönlendirme stratejisi ile belirlenmiştir.Veri üretimi ve analizleri için R programlama dili kullanılmıştır. Araştırmanın amacı doğrultusunda, uzaklık ölçülerinin, I. Tip hata oranı ortalamalarının elde edilmesi için, teşhir olan maddelerin güçlük düzeyi (orta güçlük, zor) koşulu altında 30 yineleme ile 60 veri seti oluşturulmuştur. Modül 1'de kullanılan maddeler, bireye uyarlanmış testin madde havuzundan, ikinci ve üçüncü aşamadaki modüllerde kullanılacak maddeler ise teşhir olmayan ve teşhir olan maddelerin yer aldığı madde havuzundan çekilmiştir. Teşhir olan maddelerdeki veri manipülasyonu ise ikinci aşamada, üçüncü aşamada ve hem iki hem de üçüncü aşamada olmak üzere üç farklı senaryo için gerçekleştirilmiştir. Her hile senaryosu için yöntemlerin, güç oranı ortalamalarının elde edilmesi amacıyla teşhir olan maddelerin güçlük düzeyi ve teşhir olan maddelerin oranı (%10, %40 ve %80) koşulları altında 30 yineleme ile 180 veri seti olmak üzere toplamda 540 veri seti oluşturulmuştur. Senaryoda, hile yapan bireyler 10,000 örneklem büyüklüğünün %5 olacak şekilde düşük yetenek düzeyindeki bireylerden seçilmiştir. Madde ön bilgisine sahip olan bireylerin 2. ve 3. panelde orta güçlük düzeyi ve zor maddelere ilişkin tepkileri doğru olarak değiştirilmiştir. Bireylerin yetenekleri veri manipülasyonu öncesinde ve sonrasında kestirilmiştir. Yetenek kestirimleri kullanılarak yöntemlerin , test hilesi belirleme performansları incelenmiştir.

Beklenen/Geçici Bulgular

Araştırma sonucunda, BÇAT uygulamasında, madde ön bilgisinden kaynaklı test hilesinin belirlenmesinde, teşhir olan maddelerin güçlük düzeyi ve oranı koşulları altında, uygun olmayan madde tepki örüntülerinin tespit edilmesinde, Jensen-Shannon uzaklık ölçüsünün Kullback-Leibler uzaklık ölçüsüne göre daha yüksek performans göstermesi beklenmektedir. Bununla birlikte, teşhir olan maddelerin güçlük düzeyi ve oranı arttıkça yöntemlerin performanslarının artması beklenmektedir.

Kaynakça

Cizek, G. ve Wollack, J. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. G. Cizek ve J. Wollack (Ed.), Handbook of Quantitative Methods for Detecting Cheating on Tests, (2.Baskı, s.217-233) içinde. New York: Routledge.

Drost, Hajk-Georg. ve Nowosad, Jakub.(2022). 'philentropy: Similarity and Distance Quantification Between Probability Functions.' R package version 0.7.0.
https://cran.r-project.org/web/packages/philentropy/philentropy.pdf

Haberman, S. ve Lee, Y. (2017). A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses. (Research Report No: RR-17-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12150.

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

Kingston, N. ve Clark, A. (2014). Test fraud: Statistical detection and methodology. New York, NY: Routledge.

Lee, Y. ve Haberman, S. (2016). Investigating test-taking behaviors using timing and process data. International Journal of Testing, 16(3), 240–267. https://doi.org/10.1080/ 15305058.2015.108538

Qian, H., Staniewska, D., Reckase, M. ve Woo, A. (2016). Using response time to detect item preknowledge in computer based licensure examinations. Educational Measurement: Issues and Practice, 35, 38–47. https://doi.org/ 10.1111/emip.12102

Raton-Lopez, M., Rodriquez-Alvarez, X. M., Suarez- Cadarso, C. ve Sampedro-Gude, F. (2014). 'OptimalCutpoints: Computing optimal cutpoints in diagnostic tests.' R package version 1.1-5. https://cran.r-project.org/web/packages/OptimalCutpoints/OptimalCutpoints.pdf

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge wh enthe set of compromised items is known? Applied Psychological Measurement, 41(6), 403–421. https://doi.org/ 10.1177/0146621617698453

Sinharay,S. (2020).Detection of item preknowledge using response times. Applied Psychological Measurement, 44(5), 1-17. https://doi.org/10.1177/0146621620909893

Ucar, A. ve Dogan, C.D. (2021). Defining cut point for Kullback-Leibler divergence to detect answer copying. International Journal of Assessment Tools in Education, 8(1), 156-166. https://doi.org/10.21449/ijate.864078

van der Linden, W. J. ve Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. Psychometrika, 73, 365–384. https://doi.org/10.1007/s11336-007-9046-8.

van Krimpen-Stoop, E. M. L. A. ve Meijer, R. R (2001). CUSUM-based person-fit statistics for adaptive testing. Journal of Educational and Behavioral Statistics, 26(2), 199–217. https://doi.org/ 10.3102/10769986026002199

Anahtar Kelimeler: bireyselleştirilmiş çok aşamalı test, , kullback-leibler, jensen-shannon, test güvenliği

Bireyselleştirilmiş Test Uygulamalarında Anormal Tepki Örüntü Benzerliğinin M4 Benzerlik indeksi ve Jensen-Shannon Uzaklık Ölçüsü Kullanılarak İncelenmesi

Önder Sünbül[1] Ebru Balta[2] Arzu Uçar[3]

ÖZET

Giriş

Uygun olmayan tepki örüntülerinin ortaya çıkmasına sebep olan madde ön bilgisi, sınav güvenlik önlemlerinin ihlal edilerek sınav uygulaması öncesinde, madde havuzundaki bazı sınav maddelerine yasa dışı erişilmesi ve böylelikle, maddelerin ezberlenmesi sonucunda oluşmaktadır. Madde ön bilgisine sahip bireylerin, teşhir olan maddelere doğru cevap verebilecek yeterlik düzeyine sahip olmadıkları durumda maddeyi doğru cevapladığı ve böylelikle, test puanlarının geçerliğinin zedelendiği görülmektedir (Man, Harring, Quyang ve Thomas ,2018). . Kullanımı yaygınlaşan bireyselleştirilmiş test uygulamalarında, test puanlarının geçerliğinin tespiti için, madde ön bilgisinden kaynaklı uygun olmayan tepki örüntülerinin tespit edilmesinin önemli olduğu görülmektedir. Bu nedenle, mevcut çalışmada, madde ön bilgisine sahip olan bireylerin tepki örüntü benzerliğini belirlemede, M4 benzerlik indeksi ve Jensen-Shannon Uzaklık Ölçüsünün performanslarının belirlenmesi amaçlanmıştır.

Araştırma Yöntemi

Araştırmada kullanılan veriler, Monte-Carlo simülasyon yöntemi ile elde edilmiştir. BOBUT simülasyon çalışması, Hukuk Fakültesine Kabul Testinin (LSAT) açıklanan Okuduğunu Anlama (RC), Analitik Akıl Yürütme (AR) ve Mantıksal Akıl Yürütme (LR) maddeleri kullanılarak gerçekleştirilmiştir. BOBUT madde havuzu 1500 maddeden oluşmaktadır. N(0, 1)'den alınan yetenek düzeylerinde 10,000 dürüst birey simüle edilmiştir. Veri üretimi ve analizleri için R programlama dilinden faydalanılmıştır. Araştırmanın amacı doğrultusunda, yöntemlerin, I. Tip hata oranı ortalamalarının elde edilmesi için, veri setleri, madde tepkilerinin üç parametreli lojistik model (3PLM) ile modellendiği, teşhir olan maddelerin güçlük düzeyi (orta güçlük,zor) koşulları altında 50 yineleme ile 100 veri seti oluşturulmuştur. Yöntemlerin, güç oranı ortalamalarının elde edilmesi için ise teşhir olan maddelerin güçlük düzeyi, teşhir olan maddelerin oranı (%10, %40 ve %80) ve hile yapan bireylerin oranı (%5, %10, %20) koşulları altında 50 yineleme ile 900 veri seti oluşturulmuştur. Verilerin analizinde, hile senaryosunda, hile yapan bireyler düşük yetenek düzeyindeki bireylerden seçilerek oranı, 10,000 örneklem büyüklüğünün %5, %10 ve %20 ‘si olacak şekilde hileli veriler oluşturularak yöntemlerin , test hilesi belirleme performansları incelenmiştir. Madde ön bilgisine sahip olan bireylerin, orta güçlük düzeyi ve zor maddelere ilişkin tepkileri doğru olarak değiştirilmiştir. Bireylerin yetenekleri veri manipülasyonu öncesinde ve sonrasında kestirilmiştir. Yetenek kestirimleri kullanılarak, yöntemlerin , test hilesi belirleme performansları incelenmiştir.

Beklenen/Geçici Bulgular

Bireyselleştirilmiş test uygulamasında, madde ön bilgisinden kaynaklı test hilesinin belirlenmesinde, uygun olmayan tepki örüntülerinin tespit edilmesinde, teşhir olan maddelerin güçlük düzeyi, teşhir

[1] Mersin Üniversitesi
[2] Ağrı İbrahim Çeçen Üniversitesi
[3] Hakkari Üniversitesi

olan maddelerin oranı ve hile yapan birey oranı değişimlemesi koşulları altında, M4 benzerlik indeksinin Jensen-Shannon uzaklık ölçüsüne göre daha yüksek güç oranı ortalama değerlerine ve I. Tip hata oranı ortalamaları değerlerine sahip olabileceği düşünülmektedir. M4 benzerlik indeksinin en yüksek güç oranı ortalama değerlerinin teşhir olan maddelerin zor olduğu, teşhir olan maddelerin oranının orta düzeyde olduğu ve hile yapan birey oranının yüksek olduğu durumda gözlenmesi beklenmektedir.

Kaynakça

Cizek, G. ve Wollack, J. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. G. Cizek ve J. Wollack (Ed.), Handbook of Quantitative Methods for Detecting Cheating on Tests, (2.Baskı, s.217-233) içinde. New York:Routledge.

Drost, Hajk-Georg. ve Nowosad, Jakub.(2022). 'philentropy: Similarity and Distance Quantification Between Probability Functions.' R package version 0.7.0.
https://cran.r-project.org/web/packages/philentropy/philentropy.pdf

Haberman, S. ve Lee, Y. (2017). A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses. (Research Report No: RR-17-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12150.

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

Kingston, N. ve Clark, A. (2014). Test fraud: Statistical detection and methodology. New York, NY: Routledge.

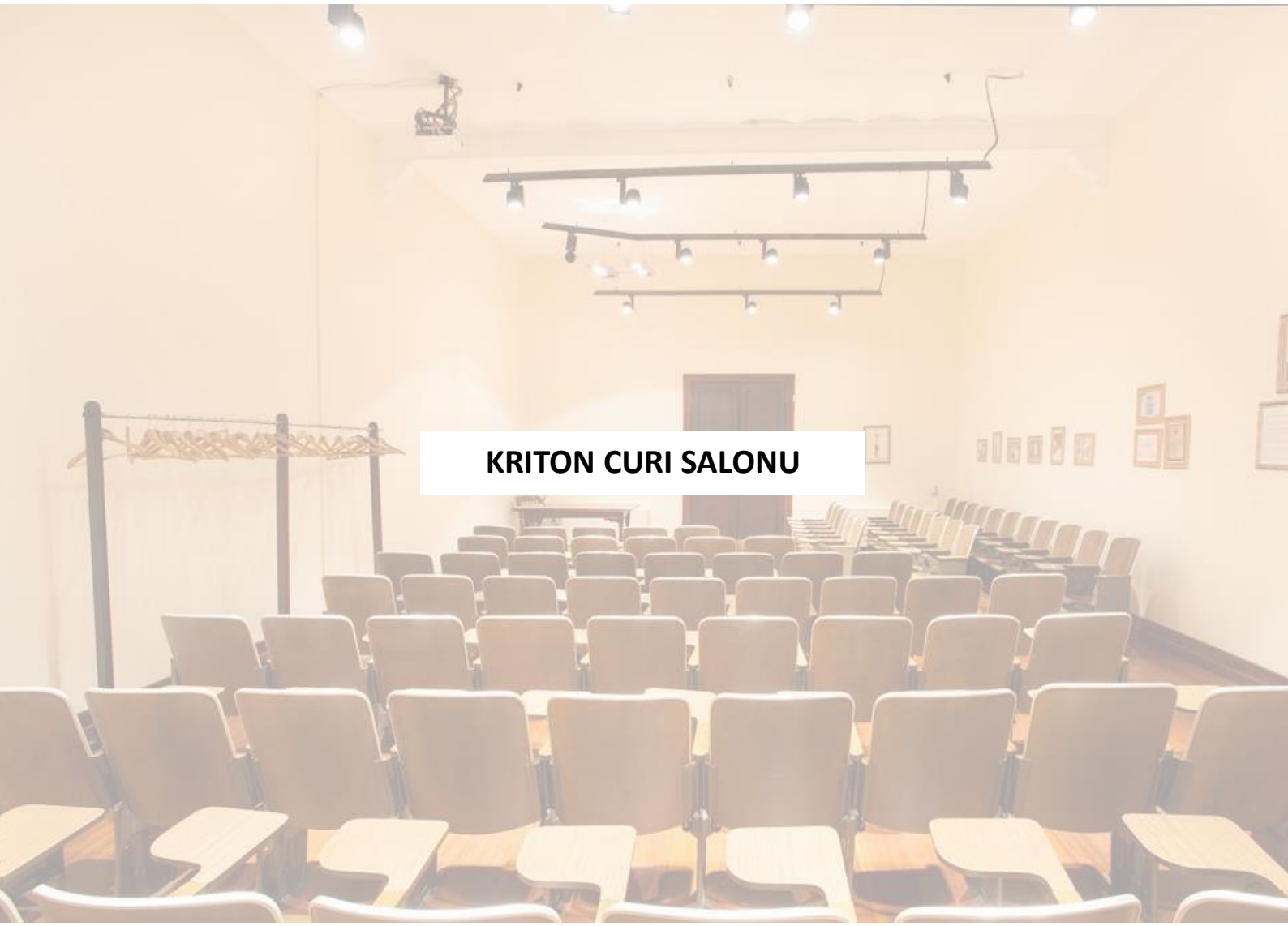
Man, K., Harring. J., Quayang.Y. ve Thomas, S. (2018). Response time based nonparametric Kullback-Leibler divergence measure for detecting aberrant test taking behavior. International Journal of Testing, 18(2), 155-177. https://doi.org/ 10.1080/15305058.2018.1429446

Maynes, D. (2017). Detecting potential collusion among individual examinees using similarity analysis. In GJ Cizek and JA Wollack (Eds.), Handbook of Quantitative Methods for Detecting Cheating on Tests (pp. 47-69). Routledge, New York, NY

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? Applied Psychological Measurement, 41(6), 403–421. https://doi.org/ 10.1177/0146621617698453

van Krimpen-Stoop, E. M. L. A. ve Meijer, R. R (2001). CUSUM-based person-fit statistics for adaptive testing. Journal of Educational and Behavioral Statistics, 26(2), 199–217.
https://doi.org/ 10.3102/10769986026002199

Anahtar Kelimeler: bilgisayarlı bireyselleştirilmiş test, M4 benzerlik indeksi, Jensen-Shannon ,test güvenliği


KRITON CURI SALONU

Kağıt Kalem ve Bilgisayar Ortamında Bireye Uyarlanmış Testlerinin Karşılaştırılması

İbrahim Hakkı Tezci[1] Bayram Bıçak[1]

ÖZET

Giriş

Bilgisayar ortamında bireye uyarlanmış testler sahip olduğu birçok özellik ve avantaj sebebiyle ölçme ve değerlendirme alanında önemli bir yer edinmeye devam etmektedir. Özellikle çevrim içi öğrenme ve ölçme ve değerlendirme alanlarında yaşanan gelişmeler bu alanda çeşitli uygulamaların değişimine ve gelişimine hız kazandırmıştır. Bilgisayar ortamında bireye uyarlanmış testler, kişinin bilişsel ve davranışsal özelliklerini ölçmek için kullanılan etkili bir yöntemdir (Hambleton Swaminathan, 1989). Bu testler, kişinin performansını belirlemek için standartlaştırılmış ve önceden hazırlanmış sorular içermektedir. Testler zorluk seviyesinin kişinin yetenek ve becerilerine göre otomatik olarak ayarlandığı, zaman kaydı yapıldığı ve sonuçların hızlı bir şekilde analiz edildiği bir bilgisayar ortamında gerçekleştirilir. Bu testlerin avantajları arasında ölçüm doğruluğu, tekrarlanabilirlik, nesnellik ve testlerin daha hızlı ve kolay bir şekilde yapılabilmesi söylenebilir (Demirtaşlı, 1999; Linacre, 2000; Lord ve Stocking, 1988). Bireye uyarlanmış testler ayrıca, test sonuçlarının dijital bir formatta saklanabilmesi sayesinde, kişinin ilerlemesini izlemek ve sonuçları takip etmek için daha kolay bir yol sunar. Bu testler, öğrenme engelleri gibi özellikleri olan kişilerin ihtiyaçlarını daha iyi anlamak ve onlara uygun öğrenme ortamları sağlamak için de kullanılabilir. Bireye uyarlanmış testler, herkesin sabit sayıda soru cevapladığı, farklı cevap örüntülerini ihmal eden geleneksel kâğıt-kalem testlerine göre daha hızlı, daha doğru ve daha verimlidir (Rudner, 1998). Sonuç olarak, bireye uyarlanmış testler, kişinin performansını doğru bir şekilde ölçerek, bireyselleştirilmiş eğitim veya tedavi planlarının oluşturulmasına yardımcı olur ve sonuçların daha hızlı ve kolay bir şekilde analiz edilmesine olanak tanır. Ancak sahip olduğu avantajlarının yanı sıra bazı uygulama zorluklarının bulunması, bu konudaki çalışmaların yeterli düzeye ulaşmaması sebebiyle kâğıt kalem testleriyle olan farklılıklarının incelenmesi yapısal ve kestirimsel sonuçlarının değerlendirilmesi önemli görülmektedir.

Yöntem

Bu çalışmada kâğıt kalem testleriyle bilgisayar ortamında bireye uyarlanmış testlerin karşılaştırılması amaçlanmaktadır. İki farklı ölçme ve değerlendirme biçiminin karşılaştırılması dolayısıyla birbiriyle olan ilişkileri ve üstünlüklerinin incelenmesi sebebiyle araştırmanın modelinin bir temel araştırma olduğu söylenebilir. Araştırmanın amacına uygun olarak ortaokul 8. Sınıf öğrencilerinin Fen Bilgilerini ölçme amacıyla Fen Bilgisi başarı testi oluşturulacaktır. Söz konusu madde havuzunun çeşitli yetenek düzeylerinde yer alan bireylerin performansının ölçebilmesi amacıyla geniş tutulması amaçlanmaktadır. Geliştirilecek başarı testinin psikometrik niteliklerinin belirlenmesinde madde tepki kuramı, klasik test kuramı ve faktör analizi yöntemleri kullanılarak incelenecektir. Özellikle bilgisayar ortamında bireye uyarlanmış testlerin yapılarının madde tepki kuramı modeliyle örtüşmesi sebebiyle bu ölçme aracı geliştirme sürecinde boyutluluk durumunun analizler üzerine olan etkileri incelecektir. Ölçme aracının geliştirilmesi süreci sonrasında test formu öğrencilere belirli zaman aralığıyla geleneksel kâğıt kalem ve bilgisayar ortamı aracılıyla sunulacaktır. Bilgisayar ortamında bireye uyarlanmış testler R paketleri kullanılarak Concerto platformu aracılığıyla gerçekleştirilecektir. Burada öğrencilerin her iki ölçme yönteminde gösterdikleri performansları arasında ilişki olup olmadığı, aldıkları madde sayılarının dağılımı ve güvenirlik durumları incelenecektir. Bir sonraki aşamada farklı

[1] Akdeniz Üniversitesi

CAT koşulları olan başlatma, sonlandırma, madde seçim ve madde kestirim yöntemlerinin yetenek parametrelerine ve madde sayısına olan etkileri incelenecektir. Başlatma kuralları olarak öğrencinin not ortalaması ve belirli bir yetenek aralığı, sonlandırma kuralı olarak standart hata ve theta değişimi, madde seçim yöntemi olarak geleneksel ve bayesci yöntemler ve son olarak yetenek kestirim yöntemi olarak ML, EAP ve MAP yöntemleri kullanılacaktır. Bir sonraki aşamada ise CAT ve kâğıt kalem test yöntemlerinin maddelerde değişen madde fonksiyonuna sebep olup olmadığı irdelenecektir. Maddeler farklı DMF belirleme yöntemleri aracılığıyla incelerek öncelikle DMF içerip içermediği eğer içeriyorsa uygulama koşullarının bir fark yaratıp yaratmadığı üzerine incelemelerde bulunulacaktır. Araştırmanın son aşamasında öğrencilerin araştırmada elde edilen yetenek düzeyleriyle LGS kapsamında girdikleri Fen Bilgisi puanları arasında bir ilişki içerip içermediği, bu ilişkinin miktarının kâğıt kalem ya da bilgisayar ortamında bireye uyarlanmış teste göre farklılaşıp farklılaşmadığı incelenecektir.

Beklenen Bulgular

Araştırmanın amacı ve alt problemlerine uygun olarak yapılacak veri analizi sonrasında

1.) Fen Bilgisi başarı testinin kâğıt-kalem ve bilgisayar ortamında uygulanması sonucu elde edilen yetenek düzeyleri arasında ilişkinin ne olduğu ve hangi yöntemde yeteneklerin daha az ya da fazla kestirildiği

2.) Fen Bilgisi başarı testinin kâğıt-kalem ve bilgisayar ortamında uygulanması sonucu elde edilen güvenirlik düzeylerinin ne olduğu

3.) Fen Bilgisi başarı testinin bilgisayar ortamında uygulanması sonucu öğrencilerin aldıkları soru sayılarının dağılımının ne olduğu, yetenek düzeyiyle bir ilişki içerip içermediği

4.) Fen Bilgisi başarı testinin kâğıt-kalem ve bilgisayar ortamında uygulanması sonucu elde edilen yetenek düzeyleriyle LGS Fen Bilgisi testi arasında ilişki olup olmadığı eğer varsa bu ilişkinin miktarının ne olduğu

5.) Farklı CAT koşullarının (başlatma-sonlandırma-madde seçim-yetene kestirim) ve bunlarının varyasyonlarının öğrencilerin aldıkları madde sayısında ve kestirilen yetenek düzeyinde bir farklılaşmaya sebep olup olmadığı

6.) Bilgisayar ortamında bireye uyarlanmış ve kâğıt-kalem test sonuçlarının maddelerde değişen madde fonksiyonuna sebep olup olmadığı konularına cevap bulunacağı düşünülmektedir.

Kaynakça

Demirtaşlı, N. (1999). Psikometride yeni ufuklar: Bilgisayar ortamında bireye uyarlanmış test. Türk Psikologlar Derneği. http://www.psikolog.org.tr/articles_detail.asp?cat=2&id=25.

Hambleton, R.K., and Swaminathan, H. (1989). Item response teory: Principles and applications. USA: Kluwer Nijhoff Publishing.

Linacre, J.M. (2000). Computer adaptive testing: A methodology whose time has come. Chae, S., Kang, U., Jeon, E., and Linacre, J.M. (Ed.). Devolopment of computerized middle scholl achievement test. Seoul: Komesa Press.

Lord, F.M., and Stocking, M.L. (1988). Item response theory. J.P. Keeves (Ed.). Educational research, methodology, and measurement: An international handbook . New York: Pergamon Press.

Rudner, L. M. (1998). An on-line, Interactive computer adaptive testing mini tutorial. http://edres.org/scripts/cat/catdemo.htm

Anahtar Kelimeler: kağıt kalem testi, bobut, gerçek veri, değişen madde fonksiyonu, bireye uyarlanmış test

Ders ve Öğretim Elemanı Değerlendirme Formu Bilgisayar Ortamında Bireyselleştirilmiş Test Olarak Uygulanabilir mi?

Semih Topuz[1] Giray Berberoğlu[1] Esra Kınay Çiçek[1] Kadriye Belgin Demirus[1]

ÖZET

GİRİŞ

Pek çok üniversite ders ve öğretim elemanının yeterliğini değerlendirmek üzere öğrencilerden belli sayıda soru içeren anketleri yanıtlamalarını istemektedir. Üniversite yönetimleri bu değerlendirme sonuçlarına dayanarak öğretim elemanları ve derslerle ilgili önemli kararlar alabilmektedir. Bu nedenle anketlere verilen öğrenci yanıtlarının geçerliği oldukça önemlidir. Literatürde bu tür değerlendirmelerin geçerliği sıklıkla tartışılmaktadır. Sınıf büyüklüğü, öğretim elemanının yaşı, deneyimi, akademik unvanı, dersin iş yükü, öğretim elemanının not vermedeki esnekliği, öğretim elemanı ile daha önceki deneyimler, öğrencinin dersten aldığı not gibi faktörlerin değerlendirmelere yansıdığı belirtilmektedir (d'Apollonia & Abrami, 1996, akt. Algozzine vd., 2004; Greenwald & Gillmore, 1997; Griffin, 2004; Hoffman, 1978; Liaw & Goh, 2003; Marsh, 1982, 1983; Zhao & Gallant, 2012). Bunun yanında öğrencilerin anketleri yanıtlamadaki motivasyon eksiklikleri ya da dikkatsiz değerlendirmeleri de geçerliği etkileyen önemli faktörler olarak belirtilmektedir (Bassett vd., 2017). Öğrencilerin anketleri yanıtlamak istememelerindeki en temel neden öğretim sürecinde alınan ders sayısına bağlı olarak yanıtlanması gereken anket sayısının ve dolayısıyla anketleri yanıtlamak için harcanması gereken zamanın fazla olmasıdır (Hoel & Dahl, 2019). Bu durum da anketlerle toplanan bilginin geçerliği konusunda sorun yaratabilmektedir. İlgili değerlendirmelerin daha kısa zaman diliminde, daha az soruyla uygulanmasının özellikle geçerliği arttırabileceği ve uygulama kolaylığı getireceği düşünülmektedir.

Bu çalışmada bilgisayar ortamında bireye uyarlanmış test uygulamaları kapsamında farklı uygulama stratejileri kullanılarak daha az sayıda soru ile uzun anket formlarından elde edilen sonuçlara ne ölçüde ulaşılabildiği incelenecektir. Bu incelemede bir vakıf üniversitesinde uygulanan 16 soruluk tek boyutlu ders ve öğretim elemanı değerlendirme formundan elde edilen veri seti kullanılarak bireye uyarlanmış test uygulamaları simülasyon çalışması ile gerçekleştirilecektir.

YÖNTEM

Ölçek

Çalışmada kullanılan ölçek öğretim üyesinin dersi yürütme süreci, materyal kullanımı, konuya hakimiyeti, iletişim ve ölçme ve değerlendirme süreçleri boyutlarında öğrencilerin verilen durumları ne ölçüde yeterli algıladıklarını değerlendirmektedir. On altı sorudan oluşan ölçek verilen durumların ne ölçüde yeterli algılandığını dörtlü dereceleme ölçeği ile "Çok", "Genellikle", "Nadiren" ve "Hiç" seçenekleri ile değerlendirmeye olanak sağlamaktadır. Öğrenciler bu ölçeği her dönem aldıkları tüm dersler için yanıtlamaktadır.

Çalışma grubu

Araştırmada, bir vakıf üniversitesinde yaklaşık 12000 öğrenciden elde edilen değerlendirme sonuçları kullanılacaktır.

Verilerin analizi

[1] Başkent Üniversitesi

Veri analizleri R paket programında “psych”, “mirt” ve “catR” paketleriyle yapılacaktır (Chalmers, 2012; Magis & Barrada, 2017; R Core Team, 2023; Revelle, 2023). Bu çalışma gerçek veriye dayalı bir post hoc simülasyon araştırmasıdır.

Post Hoc simülasyon çalışması öncesi 16 sorunun faktör yapısı temel eksen faktör analizi yöntemi ile incelenmiştir. Faktör analizine ait ilk dört özdeğer ve faktörlerin açıkladığı varyans miktarları Tablo 1’de verilmiştir.

Tablo 1.

Özdeğerler

Boyut	Özdeğer	Açıklanan varyans
1	14,97	0,936
2	0,18	0,011
3	0,08	0,005
4	0,05	0,003

Tablo 1’de ilk faktördeki özdeğerin ve açıklanan varyansın oldukça yüksek, diğer faktörlerdeki özdeğerlerin birin altında ve varyans miktarlarının oldukça küçük olduğu görülmektedir. Bu analiz ilgili ölçekte kullanılan soruların tümünün ağırlıklı olarak tek boyutlu bir değişkeni ölçtüğünü göstermektedir. Bu analiz ilgili ölçekte kullanılan soruların tümünün tek boyutlu bir yapıda olduğunu göstermektedir. Çalışmada kullanılacak veri setinin ölçeklenmesi ilgili alanda kullanılan Graded Response (Samejima, 1997), Graded Ratings Scale, Generalized Partial Credit (Muraki, 1992) ve Rasch Rating Scale (Andrich, 1978) modellerinden en iyi uyum veren bir model kullanılarak gerçekleştirilecektir. Bir sonraki aşamada modelden elde edilen parametreler kullanılarak simülasyon çalışması farklı stratejilerle yürütülecektir. Bu stratejiler genel olarak alanda benzer çalışmalarda kullanılan başlangıç noktasının sabitlenmesi, hata kestiriminin sabitlenmesi, sabit sayıda soru kullanılması vs. gibi yöntemleri içerecektir. Çalışmada daha az soru ile elde edilen kestirimlerin 16 sorudan elde edilen sonuçları ne ölçüde yansıttığı üzerinde durulacaktır.

BEKLENEN BULGULAR

Bu araştırmadan beklenen bulgular, öncelikle araştırmanın geleneksel uygulamalara göre sağlayacağı bir dizi avantaj çerçevesinde açıklanabilir. Geleneksel test uygulamalarında zaman, harcanan enerji ve maliyet açısından verimlilik düşük olabilmektedir. Daha önce de söylendiği gibi bir dönemde çok sayıda ders alan öğrencilerin dönem sonunda her ders için aynı sorulardan oluşan ölçeği defalarca yanıtlaması önemli geçerlik sorunu yaratabilmekte, ayrıca bu tür bir uygulama ciddi ölçüde uzun yanıtlama süresi gerektirmektedir. Bu anlamda daha ekonomik bir uygulamaya olanak sağlaması açısından bireye uyarlanmış test yaklaşımının ilgili değerlendirmede kullanılabilirliğinin incelenmesi önem kazanmaktadır. Araştırmacıların hipotezi daha az sayıda soru ile öğrencilerin ders değerlendirme sonuçlarının çok sayıda sorudan elde edilen bilgileri kestirebileceği yönündedir. Ayrıca, bu uygulamanın bireye uyarlanmış olması, öğrencilerin aynı dönemde farklı dersleri değerlendirirken farklı soru gruplarını yanıtlamalarına da olanak sağlayacaktır. Bunun da geçerliği olumlu yönde etkileyebileceği düşünülmektedir.

KAYNAKÇA

Algozzine, B., Gretes, J., Flowers, C., Howley, L., Beattie, J., Spooner, F., Mohanty, G., & Bray, M. (2004). Student Evaluation Of College Teaching: A Practice In Search Of Principles. College Teaching, 52(4), 134-141. https://doi.org/10.3200/CTCH.52.4.134-141

Andrich, D. (1978). Application of a Psychometric Rating Model to Ordered Categories Which Are Scored with Successive Integers. Applied Psychological Measurement, 2(4), 581-594. https://doi.org/10.1177/014662167800200413

Bassett, J., Cleveland, A., Acorn, D., Nix, M., & Snyder, T. (2017). Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations. Assessment & Evaluation in Higher Education, 42(3), 431-442. https://doi.org/10.1080/02602938.2015.1119801

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

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. Journal of Educational Psychology, 89(4), 743-751. https://doi.org/10.1037/0022-0663.89.4.743

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. Contemporary Educational Psychology, 29(4), 410-425. https://doi.org/10.1016/j.cedpsych.2003.11.001

Hoel, A., & Dahl, T. I. (2019). Why bother? Student motivation to participate in student evaluations of teaching. Assessment & Evaluation in Higher Education, 44(3), 361-378. https://doi.org/10.1080/02602938.2018.1511969

Hoffman, R. G. (1978). Variables Affecting University Student Ratings of Instructor Behavior. American Educational Research Journal, 15(2), 287-299. https://doi.org/10.3102/00028312015002287

Liaw, S., & Goh, K. (2003). Evidence and control of biases in student evaluations of teaching. International Journal of Educational Management, 17(1), 37-43. https://doi.org/10.1108/09513540310456383

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

Marsh, H. W. (1982). Factors Affecting Students' Evaluations of the Same Course Taught by the Same Instructor on Different Occasions. American Educational Research Journal, 19(4), 485-497. https://doi.org/10.3102/00028312019004485

Marsh, H. W. (1983). Multitrait-Multimethod Analysis: Distinguishing between Items and Traits. Educational and Psychological Measurement, 43(2), 351-358. https://doi.org/10.1177/001316448304300204

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. ETS Research Report Series, 1992(1), i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2023). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University. https://CRAN.R-project.org/package=psych

Samejima, F. (1997). Graded Response Model. Içinde W. J. van der Linden & R. K. Hambleton (Ed.), Handbook of Modern Item Response Theory (ss. 85-100). Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_5

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. Assessment & Evaluation in Higher Education, 37(2), 227-235. https://doi.org/10.1080/02602938.2010.523819

Anahtar Kelimeler: bireye uyarlanmış test, öğretim elemanı değerlendirme formu, ders değerlendirme formu

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) Uygulamaları ile Sınıflandırma: Farklı Faktöriyel Modellere Dayalı İngilizce Düzey Belirleme Sınavının Karşılaştırmalı Analizi

Ufuk Akdemir[1] İlker Kalender[1]

ÖZET

Giriş

Eğitimdeki kararlar genellikle tanılama, seçme ve sınıflandırma ile ilgilidir (Chang, 2015), bu durum çok yönlü ve karmaşık olan dil eğitimi alanındaki ölçme ve değerlendirme çalışmaları için de geçerlidir. Bahsedilen kararlar esas olarak tanımlama, kestirim, başarı vb. çeşitli amaçlarla yapılan dil testlerinin sağladığı bilgilere dayanmaktadır.

İdeal bir testin, geçerli ve güvenilir olmasının yanı sıra, öğrencinin yetenek düzeyine de uygun olması önem taşımaktadır (Chang, 2015; Thorndike & Thorndike-Christ, 2014; Weiss, 2011). Bununla birlikte, geleneksel sabit soru sayılı testlerde temel amaç, ortalama zorlukta ve yüksek ayırt edici maddelerle güvenilirliği en üst düzeye çıkarmaktır (Weiss, 2011). Bu yaklaşım, yetenek dağılımın ortalarındaki yetenek seviyelerine sahip bireyleri belirlemek için faydalı olabilir; ancak, daha yüksek ve daha düşük yetenek düzeylerindeki bireyler açısından sonuçların kesinliği ve doğruluğu bakımlarından sorunlara yol açmaktadır (Wainer et al., 2000; Weiss, 2011). Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT), bu tür sorunlara çözüm olabilme potansiyeli sunmaktadır. Barnard (2018), Tseng (2016), Kalender ve Berberoğlu (2017) tarafından yapılan çalışmalar, BOBUT ile bireylere farklı testleri dinamik olarak oluşturarak, çok farklı düzeylerdeki bireylerin yetenek kestirimlerini geçerlik ve güvenirlik kaybı olmadan hem daha az maddeyle hem de daha kısa sürede olabileceğini göstermektedir.

Dil yeterliğinin kavramsallaştırılması da dil konusundaki ölçme ve değerlendirme çalışmaları açısından göz önünde bulundurulması gereken başka bir konudur. Bu bağlamda dil yeterliği hem tek bir bütün hem de parçalar şeklinde değerlendirilebilir. (Harsch, 2014). Bu konuda, tek boyutlu, birbirleriyle ilişkili yapılar, yüksek dereceli ve çift faktörlü modeller gibi çeşitli kavramsallaştırmalar bulunmaktadır (Dunn & McCray, 2020). Yani, bazı dil testlerinin odak noktası genel dil yeterliği olabilirken, diğer testlerde dil yeterliğinin seçilen faktöriyel modele bağlı olarak okuma, dinleme, konuşma, yazma, dilbilgisi ve kelime dağarcığı gibi alt alanlara bölünebileceğini düşünebilir.

Diğer bir ifadeyle dil yeterliğinin tanımı kavramsal olarak karmaşık bir konudur; bu nedenle yeterlik için tek bir puana güvenmek gereğinden fazla basitleştirilmiş bir yapı anlamına gelebilir (Spolsky, 2008). Ayrıca Harsch'a (2014) göre dil yeterliğinin ölçümünde aşırı basitleştirmeden kaçınmak çok önemlidir . Bu bakımdan dil yeterliğinin nasıl tanımlanacağı öncelikle ölçümde kullanılan testin boyutsallığına bağlıdır.

Bu çalışma, tek bir bütün veya bölünebilir parçalar olarak değerlendirilebilecek dil testinin nasıl kavramsallaştırılacağını belirlemeyi ve ardından bir kağıt kalem testinden elde edilen gerçek verilerden yola çıkarak BOBUT yaklaşımının sınıflandırma performansının kağıt kalem testi sonuçları ile karşılaştırmalı olarak incelemeyi amaçlamaktadır.

Yöntem

[1] Bilkent Üniversitesi

Bu çalışma birbirini takip eden iki aşama şeklinde gerçekleştirilecektir. Öncelikle, dil yeterliğinin kavramsallaştırılması, farklı faktöriyel modellerin model uyum analizleri ile incelenecektir. Daha sonra, belirlenmiş optimum faktöriyel model(ler) BOBUT kullanılarak aralarındaki sınıflandırma performansları karşılaştırılacaktır.

İyi uyum gösteren faktöriyel model(ler)i kullanarak BOBUT ile sınıflandırma performanslarının karşılaştırılacaktır. Bu adım, post-hoc simülasyonlar aracılığıyla gerçekleştirilecektir. Post-hoc simülasyonlar, farklı başlangıç kurallarının, sonlandırma kurallarının ve test uzunluklarının bir kombinasyonuna dayalı olacaktır. Bu aşamada sonuçlar aynı sınavın kağıt kalem uygulamasından gelen sonuçlar ile karşılaştırmalı olarak incelenecektir.

Öğrencilerin dil yeterlik düzeylerine göre doğru bir şekilde sınıflandırılmasının, dil öğretiminde müfredat geliştirme ve öğretim tasarımı bağlamında kritik bir rol oynaması bakımından önemli olduğu açıktır. Ayrıca, bu tür puanların alt puan türlerinin de ifade edilerek raporlanması, birden çok boyuta sahip olmanın daha kapsamlı bir dil öğrenen profili sağlayabileceği anlamına gelir. Bu bağlamda bu çalışmanın sonuçların program geliştiricilere de ışık tutması beklenmektedir.

Bu amaçla araştırma soruları şu şekilde yazılmıştır.:

1. Dil becerisinin ölçülmesinde farklı boyutların birbirlerini tamamlama ve telafi etme durumu düşünüldüğünde, hangi faktöriyel modeller İngilizce Düzey Belirleme Sınavına iyi uyum sağlamaktadır?

2. Farklı başlama kuralları, bitirme kuralları ve test uzunluklarının kombinasyonları uygulandığında, iyi uyuma sahip faktöriyel modellerin sınıflandırma performansları, İngilizce Düzey Belirleme Sınavı bağlamında BOBUT ve kağıt-kalem versiyonları nasıl farklılık göstermektedir?

Beklenen Bulgular

Bu çalışmanın beklenen bulguları belirli yönleri kapsamaktadır. Öncelikle, dil becerisini ölçen testlerde boyutsallık açısından testin tek boyutlu, birbiriyle ilişkili yapılar, yüksek dereceli ve çift faktörlü gibi faktöriyel modellerin hangileri ile iyi uyum sağladığını belirlemek önemlidir. Bu çalışmada ifade edilen modellerden en az ikisinin iyi bir uyum sağlaması beklenmektedir. Örnek vermek gerekirse, daha önce yapılan bir çalışmada (Dunn & McCray, 2020) yukarıda belirtilen faktöriyel modellerin hepsi iyi uyum sağlamıştır.

Bu farklı kombinasyonlar dikkate alınarak yapılacak simülasyonun canlı BOBUT kullanılarak yapılacak testlerin bu kurallar bakımından şekilde optimal olacağına işaret edeceği düşünülmektedir.

Anahtar Kelimeler: Sınıflandırma, Dil Yeterliği, Dil Testi, Boyutluluk, Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT)

Bireye Uyarlanmış Hibrit Test Desenlerinde Kullanılan Test Birleştirme Stratejilerinin Karşılaştırılması: "Anında" ve "Dinamik" Test Birleştirme

Özge Altıntaş[1]

ÖZET

Giriş

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) uygulamaları, testi alan bireylerin yetenek düzeyine uygun maddelerle karşılaşmasını sağlayan bir algoritmanın kullanıldığı bilgisayar tabanlı bir test türüdür. BOBUT, geleneksel kağıt-kalem testlerine göre kapsam geçerliğini koruyarak test süresini ve uzunluğunu azaltırken istenilen ölçüm kesinliğini sağlayabilme potansiyeline sahiptir (Hambleton ve diğ., 1991; Lord, 1980; Wainer, 2000; Weiss, 1983, 2004). Bu avantajlarının yanında BOBUT'un, bir dizi sınırlılığı da bulunmaktadır. Bunlardan biri, test güvenliğiyle ilgilidir ve madde seçimi esnasında ortaya çıkmaktadır. Çünkü bireyin yetenek kestiriminin en iyi şekilde yapılabilmesi için bilgi düzeyi en yüksek olan maddelerin seçilmesi gerekmektedir. Ayırt ediciliği yüksek kaliteli maddelerin sıklıkla kullanılması ise bu maddelerin açığa çıkmasına (item exposure) neden olarak test güvenliğine zarar vermektedir (Chang ve Ying, 1999). Bunun yanında, bireylerin yanıt verdikleri maddelere dönüp yanıtlarını kontrol edememeleri ya da değiştirememeleri de önemli bir sınırlıktır (Han, 2013; Luecht ve Sireci, 2011). Madde düzeyinde uyarlanabilir test uygulamalarının yukarıda sözü edilen sınırlılıklarını ortadan kaldırmak üzere, kâğıt-kalem testleri ile BOBUT'un avantajlarını birleştirerek oluşturulmuş bir test yaklaşımı olan Çok Aşamalı Bireyselleştirilmiş Test (ÇABT) uygulamaları her iki test yaklaşımı arasında dengeli bir uzlaşı sağlamaktadır (Hambleton ve Xing, 2006; Hendrickson, 2017; Jodoin ve diğ., 2006; Zenisky ve diğ., 2010). ÇABT, test geliştiricilerin test uygulamasından önce test formlarını gözden geçirmelerine ve test kısıtlamalarını test uygulaması esnasında değil de test geliştirme esnasında değerlendirebilmelerine olanak tanıyarak karmaşık kısıtlamaların uygulanmasını kolaylaştıran bir yaklaşım sunar. Bunun yanında, testi alan bireylerin bir aşama içindeki maddeler arasında gezinmesine izin verir, bu da doğrusal testlerdekine benzer bir test deneyimi sağlar (Luo ve Wang, 2019). Öte yandan, bu avantajların test verimliliğinin düşmesine neden olduğu belirtilmekte ve alanyazında hem BOBUT'un hem de ÇABT'ın avantajlarını birleştiren hibrit (Wang ve diğ., 2016) bazı test birleştirme stratejileri önerilmektedir. Bunlardan biri, ÇABT gibi aşamalar halinde uygulanan ve her aşamadaki modüllerin uygulamadan önce değil de anında anında bir araya getirildiği Anında Birleştirilmiş Çok Aşamalı Bireyselleştirilmiş Test (On-the-Fly Assembled Multistage Adaptive Testing - ABÇABT) stratejisidir (Chang, 2015; Zheng ve Chang, 2014, 2015; Zheng ve diğ., 2014). Bir diğeri ise hem madde düzeyinde hem de aşama düzeyinde uyarlamayı test oturumuna özgü dinamik bir ara havuz kullanarak yapan Dinamik Çok Aşamalı Test (Dynamic Multistage Testing -DiÇAT) stratejisidir (Luo ve Wang, 2019). Alanyazında her iki test birleştirme stratejisi için hem BOBUT hem de ÇABT ile karşılaştırılarak avantajlarının ortaya koyulduğu çalışmalar bulunmaktadır (Du ve diğ., 2019; Luo ve Wang, 2019; Tay, 2015; Wang ve diğ., 2016; Zheng ve Chang, 2015). Bu araştırmalarda, ABÇABT'ın ölçüm doğruluğunun BOBUT ve ÇABT ile karşılaştırılabilir olduğu; hem hesaplama kolaylığı hem de tüm kısıtlamaların aynı anda karşılanabilmesi açısından BOBUT'a göre daha üstün olduğu; aşamaların her bireyin yetenek seviyesine uyacak şekilde anında bir araya getirilmesiyle ÇABT'tan daha fazla bilgi sağladığı raporlanmıştır (Zheng ve Chang, 2015). Benzer şekilde, DiÇAT'ın da hem BOBUT hem de ÇABT'a göre üstünlüklerinin olduğu belirtilmektedir. Örneğin BOBUT ile karşılaştırıldığında, DiÇAT'in daha küçük ancak özenle seçilmiş ve test uygulaması sırasında aşama düzeyinde uyarlama yoluyla dinamik olarak oluşturulan bir madde havuzuna sahip olmasının, herhangi bir güvenlik ihmalinde tüm

[1] Ankara Üniversitesi

madde havuzunun kaybedilmesi riskini azalttığı vurgulanmaktadır. Bunun yanında, ÇABT ile karşılaştırıldığında DiÇAT'ın uygulama sırasında bir test uyarlama katmanı daha bulunmasının ölçmenin standart hatasının azaltılmasını hızlandırarak erken yönlendirme kararlarını kolaylaştırdığı da belirtilmektedir (Luo ve Wang, 2019). Özetle hem BOBUT'a hem de ÇABT'a göre avantajları örnek çalışmalarla ortaya koyulan bu iki test birleştirme stratejisinin birbirlerine karşı üstünlüklerinin de araştırılması uygulayıcıların ihtiyaçlarına uygun stratejiyi seçmelerine olanak tanıması anlamında gerekli görülmektedir. Buna göre bu araştırma, test verimliliğinin görece yüksek olduğu ve içeriğin kontrolüne iyi düzeyde imkân sağladığı bilinen ABÇABT ile DiÇAT test birleştirme stratejilerini yetenek kestiriminin doğruluğu ve test verimliliği açısından karşılaştırmayı amaçlamaktır.

Yöntem

Bu araştırmada, ABÇABT ve DiÇAT test birleştirme stratejileri, yetenek kestiriminin doğruluğu ve test verimliliği açısından farklı koşullar altında yürütülen simülasyon çalışmalarıyla incelenecektir.

Bu koşullara geçmeden ÇABT uygulamalarındaki dört temel bileşenden söz etmek gerekir: Bunlar, modül, panel, aşama ve yoldur. (Luecht ve Sireci, 2011). Birden fazla maddenin güçlük düzeylerine göre bir araya gelerek oluşturduğu madde grubuna/bloğuna modül; modüllerin bir araya getirilmesiyle oluşan modül grubuna/uygulama birimine ise panel adı verilmektedir. Yol ise, bireyin panel içerisindeki aşamalarda ve modüller arasında izleyeceği rotadır. ÇABT'ta test birleştirme süreci, maddelerin modüllere modüllerin de paneller içindeki aşamalara atanmasıyla panel düzeyinde gerçekleşmektedir (Luecht ve Nungester, 1998). Yani, maddelerin modüllere ve modüllerin paneller içindeki aşamalara atanması test birleştirme sürecinin bir parçası olarak gerçekleştirilmektedir.

Bu araştırmada koşullar, ÇABT için aşamaların yapılandırılması (aşama sayısı, aşamalardaki modül sayısı ve modüllerdeki madde sayısı) ve bölümleme stratejileriyle (eşit, ilk aşamada ve son aşamada öncelikli) ilgilidir. Araştırmanın madde havuzunu ise lise son sınıf öğrencilerinin matematik ve geometri bilgilerinin kullanımını gerektiren sayısal düşünme becerilerini ölçtüğü varsayılan maddeler oluşturacaktır. Bu maddeler, üç parametreli lojistik modele göre kalibre edilecek ve hazırlanan belirtke tablosuna göre kapsam kontrolü sağlanacaktır.

Bulgular/Beklenen Bulgular

ABÇABT ve DiÇAT, test verimliliğinin yüksek olduğu ve kapsam dengesinin sağlanabildiği görece nitelikli testler oluşturabilmede kullanılan uyarlanabilir birer test birleştirme stratejisidir. BOBUT ve ÇABT'ın avantajlarını birleştiren hibrit test desenlerinde kullanılan bu test birleştirme stratejilerinin yetenek kestiriminin doğruluğu ve test verimliliği açısından karşılaştırılmasının amaçlandığı bu çalışmadan elde edilecek bulguların uygulamada kullanılacak hibrit bir test deseni altında uygun bir test birleştirme stratejisine nasıl karar verileceği ve bunun nasıl sürdürüleceği konusunda uygulayıcılara ipuçları sunması beklenmektedir. Bunun yanında araştırmanın uygulayıcılara, belirli koşullar altında ve belirli bir madde havuzu için test birleştirme stratejilerini karşılaştırmalı olarak inceleme imkânı sunacağı düşünülmektedir.

Kaynakça

Chang H-H. (2015). Psychometrics behind computerized adaptive testing. Psychometrika, 80(1), 1-20. https://doi.org/10.1007/s11336-014-9401-5

Chang, H-H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23(3), 211-222. https://doi.org/10.1177/01466219922031338

Du, Y., Li, A., & Chang. H-H. (2019). Utilizing response time in on-the-fly multistage adaptive testing. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), Quantitative Psychology, IMPS 2017 Proceedings in Mathematics & Statistics, Vol 265, pp 107-117. Springer. https://doi.org/10.1007/978-3-030-01310-3_10

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

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

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. Applied Psychological Measurement, 37(4), 259-275. https://doi.org/10.1177/0146621612473638

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

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

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates.

Luecht R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement, 35(3), 229-249. https://www.jstor.org/stable/1435202

Luecht, R. M., & Sireci, S. G. (2011). A review of models for computer-based testing. Research Report 2011-12. College Board. https://files.eric.ed.gov/fulltext/ED562580.pdf

Luo, X., & Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. International Journal of Testing, 19(3), 227-247. https://doi.org/10.1080/15305058.2019.1621871

OECD (2013). Technical report of the Survey of Adult Skills (PIAAC). Chapter 17 (pp. 406-438). OECD Publishing. https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

OECD. (2019a). PISA 2018 Assessment and Analytical Framework. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2019b). PISA 2018 Technical Report. Chapter 2 (pp. 1-32). OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-02-Test-Design.pdf

Wainer, H. (Ed.). (2000). Computerized adaptive testing: A primer (2nd ed.). Lawrence Erlbaum Associates Publishers.

Wang, S., Lin, H., Chang, H.-H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. Journal of Educational Measurement, 53(1), 45-62. https://doi.org/10.1111/jedm.12100

Weiss, D. J. (Ed.). (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. Academic Press.

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), 71-84. https://doi.org/10.1080/07481756.2004.11909751

Yamamoto, K., Shin, H. & Khorramdel, L. (2019). Introduction of multistage adaptive testing design in PISA 2018. OECD Education Working Papers, No. 209. OECD Publishing.
https://doi.org/10.1787/b9435d4b-en

Yan, D., von Davier, A. A., & Lewis, C. (Eds.) (2014). Computerized multistage testing: Theory and applications. CRC Press.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden, & C. A. W. Glas (Eds.), Elements of adaptive testing (pp. 355-372). Springer.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H-H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized multistage testing: Theory and applications (pp. 87-99). CRC Press.

Zheng, Y., & Chang, H-H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. In Y. Cheng, & H.-H. Chang (Eds.), Advancing methodologies to support both summative and formative assessments (Chapter 2, pp. 21-39). Information Age.

Zheng, Y., & Chang, H-H. (2015). On-the-fly assembled multistage adaptive testing. Applied Psychological Measurement, 39(2), 104-118. https://doi.org/10.1177/0146621614544519

Anahtar Kelimeler: bireye uyarlanmış hibrit test, test birleştirme, anında test birleştirme, dinamik test birleştirme, çok aşamalı bireyselleştirilmiş test

CAT ve MST'nin Hibrit Modellerinin Performanslarının Karşılaştırmalı Analizi: Verimlilik, Doğruluk ve Psikometrik Özellikler

YUSUF BALCI[1]

ÖZET

Giriş

Bilgisayar tabanlı adaptif testler (CAT) ve çok aşamalı testler (MST), farklı güçlü ve zayıf yanlarıyla büyük ölçekli testlerde popüler hale gelmiştir. Bu çalışmanın temel amacı, CAT ve MST'nin avantajlarını birleştirirken kısıtlamalarını en aza indirgeyen son dönem araştırmalarda önerilen iki hibrit modelin (Wang, Lin, Chang, & Douglas, 2016; Raborn & Sarı, 2021) performanslarını karşılaştırmaktır. Bu çalışma, ilgili hibrit modellerin verimlilik, doğruluk ve diğer psikometrik özellikleri değerlendirerek, çeşitli test senaryolarında uygulanabilirlikleri konusunda tavsiyeler sunmayı ve bireye uyarlanmış test alanına katkıda bulunmayı amaçlamaktadır.

Yöntem

Bu çalışma, ayrı makalelerde sunulan iki hibrit modelin performans özelliklerinin karşılaştırmalı bir analizini gerçekleştirmektedir. Yöntemler şunları içerir:

Önerilen hibrit modellerin, tasarımlarının ve orijinal makalelerdeki simülasyon sonuçlarının kapsamlı bir incelemesi ve değerlendirilmesi.

İki hibrit modelin ve ilgili standart CAT ve MST tasarımlarının performans özelliklerinin (tahmin doğruluğu, verimlilik ve diğer ilgili özellikler gibi) karşılaştırılması.

Beklenen Bulgular

İki hibrit modelin karşılaştırmalı analizinden elde edilen bulgular, test performansında potansiyel iyileştirmelere işaret edebilecektir:

Her iki hibrit modelin de, ilgili standart CAT ve MST tasarımlarına kıyasla tahmin doğruluğu ve verimlilik bulguları özetlenecektir. Böylece, CAT ve MST yaklaşımlarını birleştirmenin potansiyeli değerlendirilecektir.

Her iki modelde de testin erken aşamalarında yetenek tahmininde iyileşme ve daha sonraki aşamalarda daha fazla adaptif nokta bulunma durumu göz önüne alınarak CAT ve MST yöntemlerini birleştirmenin faydaları özetlenecektir.

Bu çalışmayla incelenecek hibrit modellerin, geleneksel CAT ve MST tasarımlarına ümit verici alternatifler sunma potansiyeli irdelenecektir. Bununla birlikte, bu bulguların farklı test bağlamlarında doğrulanması ve bu hibrit yaklaşımların pratik uygulamasıyla ilgili zorlukları değerlendirmek için yapılabilecek araştırmaların çerçevesi belirlenmeye çalışılacaktır.

Kaynakça

Raborn, A. & Sarı, H. (2021). Mixed Adaptive Multistage Testing: A New Approach. Journal of Measurement and Evaluation in Education and Psychology, 12(4), 358-373. doi: 10.21031/epod.871014

[1] EGE ÜNİVERSİTESİ

Wang, S., Lin, H., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. Journal of Educational Measurement, 53(1), 45-62. doi: 10.1111/jedm.12100

Anahtar Kelimeler: Bilgisayar tabanlı adaptif testler (CAT), çok aşamalı testler (MST), hibrit modeller, performans karşılaştırması, verimlilik, tahmin doğruluğu, psikometrik özellikler

Çok Aşamalı Bireye Uyarlanmış Testlerde Farklı Modül Uzunluklarıyla Oluşturulan Tasarımların Karşılaştırılması

Güneş Ertaş[1] Duygu Anıl[2]

ÖZET

GİRİŞ

Günümüzde halen önemli karaların alındığı geniş ölçekli birçok test, kâğıt-kalem ortamında klasik lineer test olarak uygulanmaktadır. Bu testlerde de amaç bireyin performansını en az hata ile ölçebilmektedir. Ancak klasik testler, ortalama yetenek düzeyindeki bireylerin performansını ölçmede daha başarılı iken ortalamanın altında veya üstünde kalan bireyler için çok zor veya çok kolay olabilmektedir (Hambleton & Swaminhatan, 1985; Lord, 1980; Wainer, 2000; Weiss, 1983). Her bireyin aynı sıra ve sayıda maddeleri aldığı klasik testlere göre, bireye uyarlanmış testler, daha az soru ile tüm yetenek dağılımı boyunca bireyin yetenek düzeyine ilişkin daha kesin ve etkili ölçme yapabilmektedir.

Çok aşamalı bireye uyarlanmış testler (Multistage Test - MST), bireylerin yeteneklerine göre uygun güçlükte ve belirli ortak özelliklere sahip maddelerden oluşan "modül" adı verilen madde bloklarıyla karşılaştığı uygulamalardır. Günümüzde birçok uluslararası geniş ölçekli test, bilgisayar ortamında çok aşamalı bireye uyarlanmış test olarak uygulanmaktadır (Breirhaupt, Zhang ve Hare, 2014; Educational Testing Servise, 2019; Kirsch & Lennon, 2017; Robin, Stefan ve Liang, 2014).

Bir MST uygulamasında, birey yetenek düzeyine bağlı olarak iki ya da daha fazla aşamada yer alan farklı güçlükteki modüllere yönlendirilir. Bireyin ilk aşamadaki modülde yer alan tüm maddelere verdiği cevaplara göre birey, ikinci aşamada yer alan uygun (güçlükte) modüle yönlendirilir. Benzer şekilde ikinci aşamadan, üçüncü ve varsa sonraki aşamalara bu şekilde yönlendirilir. Bir MST oluştururken; aşama sayısı, her bir aşamada yer alacak modül sayısı ve modüllerde yer alacak madde sayıları ile birlikte bireyin performansının nasıl puanlanacağı ve aşmalar arası modüllere nasıl yönlendirileceği önceden tek tek planlanır (Hendrickson, 2007; Yan, Lewis & von Davier, 2014; Zenisky, Hambleton & Luecht, 2010). Bu özelliği ile MST farklı tasarımlara olanak sağlayan esnek bir yapıya sahiptir. Bu araştırmada da farklı madde güçlük dağılımlarına göre ve farklı modül uzunluklarıyla oluşturulan MST tasarımlarının gerçek ve kestirilen yetenek düzeyleri bazında karşılaştırılması amaçlanmıştır.

YÖNTEM

Bu çalışmada çok aşamalı bireye uyarlanmış testlerde farklı simülasyon koşulları altında elde edilen yetenek kestirimleri karşılaştırılacaktır. Bunun için, tüm koşullardan elde edilen kestirimlerin ortalama hata ve yanlılık değerleri hesaplanmıştır. Tüm veriler bu çalışma için bilgisayar programı kullanılarak türetileceğinden bu çalışma bir Monte Carlo simülasyon çalışmasıdır. Bu çalışma için veriler, açık kaynak kodlu ve profesyonel bir yazılım olan RStudio'da üretilmiş ve çok aşamalı bireye uyarlanmış test simülasyonu, yazılım içerisinde yer alan mstR paketi (Magis, Yan & von Davier, 2018) yardımıyla yapılmıştır.

Bu çalışmada üç aşamadan oluşan 1-2-3 panel tasarımı (birinci aşamada 1, ikinci aşamada 2 ve üçüncü aşamada 3 modül) kullanılmış ve 2PL modele göre madde havuzları oluşturulmuştur.

[1] Boğaziçi Üniversitesi
[2] Hacettepe Üniversitesi

Madde ayırt edicilik parametresi sabit tutulmuş ve lognormal (0.0, 0.3) olarak belirlenmiştir. Madde güçlüklerinin normal dağılımdan ve uniform dağılımdan türetildiği modüllerden oluşan iki farklı madde havuzu oluşturulmuştur. Her bir madde havuzu için, test bütününde 45 maddenin yer alacağı farklı uzunluktaki modüllerden oluşan 5 farklı tasarım oluşturulmuştur. İlk aşamadan son aşamaya artan (10-15-20) ve ilk aşamadan son aşamaya azalan (20-15-10) modül uzunlukları tasarımları ile en uzun modülün ikinci aşamada olduğu (10-20-15 ve 15-20-10) ve her aşamada eşit uzunlukta (15-15-15) modüllerin bulunduğu tasarımlar oluşturulmuştur. Her bir modülden 10, 15 ve 20'şer olmak üzere 45'er madde türetilmiştir. Her bir güçlük dağılımı için 270 ve toplam 540 madde türetilmiştir. Yönlendirme yöntemi olarak her bir modülde cevaplanan doğru madde sayısını temel alan birey odaklı statik yönlendirme yöntemi kullanılmıştır (Weissman, 2014).

Gerçek yetenek dağılımları -3, 3 aralığında standart normal dağılımdan N (0,1) 500 birey için türetilmiştir. Cevap örüntüleri mstR paketinde yer alan genPattern argümanıyla oluşturulmuştur. Her bir aşamadaki modül uzunluklarına göre birbirinden farklılaşan 5 MST tasarımı iki farklı madde havuzu (normal dağılım ve uniform dağılım) için ayrı ayrı karşılaştırılmıştır. Tüm koşullar için 100 replikasyon sonucu EAP (expected a posteriori) yöntemi ile elde edilmiş yetenek kestirimleri ile gerçek yeteneklere ait yanlılık (bias) ve hata kareleri ortalamasının karekökü (Root Mean Square Error - RMSE) değerleri hesaplanmıştır. RMSE ve yanlılık değerleri arasındaki farklılığın istatistiksel olarak anlamlı olup olmadığı varyans analizi (ANOVA) ile test edilmiştir.

BULGULAR

Madde güçlüklerinin normal dağılımdan ve uniform dağılımdan türetildiği modüllerden oluşan tasarımlar ayrı ayrı incelenmiştir. Normal dağılımdan türetilen verilerle oluşturulan tasarımlar için farklı modül uzunluklarının, RMSE değerlerinde anlamlı bir fark oluşturduğu ve etki büyüklüğünün (η^2=0,31) yüksek düzeyde olduğu bulunmuştur. Farklı aşamalarda faklı uzunluktaki modüllerden oluşan beş tasarımın birbirinden nasıl farklılaştığını görmek için Bonferroni yöntemi kullanılarak Post-Hoc analizi yapılmıştır. Buna göre ilk aşamada 10 modül ile başlanan iki tasarımdan elde edilen RMSE değerleri arasında istatistiksel olarak anlamlı farklılık olmadığı tespit edilmiştir. Bu iki tasarımda diğer tasarımlara göre istatistiksel olarak anlamlı daha az hata gözlenmiştir. Yanlılık değerleri için de modül uzunluğundan oluşan farkın istatistiksel olarak anlamlı ve etki büyüklüğünün (η^2=0,16) yüksek düzeyde olduğu bulunmuştur. Son aşamada 10 modül ile biten iki tasarım arasında yanlılık değerleri üzerinde istatistiksel olarak anlamlı bir farklılaşmanın olmadığı ve sıfıra en yakın yanlılık değerlerinin bu iki tasarımda olduğu tespit edilmiştir.

Unifrom dağılımdan türetilen tasarımlar için de modül uzunlukları faktörünün RMSE değerleri üzerinde istatistiksel olarak anlamlı bir etkisi olduğu ve etki büyüklüğünün (η^2=0,50) yüksek düzeyde olduğu bulunmuştur. Benzer şekilde yapılan analizler sonucunda normal dağılımdaki bulgulardan farklı olarak son aşamada 10 modül ile biten iki tasarımdan elde edilen RMSE değerlerinin arasında istatistiksel olarak anlamlı farklılık olmadığı ve diğer tasarımlara göre istatistiksel olarak anlamlı daha düşük RMSE değerine sahip oldukları tespit edilmiştir. Modül uzunluğunun, yanlılık değerlerinin farklılaşması üzerinde istatistiksel olarak anlamlı bir etkisi olduğu ve etki büyüklüğünün (η^2=0,14) yüksek düzeyde olduğu bulunmuştur. Her aşamada eşit madde sayısı olan modül tasarımı için yanlılık değeri istatistiksel olarak anlamlı sıfıra daha yakındır.

Svetina, Liav, Rutowski ve Rutkowski (2019), geniş ölçekli çok aşamalı bireye uyarlanmış bir test simülasyonu olarak farklı tasarımları farklı koşullar altında karşılaştırdıkları çalışmalarında, her koşulda en iyi çalışan bir yöntemin olmadığına dikkat çekmişlerdir. Bu çalışmada da benzer bir sonuca varılmaktadır. Farklı güçlü dağılımlarından oluşturulan tasarımlarda, iki koşul için en iyi sonucu veren tasarımlar farklılaşmaktadır. Bu farklılşamanın nedenleri de tartışılacaktır.

KAYNAKÇA

Breithaupt, K., Zang, O. Y., & Hare, D. R. (2014). The multistage testing approach to the AICPA uniform certified public accounting examinations. In D. Yan, C. Lewis, & A. A. von Davier (Eds.), Computerized Multistage Testing: Theory and Applications (pp. 343–354). Chapman and Hall/CRC.

Educational Testing Service. (2019). Introduction of multistage adaptive testing design in PISA 2018 (Paper OECD Education Working Paper No. 209).

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Springer.

Hendrickson, A. (2007). An NCME Instructional Module on Multistage Testing. Educational Measurement Issuses and Practice, 26(2), 44–52.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: A new design for a new era. Large-Scale Assessments in Education, 5(1), 11.

Lord, F. M. (1980). Applications of Item Response Theory To Practical Testing Problems. Lawrence Erlbaum Associates.

Magis, D., Yan, D., & von Davier, A. (2018). mstR: Procedures to Generate Patterns under Multistage Testing. R package version 1.2.

Robin, F., Steffen, M., & Liang, L. (2014). The Multistage Test Implementation of GRE Revised General Test. In D. Yan, C. Lewis, & A. A. von Davier (Eds.), Computerized Multistage Testing: Theory and Applications (pp. 325–342). Chapman and Hall/CRC.

Svetina, D., Liaw, Y.-L., Rutkowski, L., & Rutkowski, D. (2019). Routing Strategies and Optimizing Design for Multistage Testing in International Large-Scale Assessments. Journal of Educational Measurement, 56(1), 192–213.

Wainer, H. (2014). Introduction and History. In N. J. Dorans, R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), Computerized Adaptive Testing: A Primer (2nd ed., pp. 1–22). Routledge.

Weiss, D. J. (1983). New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. Academic Press.

Weissman, A. (2014). IRT-Based Multistage Testing. In D. Yan, C. Lewis, & A. A. von Davier (Eds.), Computerized Multistage Testing: Theory and Applications (pp. 153–168). Chapman and Hall/CRC.

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of Computerized Multistage Tests. In D. Yan, C. Lewis, & A. A. von Davier (Eds.), Computerized Multistage Testing: Theory and Applications (pp. 3–20). Chapman and Hall/CRC.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage Testing: Issues, Design and Research. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of Adaptive Testing (pp. 355–372). Springer.

Anahtar Kelimeler: çok aşamalı bireye uyarlanmış testler, modül tasarımları

Okuma Metinlerine Dayalı Maddelerin Çok Aşamalı Bireyselleştirilmiş Hibrit Test Desenlerine Uyarlanması

Özge Altıntaş[1] Ömer Kutlu[1]

ÖZET

Giriş

Okuduğunu anlama becerisi, bir metnin içeriğini yorumlayabilme, ilişkilendirebilme ve değerlendirebilmeyi içeren temel bir beceridir. Bunun yanında, öğrencilerin bilgiye nasıl erişeceği ve onu nasıl kullanacakları konusunda onlara öngörü sağlarken öğrenilenleri anlamlandırmalarını ve gerçek yaşam durumlarına aktarmalarına da destek olmaktadır. Yapılan araştırmalar, yaşam için temel becerilerden biri olan okuduğunu anlama becerisini geliştirmenin önemine vurgu yapmaktadır (Deshler ve diğ., 2012; Kirsch ve diğ., 1993; Kutlu ve diğ., 2019; OECD, 2013; Snow, 2002). Okuduğunu anlama becerisi bireyin soyut düşünmesine, güncel olay ve bilgileri izlemesine katkıda bulunmakta (Mullis ve diğ., 2007), sosyal ve akademik yaşamına destek olmaktadır (Kutlu vd., 2011). Bu becerinin geliştirilebilmesi için öğretmenlerin, öğrencileri sınıf içi öğrenme süreçlerinde geribildirime olanak sağlayan nitelikli metinlerle ve bu metinlere dayalı maddelerle karşılaştırmaları gerekir.

Öğrenmenin kendisi gibi bir becerinin geliştirilmesi de bireyden bireye göre değişmektedir. Yan (2020), öğrencileri daha iyi tanıyabilmenin ve öğrenciler hakkında daha fazla bilgi edinebilmenin önemine dikkat çekerek daha kısa ve hedefe yönelik değerlendirmelerle geribildirim verilmesini, kanıta dayalı değerlendirmelerle öğretim sürecinin hızlı biçimde izlenmesi gerektiğini vurgulayarak ölçme ve durum belirlemenin verimliliğini ve doğruluğunu artırmak için yeni yöntemlerin ve uygulamaların geliştirilmesini, bireyselleştirilmiş öğrenmeyi gerçekleştirebilmek için bilgisayar ortamında bireye uyarlanmış ve çok aşamalı testlerin sürece dahil edilmesini savunmaktadır.

Günümüzde bireye uyarlanmış test uygulamaları, bilgisayar teknolojilerindeki gelişmeyle birlikte, sınıf içi durum belirleme süreçlerinde de kullanılmaya başlamıştır (Chang, 2012; Choi ve Tinkler, 2002; Jiao ve Lissitz, 2012; Kingsbury ve Hauser, 2004; Way, 2005). Dolayısıyla, teknolojinin hızla geliştiği bir dönemde öğretmenlerin bireyselleştirilmiş test uygulamalarını bilmeleri ve sınıf içi durum belirleme süreçlerinde kullanmaları kaçınılmazdır. Kâğıt-kalem testlerine göre daha az maddeyle daha kısa sürede ve daha kesin ölçümler elde edebilme gibi üstünlükleri bulunan bireyselleştirilmiş test uygulamaları (Hambleton ve diğ., 1991; Wainer, 2000; Weiss, 1983; 2004), öğrencilerin okuduğunu anlama becerilerindeki bireysel gelişimlerini izlemede ve onlara geribildirim sunmada etkin bir araç olarak kullanılabilir. Buradan hareketle bu araştırmanın amacı ilkokul 4. sınıf öğrencilerinin okuduğunu anlama becerilerindeki gelişimlerini izlemede kullanılacak metinlere dayalı maddelerin çok aşamalı bireyselleştirilmiş hibrit (melez) test desenlerine uygulanmasıdır.

Yöntem

Hibrit test desenleri, Bilgisayar Ortamında Bireye Uyarlanmış Test -BOBUT- (Computerized Adaptive Test -CAT) ile Çok Aşamalı Bireyselleştirilmiş Test -ÇABT- (Multistage Adaptive Test -MSAT-) uygulamalarının güçlü yönlerini alarak karma bir yaklaşımla birleştirmektedir (Wang ve diğ., 2016; Zheng ve diğ., 2014; Zheng ve Chang, 2015). Bu test deseninin seçilme nedeni, okuduğunu anlama becerisini değerlendirmede oldukça güçlü bir ölçme yaklaşımı olan okuma metinlerine dayalı madde gruplarından oluşan testçiklerin (testlet, Wainer ve Kiely, 1987) kullanılmasıdır. Testçik, bir birim

[1] Ankara Üniversitesi

olarak geliştirilen ve testi alan bireyin önceden belirlenmiş sayıda yol içeren tek bir içerik alanıyla ilgili bir grup maddeyi ifade etmektedir. Ortak bir okuma metnine dayalı bir dizi okuduğunu anlama maddesi bu anlamda kullanılmaktadır (Keng, 2008).

Bu araştırmanın iki önemli boyutu vardır; bunlardan biri metin boyutu diğeri madde boyutudur. Metin boyutunu, metnin türü, metnin okunabilirliği (readibility), metindeki cümle ya da sözcük sayısı gibi etkenler oluştururken madde boyutunu, maddeyle ölçülmeye çalışılan bilişsel düzey ve içerik gibi etkenler oluşturmaktadır. Eğer yalnızca, metnin güçlük düzeyi kullanılarak bir desen oluşturulursa bir başka anlatımla metinler de uyarlanabilir şekilde seçilirse, bu durum aşırı zor ya da aşırı kolay maddeler içeren metinlerin seçilmesine neden olabilir. Özetle, her iki bireyselleştirilmiş test deseninin güçlü yönlerini alarak araştırmanın amacı doğrultusunda geliştirilecek hibrit test deseni hazırlık ve süreç olmak üzere iki aşamalı biçimde şu adımları içerecektir:

1. Hazırlık

- İlkokul 4. sınıf öğrencilerinin okuduğunu anlama başarısını belirleyecek okuma metinlerinin ve metinlere dayalı maddelerin yazılması

- Her bir metne ilişkin maddelerin yazılmasında Uluslararası Okuma Becerilerinin İzlenmesi Projesi (Progress in International Reading Literacy Study -PIRLS-) sınıflamasındaki dört aşamalı düzeyin dikkate alınması (Mullis ve diğ., 2016).

- Her bir metnin güçlük düzeyinin belirlenmesi ve kolaydan zora doğru hiyerarşik yapıda bir metin ve metinlere dayalı madde havuzunun oluşturulması

2. Süreç

- Metin düzeyinde MSAT testçiklerinin oluşturulması/uygulanması

- Oluşturulan/uygulanan okuma metnine ait maddelerin CAT aracılığıyla uyarlamalı olarak seçilmesi

- Yetenek kestirimi tamamlanana kadar sürecin kendi içinde devam etmesi

- Bir önceki adımla bağlantılı olarak önerilen test deseninin ölçüm doğruluğunun belirlenmesi

Bulgular/Beklenen Bulgular

Çalışmadan elde edilen bulgular, öğrencilerin okuduğunu anlama düzeylerinin (okuma yeteneğinin) belirlenmesine ve bireyselleştirilmiş geribildirime olanak sağlayacaktır. Ayrıca okuduğunu anlama becerisindeki gelişimin izlenmesinde kullanılacak bireye uyarlanmış hibrit bir testin etkililiğini de gösterecektir. Buna ek olarak sonuçların, bireye uyarlanmış hibrit bir test deseninin başarısını değerlendirmek üzere iki önemli durumu ortaya koyması da beklenmektedir. Bunlardan ilki, mevcut havuz özellikleri göz önüne alındığında, hibrit bir desenin yetenek kestiriminin doğruluğuna ilişkin performansı, aşama sayısı ve testçik boyutu da dahil olmak üzere değerlendirilmesi; bir diğeri de bireye uyarlanmış hibrit bir desenin uyarlanabilirliğinin değerlendirilmesidir. Bu çalışma öğrencilerin okuma düzeylerinin metin bazında belirlenmesine katkı sağlayacaktır. Okullar metin havuzlarından, ders kitaplarındaki standart metinlere bağlı kalmayarak, öğrencilerinin okuma düzeylerine göre metin seçebilecekler ve öğrencilerinin okuma başarısındaki gelişimlerini bu yolla sağlayabileceklerdir.

Kaynakça

Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R.W. Lissitz & H. Jiao (Eds.), Computers and their impact on state assessments: recent history and predictions for the future (pp. 195-226). Information Age Publisher.

Choi, S. W., & Tinkler, T. (2002, April). Evaluating comparability of paper-and- pencil and computer-based assessment in a K-12 setting [Paper presentation]. The Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Deshler, D. D., Ihle, F., Mark, C., Pollitt, D. T., & Kennedy, M. J. (2012). Literacy and learning. In N. M. Seel, (Ed.), Encyclopedia of the sciences of learning. Springer. https://doi.org/10.1007/978-1-4419-1428-6_553

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

Jiao, H., & Lissitz, R. W. (2012). Computer-Based testing in K-12 state assessments. In R.W. Lissitz & H. Jiao (Eds.), Computers and their impact on state assessments: recent history and predictions for the future (pp. 1-21). Information Age Publisher.

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

Kingsbury, G. G., & Hauser, C. (2004, April). Computerized adaptive testing and "No Child Left Behind" [Paper presentation]. The Annual Meeting of the American Educational Research Association. San Diego, CA.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). Adult literacy in America: A first look at the results of the national adult literacy survey. US Department of Education, National Center for Education Statistics.
https://www.scirp.org/(S(czeh2tfqw2orz553k1w0r45))/reference/referencespapers.aspx?referenceid=463592

Kutlu, Ö., Yıldırım, Ö., Bilican, S., & Kumandaş, H. (2011). İlköğretim 5. sınıf öğrencilerinin okuduğunu anlamada başarılı olup olmama durumlarının kestirilmesinde etkili olan değişkenlerin incelenmesi. Journal of Measurement and Evaluation in Education and Psychology, 2(1), 132 139. https://dergipark.org.tr/tr/pub/epod/issue/5806/77235

Kutlu, Ö., Altıntaş, Ö., Kula-Kartal, S., Özyeter, N. T., & Alpayar, Ç. (2019). Okuduğunu Anlama Becerisinin Ölçülmesi ve Değerlendirilmesi [Measuring and assessing reading comprehension skills]. Ankara University Publishing No: 678. ISBN: 978-605-136-474-2. Ankara University Printing House.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement, 35(3), 229-249. https://www.jstor.org/stable/1435202

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

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). PIRLS 2006 international report. International Study Center, Lynch School of Education, Boston College.

Mullis, I. V., Martin, M. O., & Sainsbury, M. (2016). PIRLS 2016 reading framework. In I. V. S. Mullis & M. O. Martin (Eds.). PIRLS 2016 assessment framework (pp.11 26). International Association for the Evaluation of Educational Achievement (IEA).

OECD. (2013). OECD skills outlook 2013: First Results from the survey of adult skills. OECD Publishing. https://doi.org/10.1787/9789264204256-en

Snow. C. E. (2002). Reading for understanding: Toward a research and development program in reading comprehension. RAND Reading Study Group Research Report. Prepared for the Office of Education Research and Improvement (OERI).
https://www.rand.org/pubs/monograph_reports/MR1465.html

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

Wainer, H. (Ed.). (2000). Computerized adaptive testing: A primer (2nd ed.). Lawrence Erlbaum Associates Publishers.

Wang, S., Lin, H., Chang, H.-H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. Journal of Educational Measurement, 53(1), 45-62. https://doi.org/10.1111/jedm.12100

Way, W. D. (2005). Practical questions in introducing computerized adaptive testing for k-12 assessments. Research Report 05-03. Pearson.
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.3715&rep=rep1&type=pdf

Weiss, D. J. (Ed.). (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. Academic Press.

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), 71-84. https://doi.org/10.1080/07481756.2004.11909751

Yan, D. (2020). Multistage adaptive testing in practice. In H. Jiao, & R. W. Lissitz (Eds.), Application of artificial intelligence to assessment (pp. 141-160). Information Age Publishing, Inc.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H.-H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized multistage testing: Theory and applications (pp. 87-99). CRC Press.

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. Applied Psychological Measurement, 39(2), 104-118. https://doi.org/10.1177/0146621614544519

Anahtar Kelimeler: okuduğunu anlama becerisi, metne dayalı uyarlanabilir test, testçik, hibrit test deseni, çok aşamalı bireyselleştirilmiş test

Türkiye'de Bireye Uyarlanmış Test Uygulamalarının Kapsamı ve Gelişimi için Gereksinimler

Ömer KUTLU[1] Selma ŞENEL[2]

ÖZET

Tüm dünyada ulusal ve uluslararası bağlamda, bireye uyarlanmış testler çok sayıda test kuruluşu ve kurumu tarafından tercih edilmektedir. Bu uygulamalar, bilgisayar ortamında bireye uyarlanmış test (BOBUT) ya da çok aşamalı test şeklinde olabilmektedir. Örneğin, sonuçları eğitim politikalarının düzenlenmesinde dikkate alınan Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment [PISA]) çok aşamalı test uygulamasıdır (Organisation for Economic Co-operation and Development, 2018). Lisansüstü eğitim almak isteyen bireylerin girdiği uluslararası kabulü olan sınavlardan olan GRE ve GMAT da bireye uyarlanmış testler olarak uygulanmaktadır (Educational Test Servise, 2023; Luecht & Sireci, 2012). Bu örnekler her geçen gün artmaktadır. Farklı ülkelerde, ölçme kuram ve uygulamalarındaki bu gelişmeler doğrultusunda bilgisayar tabanlı ve bireye uyarlanmış testlerin ulusal olarak uygulanabilirliğini sorgulamak amacıyla ön değerlendirmeler yapıldığı ve gelecek için senaryolar geliştirildiği gözlenmektedir (Breiter et al., 2012).

Bireye uyarlanmış testlerin yaygınlaşmasının altında, testlerin geleneksel test uygulamalarına göre çeşitli avantajları barındırması yatmaktadır. BOBUT uygulamalarının, geleneksel testlere göre bireysel güvenirlik kestirimi, güvenirliği yüksek kestirim, kısa test ve test süresi, test eşitlemeye ihtiyaç duymadan puanların karşılaştırılabilirliği, bireysel test zamanı ve programı yüksek test güvenliği gibi çok sayıda avantajı (Şenel, 2021) sıralanabilir. Ancak Türkiye'de bireye uyarlanmış testler daha ziyade, araştırma projeleri, geniş ölçekli testlerde BOBUT'un uygulanabilirliği (Bulut & Kan, 2012; Kalender & Berberoglu, 2017) ya da küçük ölçekli uygulamalar şeklinde yürütülmektedir (Aybek, 2016; Özbaşı, 2014; Şenel & Kutlu, 2018). BOBUT'un psikometri alanındaki gücüne ve bu alanda ulusal düzeydeki bilimsel birikimin de gün geçtikçe artmasına rağmen, Türkiye'deki test kuruluşlarının bu gelişmeleri test uygulamalarına yansıtmadıkları görülmektedir. Bu araştırmayla, Türkiye'de test kurum ve birimlerinin test uygulamalarında bireye uyarlanmış testlerin tercih edilmesi için duyulan gereksinimlerin neler olduğunun belirlenmesi ve bu konuda halihazırda yürütülen planlamalar ve çalışmalar varsa bunların ortaya konulması amaçlanmıştır.

Yöntem

Araştırma tarama modelindedir. Araştırma verileri, Türkiye'de test kurumları ve birimlerinin uzmanlarıyla ve bu kurumlarda danışmanlık yürüten akademisyenlerle yapılan görüşmelerden elde edilecektir. Verilerin toplanmasında, araştırmacılar tarafından geliştirilecek yarı yapılandırılmış görüşme formu kullanılacaktır. Formda; kurumlardaki farkındalık, yürütülen çalışmaların kapsamı ve özellikleri, uzman ve yazılım gereksinimleri gibi boyutları yoklayan sorular ve sondalara yer alacaktır. Verilerin bulgulara dönüştürülmesinde, betimsel istatistikler ve içerik çözümlemesi kullanılacaktır.

Beklenen Bulgular

Bireye uyarlanmış testlerin güçlü psikometrik özelliklerinin yanında bazı karşılanması güç gereklilikleri bulunmaktadır. Bu testler geleneksel test uygulamalarına göre karmaşık bir algoritması olması nedeniyle ancak bilgisayar tabanlı yazılımlar yoluyla uygulanabilmektedir. Algoritmaları uygulayabilmek için açık kaynak kodlu yazılımlardan yararlanmak mümkün olsa da bu tür yazılımlar

[1] Ankara Üniversitesi
[2] Balıkesir Üniversitesi

test kurumlarının veri saklama, güvenliği ve sürdürebilirlik gereksinimleri karşılayamamaktadır. Bunun yanında madde parametreleri kestirilmiş geniş madde havuzları gerektirmektedir. Bu nedenle, BOBUT uygulamalarını test kurumlarının yürüttüğü testlerde kullanmak için, bireye uyarlanmış testler alanında eğitimi ve deneyimi olan uzmanlarla ve yazılım ekipleriyle çalışmak gerekmektedir. Araştırma sonucunda Türkiye'deki bu gereksinimlerin ortaya konulmasına ve ulusal düzeyde bireye uyarlanmış testlerin kullanımının hızlanmasına kaynaklık edecek verilerin neler olduğunun belirlenmesi beklenmektedir.

KAYNAKÇA

Aybek, E. C. (2016). Kendini değerlendirme envanteri'nin bilgisayar ortamında bireye uyarlanmış test (BOBUT) olarak uygulanabilirliğinin araştırılması. Ankara Üniversitesi.

Breiter, A., Gross, L. M., Stauke, E., Breiter, A., Gross, L. M., Stauke, E., Assessments, C. L., Breiter, A., Groß, L. M., & Stauke, E. (2012). Computer-Based Large-Scale Assessments in Germany. 10th Next Generation of Information Technology in Educational Management, 41–54.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. Egitim Arastirmalari - Eurasian Journal of Educational Research, 49, 61–80.

Educational Test Servise. (2023). GRE General Test Structure.

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey? Educational Sciences: Theory & Practice, 17(2), 573–596. https://doi.org/10.12738/estp.2017.2.0280

Luecht, R., & Sireci, S. (2012). A Review of Models for Computer-Based Testing. College Board Research Reports, 1, 1–56.

Organisation for Economic Co-operation and Development. (2018). PISA 2018 technical report Chapter 2: Test design and test development.

Özbaşı, D. (2014). Bilgisayar Okuryazarlığı Testinin Bilgisayar Ortamında Bireye Uyarlanmış Test olarak uygulanabilirliğine ilişkin bir araştırma. Ankara Üniversitesi.

Şenel, S. (2021). Bilgisayar Ortamında Bireye Uyarlanmış Testler [Computerized Adaptive Testing] (2nd ed.). Pegem Akademi. https://doi.org/10.14527/9786257676861

Şenel, S., & Kutlu, Ö. (2018). Comparison of two test methods for VIS: paper-pencil test and CAT. European Journal of Special Needs Education, 33(5), 631–645. https://doi.org/10.1080/08856257.2017.1391014

Anahtar Kelimeler: bireye uyarlanmış testler, geniş ölçekli test uygulamaları, test geliştirme

Bilgisayarda Bireyselleştirilmiş Test Uygulamalarında Madde Takımı Seçme Yöntemlerinin Karşılaştırılması

Sebahat Gören[1] Selahattin Gelbal[1]

ÖZET

Giriş

Bilgisayarda bireyselleştirilmiş test uygulamaları, test etkililiğini artırması ve test uzunluğu, test süresi vb. konulardaki ekonomikliği nedeniyle eğitim alanında son yıllarda aktif bir şekilde kullanılmaktadır. BBT uygulamaları sayesinde bireyler sadece kendi yetenek düzeylerine uygun test maddelerini yanıtladığından istenilen ölçüm hassasiyeti daha az sayıda maddeyle daha az sürede sağlanabilmektedir (Thompson, 2007; Wang, 2017). Bu avantajların yanı sıra Madde Tepki Kuramı (MTK) temelli BBT uygulamalarında bazı özel durumlarda bazı sorunlarla karşılaşılabilmektedir. Bunlardan biri de ortak bir senaryo veya okuma parçasını paylaşan madde takımlarından oluşan BBT uygulamalarıdır.

Eğitim ve psikoloji alanında özellikle karmaşık bilişsel süreçlerin kullanımını gerektiren okuduğunu anlama becerilerinin ölçülmesinde, okuma pasajlarını paylaşan gruplanmış maddeler madde demeti, madde takımı ya da madde kümesi olarak adlandırılır. (Rosenbaum, 1988). Bu madde takımlarının kullanımı, MTK'deki yerel bağımsızlığın temel varsayımı ihlali nedeniyle BBT uygulamaları için önemli bir sorun teşkil eder. Çünkü yerel bağımlılık, maddeler ortak bir uyaran içerdiğinde (metin, şekil, grafik vb.) ortaya çıkan madde takımlarının ana özelliğidir (Wilson, 1988). Yerel bağımlılığın varlığı, ölçüm kesinliğinin fazla veya eksik tahminine (Braeken, 2011; Marais & Andrich, 2008; Sireci, Thissen, & Wainer, 1991; Wainer & Thissen, 1996; Yen, 1993), madde parametre tahmininde yanlılığa (Tuerlinckx & De Boeck, 2001; Wainer & Wang, 2000) sebep olmaktadır. Madde ve madde takımı bilgi fonksiyonlarının yanlı hesaplanması, BBT uygulamalarında madde seçimini etkilediğinden BBT'nin tüm performansını olumsuz etkiler ve bu da yanlış gizil özellik tahminine neden olur (Chen ve Wang, 2007; Keller vd., 2003; Sireci vd., 1991; Thissen vd.,1989; Yen, 1993). Alan yazında yerel bağımsızlık ihlalinin yol açtığı problemleri ortadan kaldırmak için madde takımlarını çok kategorili puanlanan madde olarak değerlendirme (Thissen vd.,1989; Wainer and Lewis, 1990) ve yerel bağımlılık derecesini dikkate alan Madde Takımı Tepki Kuramını (MTTK) kullanma (Wainer vd., 2007) gibi yaklaşımlar bulunmaktadır.

Bu çalışmada madde takımlarından oluşan BBT uygulamalarında madde seçim yönteminde madde takım etkisini yani yerel bağımsızlık derecesini dikkate alan MTTK ve standart MTK modelleri kullanılarak sonuçlar karşılaştırılacaktır.

Yöntem

Bu çalışmanın amacı, madde takımlarından oluşan BBT çalışmalarında madde takımı seçimi esnasında bağımlılığı ele alan MTTK yaklaşımlarının ve yetenek kestirim yöntemlerinin performanslarını incelemektir. Her bir madde takımı aslında ayrı bir faktör olarak düşünülse de bu çalışmadaki veri yapısı genel bir faktöre bağlı olduğundan tek boyutlu olarak kabul edilecektir. Çalışmanın madde havuzu, özel bir üniversitede hazırlık öğrencilerine uygulanan her biri tek bir pasaja bağlı üç madde içeren 77 madde takımından oluşan toplam 231 maddeyi içerecektir.

[1] Hacettepe Üniversitesi

Gerçek veriden elde edilen parametre kestirimlerinde R programında yer alan “mirt” paketi (Chalmers, 2020) kullanılacaktır. R programında tanımlı BBT uygulamalarında madde takımı modellerini içeren herhangi bir paket olmadığından tüm fonksiyonlar “MASS” paketinde (Rilpert, 2022) yazılacaktır. BBT uygulaması analizleri veri yapısına uygunluğu sebebiyle Rasch modele göre yapılacaktır. Madde takımı seçiminde Fisher’in maksimum bilgi yöntemi kullanılacak olup burada madde takımı bilgisi o madde takımındaki maddelerin bilgi fonksiyonlarının toplamı olarak ele alınacak ve maksimum bilgiyi veren madde takımı seçilecektir. Burada dikkat edilmesi gereken nokta ise madde takımı seçiminde madde bilgi fonksiyonu hesaplanırken bireye ait madde takımı etkilerinin hesaba katılmamasıdır. Yetenek kestirimi için EAP ve MAP kullanılacak olup burada MTK ve MTTK’ye göre yetenek kestirilecektir. Sonlandırma kuralı olarak da sabit test uzunluğuna (n=12 ve n=24) dayalı sonlandırma kuralı kullanılacaktır. Sonuç olarak BBT simülasyonu, 2 madde seçim yöntemi (MTK, MTTK) X 2 yetenek kestirimi (EAP, MAP) X 2 sonlandırma kuralı (n=12, n=24) olmak üzere 8 koşulu içerecektir.

BBT uygulaması simülasyonu için öncelikle başlangıç noktası olarak theta = 0 ve γ=0 kabul edilecektir. Ardından verilen theta ya göre maksimum fisher yöntemi ile havuzdan bir madde takımı seçilecektir. Madde seçim yöntemine göre yetenek kestirimi yapılacaktır. Bu adımlar tekrarlanacak ve belirlenen test uzunluğuna ulaşıldığında test sonlandırılacaktır. Değerlendirme, yanlılık (bias), RMSE ve ortalama hata değerleri hesaplanarak yapılacaktır.

Beklenen Sonuçlar

Özellikle madde takımı içeren BBT uygulamalarında yerel bağımlılığın bir ölçüsü olarak görülen madde takımı etkisinin varyans değerinin yüksek çıkması beklenmemektedir. Çünkü alan yazında gerçek verilerle yapılan çalışmalarda bu değerin düşük ya da orta derecede çıktığı görülmüştür (Yılmaz-Koğar, 2016 ;Yamamoto et al., 2018).Ayrıca madde seçme yöntemi olarak MTTK kullanıldığında madde takımlarının yerel bağımlılık dereceleri de hesaba katıldığından daha az hata ile daha doğru ve hassas yetenek kestirimi yapılması beklenmektedir (Erşan, 2022; Keng, 2008). Fakat buradaki durum yerel bağımsızlık derecelerine bağlı olacaktır. Yani madde takımı etki parametre değeri düşük çıkarsa madde seçme yöntemleri arasında herhangi bir farklılık meydana gelmeyebilir. Ayrıca madde takımlarındaki madde sayısının az olması da farklılıkların küçük çıkmasına sebep olabilir. Yetenek kestiriminde MAP yönteminin, EAP yönteminden daha iyi sonuç vermesi beklenmektedir. Alan yazında az sayıda bulunan madde takımı temelli BBT uygulama çalışmaları; yerel bağımlılık derecelerini içeren farklı simülasyon koşulları ve madde seçme yöntemlerine göre içerik dengeleme, madde kullanım sıklığı kontrolü gibi kısıtlamalar altında da çalışmalar yapılabilir.

Kaynakça

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. Psychometrika, 76, 57-76.

Chalmers, P. (2020). mirt: an R package for multidimensional item response theory (v.1.33.2). https://CRAN.R-project.org/package=mirt

Chen, C.–T. & Wang, W.–C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. Applied Psychological Measurement, 31(5), 388–411.

Ersan Cinar, O. (2022). The Impact of Local Item Dependence on Computer Adaptive Testing given Between and Within Testlet Adaptivity.

Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. Applied Measurement in Education, 16(3), 207–222.

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

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas, 9(3), 200-15.

Rosenbaum, P. R. (1988) A note on item bundles. Psychometrika, 53(3), 349–359.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237– 247.

Tuerlinckx, F. & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. Psychological Methods, 6(2), 181–195.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. Journal of Educational Measurement, 26, p.247-260.

Thompson, N. A. (2009). Item selection in computerized classification testing. Educational and Psychological Measurement, 69(5), 778-793. doi: 10.1177/0013164408324460

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence of reliability? Educational Measurement: Issues and Practice, 16, 22-29.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C.A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 245–269). Kluwer Academic Publishers.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). Testlet response theory and its applications. New York, NY: Cambridge University Press.

Wang, K. (2017). A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing (Doctoral Dissertation). Michigan State University.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187–214.

Anahtar Kelimeler: bilgisayarda bireyselleştirilmiş test, madde takımı, madde takımı tepki kuramı,madde takımı etkisi

Madde Takımı Tabanlı BOBUT: Madde Takımları Arası ve İçi Uyarlanabilirlik

Özge Erşan Çınar[1]

ÖZET

Madde takımı tek bir uyaranla ilişkili olan test maddelerini ifade eder (Wainer & Kiely, 1987). Yaygın örnekler arasında bir okuma testinde bulunan birden fazla maddenin takip ettiği okuma metinleri yer alır.

Yerel bağımsızlık, Madde Tepki Kuramı' nın temel bir varsayımıdır. Yerel bağımsızlıkta bireylerin maddelere verdiği yanıtlar yalnızca gizil değişkenin bir fonksiyonudur. Ancak, Madde Tepki Kuramı'na dayalı bilgisayar ortamında bireye uyarlanmış testler (BOBUT) içinde madde takımları kullanıldığında, gizil değişken göz önüne alındıktan sonra bile bu maddeler ortak bir uyaranla ilişkilendirildiği için aynı madde takımındaki maddelere verilen yanıtlar bağımsız olmayabilir. Bu durumda yerel bağımsızlık varsayımı ihlal edilmiş olur ve bu ihlal BOBUT' un tüm performansını etkileyerek hatalı test puanlarının hesaplanmasına neden olabilmektedir (Thissen vd., 1989; Yen, 1993; Keller vd., 2003; Murphy vd., 2010; Yao, 2019).

Testlet Tepki Kuramı modelleri, araştırmacıların yerel bağımlılığı ele almalarına izin verir (Bradlow vd., 1999). Ancak önceki araştırmacılar BOBUT' ta madde seçimi sırasında madde takımı bağımlılığını hesaba katmamışlardır, bu da sınava girenlerin yeteneklerinin yanlı tahmin edilmesine yol açmaktadır (Boyd vd., 2013; Keng, 2008; Murphy vd., 2010; Yao, 2019).

Madde takımlarında yaygın uygulama madde takımının uyarlamalı bir şekilde seçilmesi ve madde takımı içinde kümelenmiş tüm maddeleri uygulamaktır. Örneğin, bir okuma testinde beş madde takımı varsa uyarlama yalnızca dört kez madde takımları arasında gerçekleşir (Glas vd., 2000).

Madde takımı tabanlı BOBUT' da doğru ve verimli yetenek kestirimi için Keng (2008) ve Ma (2020) madde takımlarını uyarlamalı olarak seçme ve seçilen madde takımları içinden de maddeleri uyarlamalı olarak seçme üzerinde çalıştılar. Ancak bu çalışmalarda madde takımı seçiminde ve puanlamada koşullu bağımlılık tümden ya da kısmi olarak göz ardı edilmiştir.

Yöntem

Simulasyon Koşulları

Bu çalışmada yetenek parametresinin ne ölçüde doğru kestirildiği bir simülasyon çalışması ile incelenmiştir. Aşağıda incelenen simülasyon koşulları verilmiştir.

Uyarlanabilirlik: (1) madde takımları arası ve (2) madde takımları içi uyarlanabilirlik.

Madde takımı seçimi yaklaşımı: Madde takımı seçimi ve seçilecek ilk madde için (1) MTK yaklaşımının kullanılması (IRT), (2) tüm bireyler için madde takımı etkisinin 0 kabul edildiği Keng (2008) yaklaşımının kullanılması (TRT), (3) Fisher bilgi fonksiyonunun marjinalinin alındığı Ip (2010) yaklaşımının kullanılması (TRT-m).

Gizil değişken kestirim metodu: expected a posteriori (EAP), maximum a posteriori (MAP).

Yerel madde bağımlılığının büyüklüğü: Madde takımı varyansı (1) düşük bağımlılık için 0,01 ile 0,50 arasında, (2) yüksek bağımlılık için 1,00 ile 1,50 arasında değişmektedir.

[1] MEB

Test uzunluğu/madde takımı sayısı: (1) 20 madde/4 madde takımı, (2) 40 madde/4 madde takımı, (3) 40 madde/8 madde takımı.

Simülasyon Verilerinin Üretimi

Gerçek gizil değişken değerleri 0,5 birim eşit aralıklarla -3.5 ve 3,5 arasında üretildi ve her değer 100 kez tekrar edildi. Üretilen madde havuzu, her biri 25 maddelik 40 madde takımından (toplam 1000 madde) oluşacak şekilde üretildi. Bu çalışmada, madde takımları içinde yer alan madde sayıları hem madde takımı arası hem de madde takımı içi uyarlamalarda eşit olarak tutulmuştur. Bu nedenle, madde takımı arası uyarlamalı BOBUT için, 40 maddelik madde takımı hedeflenen sayıda maddelerden oluşan mini paralel formlara bölünmüş ve her defasında bu mini formlardan biri rastgele seçilerek uygulanmıştır.

Simülasyon Süreci

Simülasyon algoritması aşağıdaki gibidir:

1. Başlangıçta tüm simüle edilmiş bireyler için ölçülen gizil değişken sıfır kabul edildi.
2. TRT yaklaşımıyla madde takımı seçerken, bireylerin madde takımı parametresi sıfır kabul edildi.
3. Maximum Fisher Bilgisi yöntemi kullanılarak mevcut gizil değişken kestirimine bağlı olarak madde havuzundan bir madde takımı seçildi. TRT-m yaklaşımında, madde bilgi fonksiyonlarını hesaplamak için marjinal madde kullanıldı (Ip, 2010). TRT yaklaşımında, bireylerin madde takımı parametresi madde takımı seçimi aşamasında sıfır kabul edildi.

Madde takımı arası-ve-içi uyarlamalı BOBUT:

4. Bir madde takımı seçildikten sonra, bu madde takımı içinden bir madde seçildi. Bu adım için, maksimum bilgi sağlayan madde seçildi ve uygulandı.

5. Simüle edilmiş birey tarafından seçilen madde takımı içinden uygulanan maddelere verilen yanıtlar üretilmiş madde yanıt verisi matrisinden alındı.

6. Gizil değişken ve birey madde takımı parametresi kestirildi.

7. 4'ten 6'ya kadar olan adımlar belirlenen her bir madde takımında yer alacak madde sayısı tamamlanana kadar tekrar edildi.

8. 4'ten 7'ye kadar olan adımlar testte belirlenmiş madde sayısına ulaşana kadar tekrar edildi.

Madde takımı arası uyarlamalı BOBUT:

4. Madde takımına karar verildiğinde, çalışılan madde takımı için oluşturulmuş paralel testlerden biri rastgele seçildi.

5. Seçilen madde takımı için madde yanıtları, oluşturulan madde yanıt veri matrisinden elde edilmiştir.

6. Gizil değişken ve birey madde takımı parametresi kestirildi.

7. 4'ten 6'ya kadar olan adımlar testte belirlenmiş madde sayısına ulaşana kadar tekrar edildi.

Sonuçlar

Sonuçlar, BOBUT'da madde takımları içinde uyarlanabilirliğin kullanılmasının kesinlik ve verimlilik üzerindeki etkileri hakkında bilgi sağlamaktadır.

Daha ayrıntılı bulgular aşağıda sunulmuştur:

1. Madde takımları arası-ve-içi uyarlamalı BOBUT, madde takımları arası uyarlamalı BOBUT' a göre gizil değişken kestirimini daha gerçek değerlere daha yakın yapmıştır ve aynı zamanda tüm simülasyon koşulları altında ölçmenin standart hatası daha küçük olarak hesaplanmıştır.

2. Madde takımı sayısını artırmak madde takımı uzunluğunu artırmaya kıyasla gizil değişken kestiriminin gerçek değerlere daha yakın olmasını ve ölçmenin standart hatasının daha küçük olarak hesaplanmasını sağlamıştır.

3. Madde takımı bağımlılığı yüksek olduğunda, tüm simülasyon koşulları altında, gerçek ve kestirilen gizil değişkenler arasındaki farkın daha büyük olduğu ve ölçmenin standart hatasının daha büyük olarak hesaplandığı görülmüştür.

4. Tüm simülasyon koşulları altında TRT ve TRT-m yaklaşımları arasında dikkate değer bir farklılık gözlenmemiştir. Her iki yaklaşım da IRT yaklaşımından daha iyi performans göstermiştir.

5. Gizil değişken kestirim yöntemlerinden MAP özellikle test uzunluğu daha kısa olduğunda veya madde takımı bağımlılığı yüksek olduğunda EAP'den biraz daha iyi performans göstermiştir.

6. Özetle, bu çalışma hem madde takımları arası hem de içi için uyarlamanın madde takımı tabanlı BOBUT' un performansını artırdığını göstermiştir.

Kaynakça

Boyd, A. M., Dodd, B., & Fitzpatrick (2013). A comparison of exposure control procedures in CAT systems based on different measurement models for testlets. *Applied Measurement in Education, 26*(2), 113–135.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168.

Glas, C.A.W., Wainer, H., & Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Boston, MA: Kluwer Academic Publishers.

Ip, E. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34*(7), 467–482.

Keller, L.A., Swaminathan, H., & Sireci, S.G. (2003). Evaluating scoring procedures for contextdependent item sets. *Applied Measurement in Education, 16*(3), 207–222.

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

Ma. Y.C. (2020). Investigating hybrid test designs in passage-based adaptive tests (Unpublished doctoral dissertation). University of Iowa, Iowa City, Iowa.

Murphy, D.L., Dodd, B.G., & Vaughn, B.K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement, 34*(6), 424–437.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple categorical-response models. *Journal of Educational Measurement, 26*, 247–260.

Yao, L. (2019). Item selection methods for computer adaptive testing with passages. *Frontiers in Psychology*, *10*, 1–10.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–214.

Anahtar Kelimeler: CAT, testlet

Bilgisayarda Bireyselleştirilmiş Sınıflama Testlerinde Madde Kullanım Sıklığı Kontrol Yönteminin Çok Kategorili Sınıflamada Test Etkililiğine ve Ölçme Kesinliğine Etkisi

Demet ALKAN[1] Sebahat GÖREN[2] Nuri DOĞAN[2]

ÖZET

Giriş

Sınıflandırma testi prosedürlerinin amacı, bir sınav katılımcısını önceden belirlenmiş bir kesme puanına göre değerlendirmek ve kategorik bir sonuç elde etmektir (Wainer, 1990;Weiss, 1983). Bireyselleştirilmiş bilgisayarda sınıflama testleri (BBST), özellikle değerlendirme sonuçları yüksek riskli olduğunda sınıflandırma kararı vermek için kağıt kalem testlerine tercih edilebilir. Ayrıca bu testlerde kullanılan yetenek belirleme yöntemleriyle daha az sayıda madde ile daha doğru sınıflamalar elde edilebilir (Lewis & Sheehan, 1990).

BBST uygulamaları psikometrik model, kalibre edilmiş madde havuzu, başlama noktası, madde seçme yöntemi ve sınıflama kriterleri olmak üzere beş temel bileşenden oluşmaktadır (Thompson,2007). Bu temel bileşenlerin yanında daha az madde ile daha geçerli ve güvenilir ölçümler yapabilmek için madde kullanım sıklığı gibi pratik kısıtlamaların dikkate alınması önemlidir. Madde kullanım sıklığı kontrolü BBST' leri için oldukça önemlidir. Madde kullanım sıklığındaki artış testlerin geçerliğini düşürmektedir bu da o maddenin bireyin yeteneğini değerlendirmesinde daha az etkili olmasına sebep olur (Barret,2015). Madde kullanım sıklığı kontrolünün temel amacı test maddelerinin güvenliğini sağlamaktır (Lin, 2011). BBST çalışmalarında kullanılan madde kullanım sıklığı, kontrolü istenen maksimum madde kullanım sıklığı oranı (r max) belirlenerek havuzdaki her maddenin sıklığını bu orana eşitlemekle ya da bu orandan daha düşük tutmakla sağlanmaya çalışılır. En sık kullanılan madde kullanım sıklık yöntemleri ise Sympson-Hetter yöntemi (SH; Sympson ve Hetter, 1985) ve madde uygunluk yöntemidir (van der Linden &Veldkamp,2004). Bu araştırmada Bilkent üniversitesi yabancı dil yüksekokulunda 919 öğrenciye uygulanan İngilizce seviye belirleme sınavına ait 256 madde içeren madde havuzundaki gerçek veri seti ile BBST uygulamalarında kullanılan farklı madde kullanım sıklığı yöntemlerinin iki üç ve dört kategorili sınıflamada beklenen sonsal dağılım yetenek kestirim yöntemi ile farklı sınıflama kriterleri ve madde seçme yöntemlerinde ölçme kesinliği ve test etkililiğine etkisi araştırılacaktır.

Yöntem

Bu çalışma betimsel araştırma ve gerçek veri ile yapılan simülasyon çalışmasıdır. Araştırmada yetenek kestirimi olarak beklenen sonsal dağılım; madde seçme yöntemleri ise Maksimum Fisher Bilgi kestirilen yetenek ve kesme noktası temelli yöntemler; sınıflama kriterleri ağırlıklı olabilirlik oran testi ve güven aralığı yöntemi ile madde kullanım sıklık yöntemleri için Sympson-Hetter yöntemi (SH; Sympson ve Hetter, 1985) ve madde uygunluk yöntemleri madde kullanım sıklık oranı olarak Huebner (2012) ile Li'nin (2012) çalışmaları dikkate alınarak rmax = 0,20 (Leung ve diğerleri, 2002) kullanılacaktır. Gerçek veri setinden oluşturulan madde havuzundan 24 koşul için gerçek zamanlı tek boyutlu veri ile madde kullanım sıklık yönteminin test etkililiğine, ölçme kesinliğine etkisi ve madde kullanım sıklığını aşan maddelerin yüzdesi (MKSAMY) ile bu maddelerin madde kullanım sıklıklarının

[1] MEB
[2] Hacettepe Üniversitesi

ortalaması (MKSAO) belirlenecektir. Araştırmada İki, üç, dört kategorili sınıflama için oluşturulan 24 koşul (2 sınıflama kriteri X 2 madde seçme yöntemi X 2 madde kullanım sıklığı yöntemi X 3 sınıflama kategori sayısı) için yapılan analizler 25 tekrar ile tekrarların ortalaması alınarak R yazılımı kullanılarak oluşturulacaktır. Test etkililiği için ortalama sınıflama doğruluğu, ortalama test uzunluğu değerleri hesaplanacaktır. Ölçme kesinliği için RMSE, ortalama mutlak hata, yanlılık, gerçek yetenekler ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) hesaplanacaktır. Gerçek yetenek düzeyi ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) için Pearson korelasyon katsayısı değeri hesaplanacaktır. Ortalama sınıflama doğruluğu için gerçek sınıflar ile simülasyon sonucu hesaplanan sınıflar arasındaki uyum Cohen Kappa istatistiği ile hesaplanacaktır.

Araştırma Sorusu

İngilizce seviye belirleme sınavına ait gerçek veri seti ile gerçekleştirilen çok kategorili BBST uygulamasında farklı madde kullanım sıklık kontrolü ve farklı madde seçme yöntemlerinde test etkililiği, ölçme kesinliği ve maksimum madde kullanım sıklığını aşan maddelerin yüzdesi (MKSAMY) ile bu maddelerin madde kullanım sıklıklarının ortalaması (MKSAO) nasıl değişmektedir?

Alt Problemler

1. AOOT ve GA sınıflama kriterleri ile iki, üç, dört kategorili sınıflama için SH ve MU madde kullanım sıklığı kontrol yöntemlerine göre test etkililiği, ölçme kesinliği, MKSAMY, MKSAO nasıl değişmektedir?

2.MFB-KY, MFB-KN madde seçme yöntemleri ile SH ve MU madde kullanım sıklığı kontrolü yöntemlerinin çaprazlandığı iki, üç, dört kategorili sınıflamalar için ölçme kesinliği, test etkililiği, MKSAMY ve MKSAO nasıl değişmektedir?

3.MFB-KY, MFB-KN madde seçme yöntemleri ile AOOT ve GA sınıflama kriterlerinin çaprazlandığı koşullarda iki, üç, dört kategorili sınıflamalar için madde kullanım sıklık yöntemlerinin etkisi olmadan test etkililiği ve ölçme kesinliği nasıl değişmektedir?

Beklenen Bulgular

AOOT sınıflama kriteri kullanıldığında, MU madde kullanım sıklığı kontrol yönteminin SH' ye kıyasla test etkililiğini olumsuz etkileyeceği düşünülmektedir. Alan yazında ise MU yöntemi madde kullanım sıklığını daha iyi kontrol etmesine rağmen test etkililiği üzerinde SH'ye kıyasla daha olumsuz bir etkiye sahip olduğu, MU'nun tercih edildiği koşullarda SH'ye kıyasla daha yüksek OTU ve daha düşük OSD'ler elde edileceği beklenen bulgulardandır (Huebner, 2012).

SH yöntemi GA sınıflama kriteriyle birlikte kullanıldığında daha az sayıda madde kullanım sıklığı oranını aşmasına rağmen aynı maddelerin çok fazla kullanılmasından dolayı MKSAMYO değerinin yüksek olması beklenilmektedir. Madde kullanım sıklığı MU ile kontrol edildiğinde, AOOT ve GA yöntemleri için birbirine benzer MKSAMO değerleri ve SH'ye kıyasla daha düşük MKSAMY değerleri beklenilmektedir. Sınıflama kategori sayısı arttıkça test etkililiğinin azaldığı çalışmalar alan yazında mevcuttur (Eggen,1999; Nydick vd. 2012).

MFB-KY, MFB-KN madde seçme yöntemlerinin ele alındığı koşullarda MU yöntemi SH yöntemine kıyasla madde kullanım sıklığı kontrolü bakımından daha iyi bir performans göstereceği beklenmektedir.

AOOT ve GA sınıflama kriterleri için tüm madde seçme yöntemlerinin olduğu koşullarda AOOT'nin GA'ya kıyasla daha iyi performans göstereceği beklenmektedir (Nydick vd, 2012;Thompson, 2011).

Kaynakça

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. Applied Psychological Measurement, 23, 249-261. doi: 10.1177/01466219922031365

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson–Hetter algorithm. Applied Psychological Measurement, 26(4), 376-392. doi: 10.1177/014662102237795

Lewis, C., & Sheenan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. Applied Psychological Measurement, 14(4), 367-386. doi: 10.1177/014662169001400404

Lin, C. (2011). Item selection criteria with practical constraints for computerized classification testing. Applied Psychological Measurement 71(1), 20-36. doi: 10.1177/0013164410387336

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, Nisan). Accuracy and efficiency in classifying examinees using computerized adaptive tests: An application to a large scale test. The National Council on Measurement in Education (NCME) toplansında sunulan bildiri, Vancouver, British Columbia, Canada. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3381&rep=rep1&type=pdf adresinden erişilmiştir

Sympson, J.B., & Hetter, R.D. (1985, Ekim). Controlling item exposure rates in computerized adaptive testing. 27th Military Testing Association toplantıısında sunulan bildiri, 937-977. San Diego, CA: Navy Personnel Research andDevelopment Center. http://www.iacat.org/content/controlling-item-exposure-rates-computerizedadaptive- testing adresinden erişilmiştir.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. Practical Assessment Research & Evaluation, 12(1), 1-13. http://www.iacat.org/sites/default/files/biblio/th07-01.pdf adresinden erişilmiştir.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. Practical Assessment, Research & Evaluation, 16(4), 1-7. https://pareonline.net/getvn.asp?v=16&n=4 adresinden erişilmiştir

Van der Linden, W.J., & Veldkamp, B.P. (2004). Constraining Item Exposure in Computerized Adaptive Testing With Shadow Tests. Journal of Educational and Behavioral Statistics, 29(3), 273-291. https://www.utwente.nl/nl/bms/omd/Medewerkers/artikelen/vdLinden/JEBS%202004%2C%20273-291-1.pdf adresinden erişilmiştir.

Anahtar Kelimeler: Bilgisayarda bireyselleştirilmiş sınıflama testi, madde kullanım sıklığı, test etkililiği, ölçme kesinliği

Çok Aşamalı Testlerin Lineer Testlerle Birlikte Kullanımının Farklı Koşullar Altında İncelenmesi

Çağla Alpayar[1] Deha Doğan[1]

ÖZET

Giriş

Geniş ölçekli test uygulamaları giderek bilgisayar ortamına aktarılmaktadır (Shin, Yamamoto, Khorramdel, & Robin, 2021). Özellikle çeşitli kısıtlamalar dikkate alınması gerektiğinde çok aşamalı testler (ÇAT), kullanışlı tasarımlar sunmaktadır (Yamamoto, Shin ve Khorramdel, 2018) ve giderek yaygınlaşmaktadır (Yan, Lewis, & von Davier, 2014). ÇAT'da yanıtlayıcının yanıtlanacağı maddeler tek tek değil modül adı verilen madde grubu olarak seçilir ve bir önceki modüldeki performansına bağlı olarak sonraki aşamaya yönlendirilir (Yan, von Davier, & Lewis, 2014). Çok aşamalı testlerin birleştirme literatürü incelendiğinde, sıklıkla düşük-orta ve yüksek yetenek düzeyi (-θ,+ θ) ranjının hedef alındığı görülmektedir (van der Linden, & Diao, 2011). Ancak Türkiye'nin hem ulusal hem uluslararası geniş ölçekli test uygulamalarının sonuçları, öğrencilerin orta ve düşük yeterlilik düzeylerinde yoğunlaştığını göstermektedir. Örneğin, PISA 2018 uygulamasında, Türkiye'deki öğrencilerin yaklaşık %36'sı matematik alanında temel yeterlilik düzeyinin altında yer almıştır. 2.-4. yeterlilik düzeyinde ise sırasıyla %27, %20 ve %11'i yer almaktadır. Türk öğrencilerin %0.9'u ise en üst yeterlilik düzeyindedir. Fen ve okuma alanlarında da benzer dağılımlar gözlemiştir (OECD, 2019). Söz konusu dağılımı gösteren yanıtlayıcı grubu için çok aşamalı testlerin uygulandığı bir sınavda, (-θ,+ θ) aralığında yaygın kullanılan desen yerine hedef bilgi değerleri orta yetenek düzeylerinde maximize edilmiş modüllerin uygulanması, daha yüksek kesinlikte ölçümlere olanak sağlayabilir.

Avusturya'nın ulusal sınavı olan NAPLAN'de alternatif bir test tasarımı geliştirilmiştir. NAPLAN, Avusturya üçüncü, beşinci, yedinci ve dokuzuncu sınıf öğrencilerine uygulanan bir matematik bilgisi ve okuryazarlık değerlendirmesi sınavıdır (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2017). Sınavda, yaygın olanın dışında bir çok aşamalı test deseni uygulanmaktadır. Buna göre uygulamadan önce modüller özellikle orta yetenek düzeyinde maksimum bilgiyi verecek, modüller arasında çeşitli geçişler içerecek ve düşük yeterlik düzeyindeki öğrenciler için alternatif yollar tanımlayacak şekilde bir çok aşamalı test tasarımı kurgulanmıştır (ACARA, 2022). Böylelikle, her bir öğrenci için testin süresi artırılmadan, hedeflenen zorlukta daha geniş bir soru yelpazesi kullanarak öğrencilerin değerlendirilmesine olanak tanınmaktadır (ACARA, 2017). Bu bilgiler ışığında farklı yetenek ranjlarını hedef alan modüllerden oluşmuş çok aşamalı test bir performanslarının çeşitli koşullar altında karşılaştırılması amaçlanması anlamlı bilgiler sunacaktır.

Yöntem

Araştırma Modeli: Simülasyon yöntemi ile üretilecek veri, farklı hedef bilgi değerlerine yönelik yapılandırılmış modüllerden oluşan farklı test desenleri ve çeşitli test uzunluklarında karşılaştırılacaktır.

Verilerin Üretilmesi: En yaygın kullanılan ÇAT tasarımları iki veya üç aşama içermektedir (Yan, 2020; Zenisky, Hambleton ve Luecht, 2009); bu çalışmada da NAPLAN tasarımın orijinaline sadık kalarak temel olarak 1-3-3 deseni tercih edilecektir. Hedef bilgi düzeyinin belirlenmesinde "mixed integer programming" yaklaşımı izlenecek ve minimax yöntemi uygulanacaktır. Madde çakışmasını önlemek

[1] Ankara Üniversitesi

ve sabit modül uzunluğu sunmak üzere modül düzeyinde sınırlılıklar tanımlanması nedeniyle “bottom up” yöntemine başvurulacaktır. Modüllere yönlendirmede norm dayanaklı ölçüt uygulanacaktır (Zenisky & Hambleton, 2014).

Desen-1: NAPLAN uygulamasında izlenen yaklaşımla, hedef fonksiyonunun tanımlanmasında her aşamadaki modüllerde farklı değerler ele alınacak. başlangıç modülü (Modül-A), (3.5θ, +θ) aralığında bilgiyi maksimize edecek şekilde yapılandırılacaktır. İkinci aşamada, -0.5 θ düzeyinin altındaki bireyler lineer bir teste yönlendirilecek ve doğrudan Modül-C (-3.5 θ, -0.5 θ) ve Modül-B (-2θ, .5θ)’yi yanıtlayacaktır. Diğer yanıtlayıcılar ise 1-2-3 geleneksel ÇAT desenini izleyecek ve hitap ettiği yetenek aralıklarında belli oranlarda çakışmalarla (-2, +2) θ aralığında maksimum bilgiyi verecek şekilde yapılandırılmış üç farklı modüle yönlendirilecektir.

Desen-2: 1-3-3 ÇAT deseni uygulanacaktır. Başlangıç modülü (-θ, + θ) aralığında maksimum bilgiyi verecek şekilde yapılandırılacaktır (SS=0.5). Sonraki aşamalarda ise modüller sırasıyla (-1, 0, + 1) θ değerlerini merkeze alarak önceden birleştirilecektir (SS=0.2) (Becker, Debeer, Sachse, & Weirich, 2021; Becker, & Debeer, 2022; Diao, & van der Linden, 2011).

1000 kişilik (-3.5, 3.5) aralığında uniform dağılan yetenek parametreleri üretilecektir.

Verilerin Çözümlenmesi

Ölçme kesinliği, RMSE, mutlak yanlılık, standart hata ve koşullu RMSE değerleri üzerinden yorumlanacaktır.

Beklenen Bulgular

Desen-1, düşük yeterlilik düzeyindeki yanıtlayıcılar özel bir yola yönlendirilecek ve orta yeterlilik düzeyindeki yanıtlayıcılar için hedef bilgi fonksiyonu orta ve alt yetenek ranjını hedef alarak yapılandırılmış birden çok modülü içerecektir. Bu yaklaşımla test yapılandırmada, hedef grubu ayırt etmeden daha etkili bir test yapılandırılacaktır. Böylelikle; uyarlanabilirlik, pratiklik, ölçüm hassasiyeti ve test formları üzerinde kontrol sunması sayesinde lineer test ile madde düzeyi adaptif testler arasında bir denge moldeli sunan ÇAT’ın (Zenisky, Hambleton, & Luecht, 2009) daha adaptif olmasına olanak verecektir çünkü aşama sayısı ve aşamalar içindeki modüllerin zorluk çeşitliliği arttıkça, bu durum testin uyarlanabilirliğini ve esnekliğini olumlu yönde etkilemektedir (Hendrickson, 2007; Yan, Lewis & von Davier, 2014). Buna bağlı olarak, farklı hedef bilgi değerlerine yönelik yapılandırılmış modüllerden oluşan farklı test desenlerinin ölçme kesinliği daha yüksek ölçümler sunması beklenmektedir.

Kaynakça

Australian Curriculum Assessement and Reporting Authority. (2017). The Australian National Assessment Program Literacy and Numeracy (NAPLAN) Assessment Framework: NAPLAN Online 2017-2018.

Australian Curriculum, Assessment and Reporting Authority (2022). National Assessment Program – Literacy and Numeracy 2021: Technical Report, ACARA, Sydney.

Becker, B. & Debeer, D. (2022). eatATA: Create constraints for small test assembly problems. https://cran.r-project.org/web/packages/eatATA/index.html adresinden alınıştır.

Becker, B., Debeer, D., Sachse, K. A., & Weirich, S. (2021). Automated test assembly in R: The eatATA package. Psych, 3(2), 96-112. https://doi.org/10.3390/psych3020010.

Diao, Q., & van der Linden, W.J. (2011). Automated test assembly using lp_solve version 5.5 in R. Applied Psychological Measurement. 35, 398–409. https://doi.org/10.1177/0146621610392211

Hendrickson, A. (2007). An NCME instructional module on multistage testing. Educational Measurement: Issues and Practice, 26(2), 44-52. Https://doi.org/10.1111/j.17453992.2007.00093.x

Magis, D., Yan, D., & von Davier, A. (2018). mstR: Procedures to generate patterns under multistage testing. R package version 1.2. https://cran.r-project.org/web/packages/mstR/index.html adresinden alınmıştır.

OECD (2019a), PISA 2018 Results (Volume I): What Students Know and Can Do, PISA, OECD Publishing, Paris, https://doi.org/10.1787/5f07c754-en.

Shin, H. J., Yamamoto, K., Khorramdel, L., & Robin, F. (2021). Increasing measurement precision of PISA through multistage adaptive testing. In Quantitative Psychology: The 85th Annual Meeting of the Psychometric Society, Virtual (pp. 325-334). Springer International Publishing.

van der Linden, W. J., & Diao, Q. (2011). Automated test-form generation. Journal of Educational Measurement, 48(2), 206–222. https://doi.org/10.1111/j.1745-3984.2011.00140.x.

Yamamoto, K., Shin, H., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. Educational Measurement: Issues and Practice, 37, 16–27.

Yan, D. (2020). Multistage Testing in Practice. In H. Jiao, & R. W. Lissitz (Eds.), Application of Artificial Intelligence to Assessment, (pp.141-160). USA: Information Age Publication.

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von da Vier & C. Lewis (Eds.), Computerized multistage testing: Theory and applications, (pp. 3-20). London, England: Chapman & Hall.

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. Von da Vier & C. Lewis (Eds.), Computerized multistage testing: Theory and applications, (pp. 3-20). London, England: Chapman & Hall.

Zenisky, A. L., & Hambleton, R. K. (2014). Multistage test designs: Moving research results into practice, In D. Yan, A. A. von davier & C. Lewis (Eds.), Computerized multistage testing: Theory and applications (pp. 21–37). London, England: Chapman & Hall.

Zenisky A. L., Hambleton R.K. & Luecht R.M. (2009). Multistage testing: Issues, designs, and research. In van der Linden W., Glas C. (Eds) Elements of Adaptive Testing. Statistics for Social and Behavioral Sciences. Springer, New York, NY. https://doi.org/10.1007/978-0-387-85461-818 .

Anahtar Kelimeler: Çok Aşamalı Test, Lineer Test

Rastgele ve Ardışık Madde Parametre Sapmasının Bilgisayar Ortamında Bireye Uyarlanmış Testlerde Test Etkililiği Üzerindeki Etkisi

Merve ŞAHİN KÜRŞAD[1] Seher YALÇIN[2]

ÖZET

Giriş

Bu çalışmanın amacı, rastgele madde parametre sapması (random item parameter drift-R-MPS) ve ardışık madde parametre sapmasının (compound item parameter drfit- A-MPS) bilgisayar ortamında bireye uyarlanmış testlerde (BOBUT) test etkililiğine etkisini incelemektir. Madde tepki kuramının (MTK) önemli varsayımlarından olan değişmezlik varsayımına göre madde parametre değerleri, testi alan tüm örneklemlerde benzer olmalıdır. Ancak bazı araştırmalar göstermektedir ki bazı durumlarda madde parametreleri zamana bağlı olarak değişmektedir (Bock vd., 1988; Goldstein, 1983; Risk, 2015). Madde parametrelerinin zamana bağlı değişim göstermesi madde parametre sapması (MPS) olarak adlandırılmaktadır (McCoy, 2009). MTK temelli yapılan yetenek kestirimlerinde MPS'nin etkisinin önemli düzeyde olmadığı belirtilse de bir maddenin iki veya daha fazla uygulamada madde parametrelerinin birebir aynı kalmasının da güç olduğu belirtilmektedir (Wollack vd., 2006). Bu durum, BOBUT uygulamalarında da araştırmacıların karşısına çıkmaktadır (Chan vd., 1999; Risk, 2015; Rupp ve Zumbo, 2003; Wells vd., 2002). BOBUT uygulamaları geniş bir madde havuzu gerektirmekle birlikte, madde havuzundaki bazı maddelerin çok fazla uygulanması MPS'ye neden olabilmektedir. BOBUT uygulamalarında MPS görülmesi durumunda hem madde hem yetenek parametreleri etkilenmektedir (Deng ve Melican, 2010). Bu durum da BOBUT uygulamalarında MPS'nin varlığının incelenmesini gerektirmektedir. MPS, rastgele veya ardışık olarak ortaya çıkabilmektedir. Rastgele MPS (R-MPS), bazı maddelerin madde güçlük parametrelerinin bazı uygulamalar arasında değişirken, diğer uygulamalar arasında sabit kalmasını ifade etmektedir. Bir başka ifade ile madde güçlük parametrelerinde uygulamalar arasında sabit bir azalma veya artma yoktur. Ardışık madde parametre sapması (A-MPS) ise madde parametrelerinin uygulamalar arasında sistematik olarak değişmesidir. Bir başka ifade ile madde güçlük parametrelerindeki değişim, her uygulama arasında aynıdır (ya aynı oranda artış ya da aynı oranda azalma vardır). Bu iki durumun test eşitleme çalışmalarında madde yetenek kestirimlerini etkilediği bazı araştırmalarla ortaya konmuştur (Wollack vd., 2006). MPS'nin BOBUT uygulamalarında madde ve yetenek kestirimlerini etkilemesinden dolayı, bu çalışma kapsamında R-MPS ve A-MPS'nin BOBUT uygulamalarında test etkililiği üzerindeki etkisi incelenmiştir.

Yöntem

Çalışma, simülatif verilerle yürütülmüştür. Bunun nedeni, BOBUT uygulamalarının geniş bir örneklem ve geniş bir madde havuzu gerektirmesinden dolayı gerçek veriye ulaşmanın güçlüğünden kaynaklanmaktadır (McDonald, 2002; Risk, 2015; Wang vd., 2012). Simülatif veri üretilmesi sürecinde Rstudio catIrt paketinden yararlanılmış ve BOBUT uygulamaları için SimulCat programı kullanılmıştır. Yetenek parametreleri ortalaması 0, standart sapması 1 olan normal dağılımdan ve bir uygulamada 1000 kişinin uygulamaya katılması şeklinde üretilmiştir. Madde parametreleri ise sadece güçlük parametresindeki değişimin incelenmesinden dolayı Rasch modele göre üretilmiştir. Bu bağlamda madde güçlük parametresi ortalaması 0, standart sapması 1 olan normal dağılımdan 300 madde olacak şekilde üretilmiştir. MPS açısından ise sadece güçlük parametresinde kolay yönde sapma olan durumla daha sık karşılaşılmasından dolayı (Babcock ve Albano, 2012; Hagge vd., 2011; Risk, 2015;

[1] TED Üniversitesi
[2] Ankara Üniversitesi

Stahl ve Muckle, 2007) madde güçlük parametresinde kolay yönde rastgele ve ardışık sapma olan durum ele alınmıştır. MPS'nin büyüklüğü ise hangi tür MPS'nin test etkililiği üzerinde etkisinin daha çok görüldüğünün belirlenebilmesi açısından 1.00 logit biriminde değişim olacak şekilde düzenlenmiştir. Çünkü özellikle 0.50 logit ve üzeri MPS'nin etkisinin daha önemli olduğu belirtilmektedir (Donoghue ve Isham, 1998; Wollack vd., 2005). MPS gösteren madde yüzdesi ise literatürdeki çalışmalara göre (Hagge vd., 2011; Stahl vd., 2002; Song ve ArceFerrer, 2009; Wells vd., 2002) %5 ve %25 olarak belirlenmiştir. BOBUT koşulları ise sabit koşullar olarak ele alınmıştır. Yetenek kestirim yöntemi olarak Beklenen Sonsal Dağılım (BSD) (Keller, 2000; Kingsbury ve Zara, 2009; Wang vd., 2012), başlama kuralı olarak önsel teta dağılımları (Segall, 2004), madde seçim yöntemi olarak Maksimum Fisher Bilgisi (MFB) (Van der Linden ve Glas, 2010; Wainer, 2000; Weiss ve Kingsbury,1984), madde kullanım sıklığı yöntemi olarak azalarak kaybolma yöntemi, test durdurma kuralı olarak da standart hatanın 0.40'a eşit ve küçük olduğu durum ele alınmıştır (Babcock ve Weiss, 2009; Blaise ve Raiche, 2002). R-MPS ve A-MPS'nin BOBUT uygulamasındaki test etkililiği üzerindeki etkisini incelemek için de ölçeklenmiş X2 istatistiği, test uzunluğu ve madde kullanım sıklığı oranları hesaplanmıştır.

Beklenen Bulgular

Çalışmaya ilişkin analizler henüz tamamlanmamıştır. MPS'li madde oranın %25 ve maddelerde A-MPS olduğu durumda, R-MPS olduğu duruma göre test etkililiğinde daha çok azalma beklenmektedir. Bir başka ifade ile madde havuzunda A-MPS gösteren madde yüzdesi arttıkça, BOBUT uygulamasına ilişkin ortalama test uzunluğu, madde kullanım sıklığı oranı ve ölçeklenmiş X2 değerlerinde artış beklenmektedir. Çünkü A-MPS olduğu durumda maddelerin güçlük parametresinde sabit bir artma veya azalma görülmekte, bu durum da MPS'nin etkisini arttırmaktadır. Test eşitleme bağlamında, A-MPS'nin, R-MPS'ye göre yanlı yetenek kestirimlerine neden olduğu bazı çalışmalarda da görülmüştür (Wollack vd., 2006). Bu bağlamda, elde edilen bulgular alan yazın desteğiyle tartışılacaktır.

Kaynakça

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. Applied Psychological Measurement, 36(7), 565-580.

Babcock, B., & Weiss, D.J. (2009). Termination criteria in computerized adaptive tests: Do variable-length CAT's provide efficient and effective measurement? International Association for Computerized Adaptive Testing, 1, 1-18.

Blais, J. & Raiche, G. (2002, April). Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules, International Objective Measurement Workshop, New Orleans.

Bock, D. B., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 25(4), 275-285.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. Journal of Applied Psychology, 84(4), 610-619.

Deng, H., & Melican, G. (2010, April). An investigation of scale drift in computer adaptive test. Paper presented at Annual Meeting of National Council on Measurement in Education. San Diego, CA.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. Applied Psychological Measurement, 22(1), 33-51.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 20(4), 369-377.

Hagge, S., Woo, A., & Dickison, P. (2011, October). Impact of item drift on candidate ability estimation. Paper presented at the Annual Conference of the International Association for Computerized Adaptive Testing, Pacific Grove, CA.

Keller, A.L. (2000). Ability estimation procedures in computerized adaptive testing. Technical Report, American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams, May.

Kingsbury, G. G., & Zara, A. R. (2009). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2(4), 359- 375.

McCoy, K. M. (2009). The impact of item parameter drift on examinee ability measures in a computer adaptive environment (Unpublished doctoral dissertation). University of Illinois at Chicago, Chicago, IL.

McDonald, P.L. (2002). Computer adaptive test for measuring personality factors using item response theory (Unpublished doctoral dissertation). The University Western of Ontario, London.

Risk, N.M. (2015). The impact of ıtem parameter drift in computer adaptive testing (CAT) (Unpublished doctoral dissertation), University of Illinois, Chicago.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. The Alberta Journal of Educational Research, 49(3), 264-276.

Segall, D. O. (2004). Computerized adaptive testing. K. Kempf-Lenard (Ed.), The Encyclopedia of social measurement. San Diego, CA: Academic Press.

Song, T., & Arce-Ferrer, A. (2009, April). Comparing IPD detection approaches in common-item nonequivalent group equating design. Paper presented at the Annual Meeting of The National Council on Measurement, San Diego, CA.

Stahl, J. A., & Muckle, T. (2007, April). Investigating displacement in the Winsteps Rasch calibration application. Paper presented at the Annual Meeting of The American Educational Research Association, Chicago, IL.

Wang, H-P., Kuo, B-C., Tsai, Y-H., & Liao, C-H. (2012). A Cerf-Based computerized testing system for chinese proficiency. TOJET: The Turkish Journal of Educational Technology, 11(4), 1–12.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21(4), 361–375.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. Applied Psychological Measurement, 26(1), 77- 87.

Wollack, J. A., Sung, H. J., & Kang, T. (2005). Longitudinal effects of item parameter drift. Paper presented at Annual Meeting of the National Council on Measurement in Education, Montreal, CA.

Anahtar Kelimeler: madde parametre sapması, bilgisayar ortamında bireye uyarlanmış testler, simülasyon yöntemi

Bilgisayar Ortamında Bireye Uyarlanmış Yabancı Dil Testlerinde Üretim Gerektiren Becerilerin Değerlendirilmesine Yönelik Çalışmalar

Ahmet Kütük[1] Hakan Koğar[1]

ÖZET

Giriş

Yerleştirmeye dayalı ya da başarıyı ölçmeye yönelik sınavlarda bilgisayar ortamında bireye uyarlanmış testlerin (BOBUT) kullanımı dünya çapında yaygınlaşmaktadır. Test uygulayıcıların yanı sıra, BOBUT'un bireyin yetenek düzeyi temelinde madde seçimi ve test süresini kısaltması gibi avantajları sayesinde testi alan kişiler tarafından da tercih edildiği bilinmektedir. Amerika'da lisansüstü düzeyde üniversite giriş sınavı olarak kullanılan Graduate Record Examination (GRE) ve Graduate Management Admission Test (GMAT) sınavları ile dünya genelinde kabul gören Test of English as a Foreign Language (TOEFL) gibi yüksek önem taşıyan sınavlar günümüzde BOBUT olarak uygulanabilmektedir. Testlerin bilgisayar ortamında kolaylıkla yapılabiliyor olması kullanışlılık açısından önemli getiriler sunmasına rağmen değerlendirme aşamasında birtakım güçlükler de ortaya koymaktadır. Özellikle bir yabancı dildeki yeteneğin kestirilmesi aşamasında üretkenlik gerektiren becerilerin bilgisayar ortamında puanlanabilmesi hala birtakım soruları ve sorunları barındırmaktadır. Bu derleme çalışmasının amacı, yabancı dil yeteneğinin ölçülmesi aşamasında üretim gerektiren beceriler olan konuşma ve yazma yetilerinin yazılım temelli değerlendirilmesine yönelik halihazırdaki BOBUT uygulamaları hakkında durum tespiti yapmak, benzer değerlendirme süreçlerinde çalışacak araştırmacıları bir araya getirmek ve daha güvenilir sonuçlar elde edebilmek için olası önerileri tartışmaktır. Elde edilecek bilgilerin, bu çalışmanın araştırmacıları tarafından yürütülmekte olan ve yabancı dil yeterliğine yönelik bir BOBUT uygulaması konulu doktora tez çalışmasına da temel oluşturması beklenmektedir. Araştırma devam etmekte olup elde edilecek veriler ve bulgular sempozyum sürecinde paylaşılacaktır.

Yöntem

Bu araştırmada, dünya çapında yabancı dilde konuşma ve yazma becerilerinin değerlendirilmesine yönelik bilgisayar ortamında yapılmakta olan ölçme ve değerlendirme uygulamaları, derleme yöntemiyle analiz edilecektir. Araştırmada veri toplama yöntemi olarak nitel araştırma yöntemlerinden doküman analizi kullanılacaktır. Doküman analizi, bir amaç doğrultusunda derlenen belgelerin içeriklerinin titizlikle ve sistem dahilinde analiz etmek için faydalanılan bir araştırma yöntemidir (Wach, 2013, Karasar, 2005). Doküman analizi, araştırmalarda kullanılan diğer yöntemler için bir tamamlayıcı unsur olmakla birlikte bağımsız bir yöntem olarak da değerlendirilmektedir. (Bowen, 2009). Araştırma kapsamında amaçlı örnekleme yöntemiyle seçilecek makale, tez çalışmaları ile resmi kurum sayfalarından elde edilecek belgeler incelenecektir. Alanyazında geçen çalışmaların derlenmesi aşamasında YÖK Tez Merkezi, DergiPark, Google Akademik ile uluslararası veri tabanlarından faydalanılacaktır.

Araştırma süresince dünya genelinde yabancı dilde konuşma ve yazma becerilerinin değerlendirilmesine için kullanılan BOBUT uygulamalarının özet olarak ortaya koyulması, bu uygulamaların konuşma ve yazma ürünlerini nasıl puanladığının incelenmesi, varsa eksik yönlerinin ortaya koyulması ve olası önerilerin tartışılması beklenmektedir. Araştırma kapsamında incelenen dokümanlar ve bunlara ait içerik bilgileri sempozyum sürecinde paylaşılacaktır. Bu sebeple, derlenen

[1] Akdeniz Üniversitesi

belgeler ve içerikleri konusundaki bilgiler henüz çalışılmakta olup elde edilecek veriler sempozyumda diğer araştırmacılar ile detaylı olarak paylaşılacaktır.

Beklenen Bulgular

Yabancı dil yeterliğinde konuşma ve yazma becerilerinin değerlendirilmesi için kullanılan BOBUT uygulamalarında son yıllarda oldukça hızlı ilerlemeler kaydedilmektedir. Yapay zekâ uygulamalarının da ölçme ve değerlendirme çalışmalarına entegre edildiği, BOBUT tarzında yapılabilen birtakım sınavların değerlendirilmesi aşamasında yapay zekâ [artificial intelligence (AI)] uygulamalarından da faydalanıldığı görülmektedir. Özellikle TOEFL-IBT gibi dünya genelinde milyonlarca insanın girdiği bir sınavın yazma ve konuşma becerilerinin değerlendirilmesi aşamasında yetkilendirilmiş puanlayıcı bireylerin yanı sıra AI uygulaması ile elde edilen puanların bir kombinasyonu kullanılmaktadır. Ancak, beklenen ürünlerin değerlendirmesi aşamasında bireysel puanlayıcılara ihtiyaç duyulması, BOBUT uygulamalarında hızlı ve güvenilir değerlendirme için bir engel durumunda olup yazma ve konuşma ürünlerinin BOBUT kapsamı dışında kalmasına neden olmaktadır. Bunların aşılabilmesi için daha kapsamlı çözüm yollarına ihtiyaç duyulmaktadır. Nitekim, bireye uyarlanmış testlerdeki adayın düzeyi dikkate alınarak çok zor ve çok kolay maddelerin sorulmaması yönündeki genel prensip (Davey, 2011) bu bölümlerde uygulanamamakta, aynı test maddelerinin testi alan tüm bireylere yöneltilmesine neden olmaktadır. Yazma ve konuşma becerilerinin değerlendirilmesine yönelik elde edilecek bulgular neticesinde, sempozyumda daha verimli tartışmaların, önerilerin ve hatta yeni çalışma gruplarının oluşturulabileceği düşünülmektedir. Çalışma devam etmekte olup, elde edilecek bulgulara sempozyumda detaylı olarak yer verilecektir.

Kaynakça

Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27-40.

Davey, T. (2011). A guide to computer adaptive testing systems educational testing service for technical issues in large-scale assessment (TILSA). State Collaborative on Assessment and Student Standards (SCASS). Alındı Nisan 21, 2023, https://files.eric.ed.gov/fulltext/ED543317.pdf

Karasar, N. (2005). Bilimsel araştırma yöntemi. Nobel Yayın Dağıtım.

Wach, E. (2013). Learning about qualitative document analysis.

Anahtar Kelimeler: BOBUT, yabancı dil, ölçme, değerlendirme, yazma, konuşma, yapay zeka