ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);422-435

TIMSS Matematik Değerlendirmeleri Bilgisayar Ortamında Bireyselleştirilmiş Test Olarak Uygulanabilir mi?

Semirhan GÖKÇE* Cees A.W. GLAS**

Öz

Son yıllarda, bilgisayar ortamında bireye uyarlanmış testlerin (BOBUT) özellikle geniş ölçekli test uygulamalarında kullanımı yaygın hale gelmiştir. Testte kullanılan maddelerin ve katılımcıların özelliklerine bağlı olarak en uygun bireye uyarlanmış test algoritmasının belirlenebilmesi amacıyla gerçek ya da türetilmiş veri setlerinin kullanıldığı çok sayıda simülasyon çalışması gerçekleştirilmiş ve çalışmalardan elde edilen bulgular sonucunda birçok test gerçek anlamda bilgisayar ortamında bireye uyarlanmış halleriyle uygulanmaya konulmuştur. Geniş ölçekli test uygulamaları dendiğinde ilk akla gelen uygulamalardan biri olan Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) 1995 yılından itibaren dördüncü ve sekizinci sınıf düzeylerinde matematik ve fen bilimlerindeki öğrenci başarısını izlemek amacıyla kağıt ve kalem testleri kullanılarak gerçekleştirilmektedir. Bu çalışmanın amacı, TIMSS sekizinci sınıf matematik değerlendirmeleri için en uygun BOBUT algoritmasının belirlenmesini sağlamaktır. TIMSS 2007, 2011 ve 2015 uygulamalarına sekizinci sınıf düzeyinde katılan Türkiye ve ABD'ye ait veri setlerinde yer alan 393 madde, marjinal maksimum olabilirlik tahmin yöntemi kullanılarak aynı ölçek üzerinde kalibre edilmiştir. Madde havuzu oluşturulduktan gerçekleştirilen simülasyonlar ile en iyi test başlangıç kuralının, yetenek kestirim yönteminin, test sonlandırma kuralının ve madde kullanım sıklığı kontrolünün belirlenmesi amaçlanmıştır. Araştırma bulguları beklenen sonsal dağılım yetenek kestirim yönteminin kullanıldığı, test başlangıç kuralı olarak rastgele 6 maddenin uygulandığı, test sonlandırma kuralı olarak 20 maddeden oluşan sabit uzunluktaki testlerin kullanıldığı algoritmanın TIMSS sekizinci sınıf matematik değerlendirmeleri için en uygun yapı olduğunu ortaya koymuştur. Bununla birlikte, madde kullanım sıklığı kontrolünün madde havuzunun etkili kullanımında çok önemli bir yere sahip olduğu belirlenmiştir. Bu çalışmanın hem ulusal hem de uluslararası düzeyde uygulanan geniş ölçekli kağıt ve kalem testlerine alternatif olarak geliştirilebilecek BOBUT uygulamalarındaki en uygun algoritmanın belirlenmesi ve geliştirilmesi süreçlerine katkı sağlayacağı düşünülmektedir.

Anahtar Kelimeler: bilgisayar ortamında bireye uyarlanmış test, madde tepki kuramı, matematik değerlendirmesi, simülasyon çalışması, TIMSS.

GİRİŞ

Son yıllardaki teknolojik gelişmeler, geleneksel kağıt ve kalem testlerine alternatif olabilecek eğitim amaçlı testlerin gelişmesine ve kullanılmasına önayak olmuştur. Başlangıçta sadece soruların gösterimi ve yanıtların toplanması amaçlı kullanılan bilgisayarlar, daha sonraları çoklu ortam sayesinde ses ve görüntü içeren farklı soru formatlarını destekleyerek (Zenisky & Sireci, 2002) ve uygulama bitiminde test puanını hemen kullanıcı ile paylaşarak test ortamlarında daha fazla tercih edilir olmuşlardır. Süreç içerisinde, katılımcılara aynı soruları sormak yerine farklı test formlarının uygulanması düşüncesi ortaya çıkmış daha sonraları ise her katılımcının önceki yanıtlarına bakarak bir sonraki sorunun belirlenmesine yol açmıştır (Davey & Pitoniak, 2006). Aslında bu süreç bilgisayar ortamında bireye uyarlanmış testlerin (BOBUT) gelişimindeki temel fikrin oluşmasını sağlamıştır. Bu tür testlerin temel amacı katılımcıların yetenek düzeyi hakkında mümkün olduğunca fazla bilgi sahibi olabilmektir. İstatistiksel açıdan bakıldığında ise, verilen doğru ya da yanlış her yanıt ölçülen nitelik açısından soruyu yanıtlayan hakkında az ya da çok bir bilgi vermektedir. Fakat sorunun güçlüğü katılımcının yetenek düzeyine uygun olursa bu durumda katılımcıdan elde edilen bilgi düzeyi

* Dr. Öğretim Üyesi, Niğde Ömer Halisdemir Üniversitesi, Niğde-Türkiye, e-posta: semirhan@gmail.com, ORCID ID: 0000-0002-4752-5598
** Prof. Dr., Twente Üniversitesi, Enschede-Hollanda, e-posta: c.a.w.glas@utwente.nl, ORCID ID: 0000-0001-6531-5503

To cite this article:
Gökçe, S., & 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
Geliş Tarihi: 25.11.2018
Kabul Tarihi: 21.12.2018

yükselmektedir. Bu nedenle, BOBUT uygulamalarında eğer bir katılımcı bir soruyu doğru yanıtlarsa katılımcının yetenek kestirim değeri artacağından bir sonraki soru öncekiyle kıyaslandığında daha zor olmaktadır. Ya da aynı katılımcı soruyu yanlış yanıtlarsa sonraki soruda daha kolay bir soru ile karşı karşıya gelmektedir (Hambleton, Swaminathan, & Rogers, 1991; Luecht & Sireci, 2012; van der Linden, 2010). Bu optimizasyon süreci, BOBUT algoritmasının katılımcının yetenek düzeyi hakkında yeterli düzeyde bilgi sahibi oluncaya dek devam etmektedir. Böylece, tüm katılımcıların aynı soruları yanıtladığı geleneksel testlerden farklı olarak, soruların katılımcıların yanıtlarına göre belirlendiği BOBUT uygulamalarında çok sayıda test formu süreç içerisinde oluşmaktadır (Sireci, Baldwin, Martone, Kaira, Lam, & Hambleton, 2008). Bu nedenle, bilgisayar ortamında uygulanan birçok farklı test türü lineer testlerden bireye uyarlanmış testlere kadar çok geniş bir spektrumda yer almaktadır.

Her katılımcının istediği anda testte yer alan soruları çözebilmesi ya da diğer bir ifadeyle tüm katılımcıların aynı anda testteki soruları çözmek zorunda olmaması (Glas & Geerlings, 2009; Hambleton vd., 1991; van der Linden, 2001; Wainer, 2000), daha az soru ile daha güvenilir ölçüm yapılabilmesi (Eggen, 2007; Hambleton vd., 1991; Meijer & Nering, 1999; Mills & Stocking, 1996; Verschoor & Straetmans, 2010), uygulama bitiminde test puanının öğrenilebilmesi (Eggen, 2007; Wainer, 2000) ve kullanıcıların test stresini azaltması (Hambleton et al., 1991, Mills & Stocking, 1996) gibi nedenler BOBUT'un lineer testler karşısındaki üstünlüğü olarak ifade edimektedir. Diğer yandan, testlerdeki soru kontrolünün tamamen bilgisayar bağlı olması (testi geliştirenlerden bağımsız olarak testteki maddelerin belirlenmesi) ve kalibre edilmiş geniş bir soru havuzuna ihtiyaç duyulması gibi sebepler de BOBUT'un lineer testler karşısında dezavantajları olarak ifade edilmektedir (Meijer & Nering, 1999).

Teorik olarak BOBUT, bir soruya verilen yanıtın katılımcının yetenek düzeyi ve soru parametreleri cinsinden bir matematiksel fonksiyon olarak ifade edilebilmesini mümkün kılan madde tepki kuramı (MTK) çerçevesi üzerine kuruludur. MTK ile katılımcıların yetenek düzeyi yöneltilen maddelerden bağımsız olarak belirli bir hata payıyla belirlenebilir. Ayrıca, MTK'nın farklı maddeleri yanıtlayan katılımcıların da yetenek kestirimlerinin karşılaştırılabilmesine imkan sağladığı belirtilmektedir (Hamleton vd., 1991). Bu nedenle, madde parametreleri ile katılımcıların yetenek kestirimlerini eşleştirebilmek için istatistik özellikleri belirlenmiş geniş bir madde havuzuna (bazı kaynaklarda madde bankası olarak da ifade edilmektedir) ihtiyaç duymaktadır. van der Linden (1995), BOBUT algoritmasının (1) test başlangıç kuralının (2) madde seçim yönteminin (3) yetenek kestirim yönteminin ve (4) test sonlandırma kuralının belirlenmesi adımlarından oluştuğunu ifade etmektedir. En uygun test başlangıç kuralı belirlenirken bazı kaynaklarda yöneltilecek ilk sorunun güçlüğü önem kazanmakta ve uygulamanın başında kolay sorular yöneltmenin katılımcıların motivasyonunu arttırdığı ve bu durumun katılımcıların test başarısına olumlu yansıdığı dile getirilmektedir (Mills & Stocking, 1996). Soru seçim yöntemleri düşünüldüğünde ise iki temel yaklaşım bulunmaktadır ki bunlar Fisher maksimum bilgi ve Bayes yöntemleridir. Her ne kadar Wainer (2000) her iki yöntemin de iyi sonuçlar verdiğini belirtse de Eggen (2004) Bayes yönteminin uygulanabilmesi için daha gelişmiş bilgisayarlara ihtiyaç duyulduğunu belirtmiştir. Yetenek kestirim yöntemleri incelendiğinde ise maksimum olabilirlik (MO), ağırlıklandırılmış maksimum olabilirlik (AMO), maksimum sonsal dağılım (MSD) ve beklenen sonsal dağılım (BSD) yöntemleri yer almaktadır. Gu ve Reckase (2007)'e göre, maksimum olabilirlik yöntemleriyle kıyaslandığında MSD ve BSD yöntemleri eş uzunluktaki testlerde daha küçük standart hata ile kestirim yapmakta fakat seçilen önsel dağılıma bağlı olarak yanlı kestirimde bulunabilmektedir. Test sonlandırma kuralları incelendiğinde ise sabit test uzunluğu ve değişken test uzunluğu olmak üzere iki farklı seçenek yer almaktadır. Sabit uzunluktaki testler tüm katılımcılara aynı sayıda soru yöneltilmesini sağlamakta ancak bu durumda yetenek düzeyleri farklı standart hatalar ile kestirilebilmektedir. Değişken uzunluktaki testlerde ise katılımcılara ait yetenek kestiriminin standart hatası daha önceden belirlenen bir değerden düşük olduğunda ya da ardışık iki yetenek kestirimi arasındaki fark yine daha önceden belirlenen bir değerden düşük olduğunda test sonlandırma kuralları devreye girebilmektedir. Bu aşamada testi geliştiren uzmanlar testin amacını, simülasyon sonuçlarını ve kapsam geçerliğini de dikkate alarak test sonlandırma kuralı olarak sabit uzunlukta testlerin mi yoksa değişken uzunluktaki testlerin mi kullanılacağına karar vermek durumundadır.

Bilgi ve iletişim teknolojilerinin gelişimine ve bilgisayarların yaygınlaşmasına bağlı olarak birçok geniş ölçekli test de bilgisayar ortamında test olarak ve hatta bireye uyarlanmış test formunda uygulanmaya başlanmıştır. Bu testlere örnek olarak Graduate Record Examinations (GRE), Graduate Management Admission Test (GMAT), Armed Services Vocational Aptitude Battery (ASVAB), ve United States Medical Licensing Examination (USMLE) testleri verilebilir. Educational Testing Service tarafından geliştirilen GRE, 1992 yılından itibaren ve Graduate Management Admission Council tarafından hazırlanan GMAT ise 1997 yılından itibaren bilgisayar ortamında bireye uyarlanmış test olarak uygulanmaktadır (Luecht ve Sireci, 2012).

Geniş ölçekli testlerde bu gelişmeler olurken, 1995 yılından günümüze kadar her dört yılda bir geniş ölçekli uygulanan Matematik ve Fen Eğilimleri Araştırması'nda (TIMSS) ise kağıt-kalem testleri kullanılmaktadır. TIMSS uygulamalarının temel amacı katılımcı ülkelerin matematik ve fen başarısını belirleyerek gelişimini takip etmektir. Çalışmaya katılan ülkelerin dördüncü ve sekizinci sınıf öğrencilerine başarı testleri uygulanmakta ve bu testlerde birbirine ortak sorular ile bağlı 14 farklı kitapçık kullanılmaktadır (Mullis, Martin, & Loveless, 2016). Kitapçıklarda çoktan seçmeli ve açık uçlu maddeler yer almaktadır. Bununla birlikte, farklı uygulama dönemlerinde kullanılan başarı testlerinin arasında da ortak maddelerin bulunması dönemler arasındaki eşitlemeyi mümkün kılmaktadır.

Çalışmanın Amacı

Bu çalışmanın amacı, TIMSS sekizinci sınıf matematik değerlendirmesinde kullanılan kağıt-kalem testlerine alternatif olarak geliştirilebilecek bilgisayar ortamında bireye uyarlanmış test uygulamaları için en uygun algoritmayı belirlemektir. Simülasyonlarda kullanılan madde havuzunun oluşturulmasında 2007, 2011 ve 2015 yıllarında gerçekleştirilen TIMSS uygulamalarına sekizinci sınıf düzeyinde katılan Türkiye ve ABD'ye ait veri setleri kullanılmıştır. Madde parametreleri belirlendikten sonra farklı senaryolardan elde edilen sonuçlar karşılaştırılmıştır. Bu senaryolarda farklı test başlama kuralları, yetenek kestirim yöntemleri, test sonlandırma kuralları değerlendirilmiştir. Ayrıca, her durum için madde kullanım sıklığı değerleri hesaplanmış ve madde kullanım sıklığı kontrol yönteminin belirlenen durumlar üzerindeki etkileri incelenmiştir. Bu araştırmada yanıt aranan sorular aşağıda verilmektedir.

1. TIMSS sekizinci sınıf matematik değerlendirmeleri için en uygun bireye uyarlanmış test algoritmasının test başlama kuralı, yetenek kestirim yöntemi ve test sonlandırma kuralı nasıl olmalıdır?

2. TIMSS sekizinci sınıf matematik değerlendirmesine alternatif olarak geliştirilen bireye uyarlanmış test algoritmasında kullanılan madde kullanım sıklığı kontrolü ne derece etkili olmaktadır?

YÖNTEM

Bu bölümde, katılımcılara, kullanılan veri toplama araçlarına (matematik başarı testlerine) ve veri analizine (madde parametrelerinin belirlenmesine ve gerçekleştirilen simülasyonlara) yönelik bilgiler yer almaktadır.

Katılımcılar

TIMSS'in 2007, 2011 ve 2015 yıllarındaki sekizinci sınıf matematik değerlendirmelerine Türkiye ve ABD'den katılan öğrenci sayıları Tablo 1'de verilmektedir.

Tablo 1. TIMSS sekizinci sınıf matematik değerlendirmelerine Türkiye ve ABD'den katılan öğrenci sayısı

Yıl	Türkiye	ABD	Toplam
2007	4498	7377	11875
2011	6928	10477	17405
2015	6079	10221	16300
Toplam	17505	28075	45580

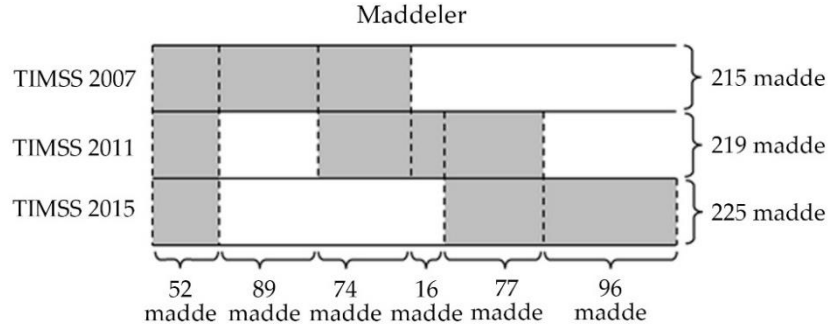
Veri Toplama Araçları

Daha önce de belirtildiği üzere, TIMSS uygulamalarında kullanılan başarı testlerinin her birinde birbirlerine ortak sorular ile bağlı olan 14 farklı test formu yer almaktadır. Bu test formlarında yer alan soru sayıları Tablo 2'de verilmektedir.

Tablo 2. TIMSS sekizinci sınıf matematik değerlerlendirmelerindeki test uzunlukları

Kitapçık	TIMSS 2007	TIMSS 2011	TIMSS 2015
1	29	26	35
2	31	32	33
3	32	32	28
4	29	29	32
5	29	32	34
6	32	33	32
7	33	30	32
8	32	34	30
9	31	34	28
10	32	31	28
11	32	32	29
12	28	32	30
13	28	33	29
14	30	27	30
Ortalama	30.6	31.2	30.7

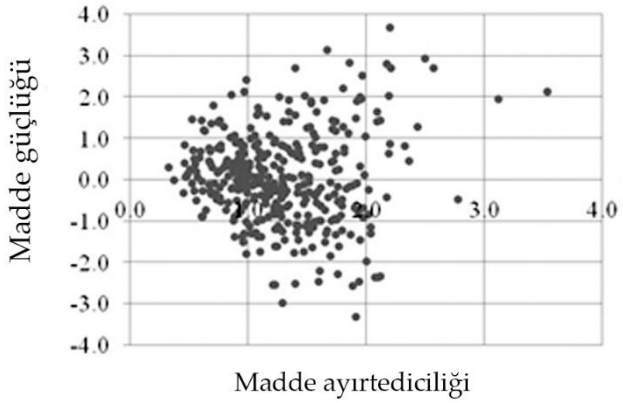
Tablo 2'ye göre, TIMSS uygulamalarında kullanılan matematik başarı testlerinde katılımcılara ortalama 30 soru yöneltilmektedir. Bu testlerdeki yanıt örüntüleri kullanılan ortak soruların yardımıyla birleştirilmiş ve Şekil 1'de verilen eksik test tasarımı elde edilmiştir.



Şekil 1. TIMSS sekizinci sınıf matematik değerlendirmesinin eksik test tasarımı

Elde edilen verilerde 45580 öğrenci ve 404 madde yer almaktadır. Bu maddelerden 11 tanesi (M042273, M062345BA, M062345BB, M062345BC, M062345BD, M062345B, M062342,

M062048A, M062048B, M062048C ve M062048) herhangi bir yanıt örüntüsü içermediğinden madde parametreleri hesaplanamamış ve araştırmanın kapsamı dışında bırakılmıştır. Geri kalan 393 maddenin 360 tanesi çoktan seçmeli olduğu ve bu maddelere verilen yanıtlar 0 ya da 1 olarak puanlandığı için 2 parametreli lojistik (2PL) model kullanılarak, 33 tanesi ise kısmi puanlama yapıldığı ve bu maddelere verilen yanıtlar 0, 1 ya da 2 olarak puanlandığı için kısmi puan modeli (partial credit model-PCM) kullanılarak kalibre edilmiştir. Havuzdaki çoktan seçmeli maddeler 0 ya da 1 olarak puanlanırken, açık uçlu maddelerin puanlanmasında iki farklı durum söz konusudur. Bazı açık uçlu maddeler 0 ya da 1 şeklinde puanlanırken, bazıları ise 0, 1 ya da 2 olarak puanlanmaktadır. Çalışmada kullanılan kısmi puan modelinin iki ya da daha fazla kategoride puanlanan yanıtlar için tek boyutlu bir model olduğu belirtilmektedir (Masters, 2016). Araştırma kapsamındaki madde parametrelerinin belirlenmesi sürecinde ise Multidimensional Item Response Theory-MIRT (Glas, 2010) programı kullanılmıştır. Çalışmada 2PL model kullanılarak kalibre edilen maddelere ait parametrelerin (madde güçlüğü ve ayırtedicilik) dağılımı Şekil 2'de verilmektedir.



Şekil 2. 2PL modele göre kestirilen madde parametrelerinin dağılımı

MIRT programı madde parametrelerinin yanı sıra her katılımcının kestirilen yetenek düzeyini ve kestirimlerin standart hatalarını da rapor etmektedir. Buna göre, örneklemde yer alan katılımcıların ağırlıklı maksimum olabilirlik (AMO) ve beklenen sonsal dağılım (BSD) yöntemleri kullanılarak elde edilen yetenek kestirimlerine ait istatistiksel bilgiler Tablo 3'te verilmektedir.

Tablo 3. Madde kalibrasyonunda elde edilen ortalama yetenek kestirimi ve ortalama standart hata değerleri

İstatistik	Yetenek kestirim yöntemi	
	AMO	BSD
Yetenek kestirimi	-.054	-.063
Ortalama standart hata	.371	.328

Tablo 3'te görüldüğü üzere AMO ve BSD ile kestirilen yetenek düzeylerinin ortalamasının sırasıyla -.054 ve -.063 değerlerini aldıkları ve sırasıyla .371 ve .328 ortalama standart hata ile kestirim yaptıkları belirlenmiştir.

Veri Analizi

TIMSS uygulamalarının test eşitleme ve ölçeklendirme çalışmaları madde tepki kuramı (MTK) temel alınarak yapıldığı için (Martin, Mullis & Hooper, 2016) bu çalışmada kullanılan veri setinin MTK

varsayımlarını sağladığı kabul edilmiş ve madde kalibrasyonu tek boyutlu MTK modeline göre gerçekleştirilmiştir. Veri analiz sürecinde öncelikle MIRT (Glas, 2010) programı kullanılarak madde parametreleri elde edilmiştir. Bu aşamada 0 ya da 1 olarak puanlanan 360 madde 2-parametreli lojistik model ile 0, 1 ya da 2 olarak puanlanan 33 madde ise kısmi puan modeli ile kalibre edilmiştir. Madde parametreleri belirlendikten sonra simülasyon çalışmalarına geçilmiştir. Öncelikle, analizlerde kullanmak üzere normal dağılımdan N(0,1) ortalaması 0 ve standart sapması 1 olan 1000 kişiye ait yetenek kestirimi çekilmiştir. Daha sonra, madde parametreleri ve yetenek kestirimleri dikkate alınarak her bireye ait yanıt örüntüsü oluşturulmuş; 1000 satır ve 393 sütundan oluşan bir yanıt matrisi elde edilmiştir.

İlk simülasyon çalışmasında, değişken uzunluklu testler kullanılmış ve standart hata için .20, .30 ve .40 referans değerleri belirlenerek (a) gerçek yetenek ile kestirilen yetenek arasındaki korelasyon (b) ortalama test uzunluğu (c) madde kullanım sıklığı (d) hata karelerinin ortalamasının karekökü *(İng. Root Mean Square Error-RMSE)* ve (e) yanlılık *(İng. bias)* açısından incelenmiştir. Burada yer alan, madde kullanım sıklığı (MKS) değeri, ilgili maddeyi yanıtlayan kişi sayısının uygulamaya katılan toplam kişi sayısına bölümü olarak ifade edilmektedir. Örneğin, 1000 kişinin katıldığı bir uygulamada eğer bir madde 130 kişiye yöneltildiyse bu maddenin MKS değeri .13 olarak hesaplanmaktadır. RMSE ve yanlılık değerleri ise gerçek yetenek düzeyleri ile uygulama sonunda kestirilen yetenek düzeylerinin birbirinden ne ölçüde farklılaştığını tanımlamaktadır.

İkinci simülasyon çalışmasında ise sabit uzunluklu testlere odaklanılmıştır. Buna göre, test uzunluğu 10, 20 ve 30 madde ile sabitlenerek (a) gerçek yetenek ile kestirilen yetenek arasındaki korelasyon (b) ortalama standart hata (c) madde kullanım sıklığı (d) hata karelerinin ortalamasının karekökü (RMSE) ve (e) yanlılık (bias) değerleri karşılaştırılmıştır.

Üçüncü simülasyon çalışmasında madde kullanım sıklığı kontrolünün etkililiği incelenirken dördüncü simülasyon çalışmasında yetenek kestirim yöntemlerinin etkililiği araştırılmıştır.

Yürütülen simülasyonlarda başlama kuralı olarak testin başında farklı sayılarda rastgele maddeler kullanılmış, madde seçim yöntemi olarak Fisher bilgi fonksiyonu kullanılmış, yetenek kestirimi olarak ağırlıklı maksimum olabilirlik (AMO) ve beklenen sonsal dağılım (BSD) yöntemleri karşılaştırılmış, test sonlandırma kuralı olarak sabit ve değişken uzunluktaki testler tanımlanmış ve madde kullanım sıklığı kontrolü olarak Randomesque (Kingsbury & Zara, 1989) yönteminin etkisi araştırılmıştır.

BULGULAR

TIMSS sekizinci sınıf uygulamaları için en uygun BOBUT algoritmasının belirlenebilmesi amacıyla yürütülen simülasyonlarda 36 farklı durum karşılaştırılmıştır. Buna göre, üç farklı test başlangıç kuralı (yetenek kestiriminin hemen rastgele seçilen ilk maddeden sonra, rastgele 3 maddeden sonra ya da rastgele 6 maddeden sonra yapılması), iki farklı yetenek kestirim yöntemi (BSD ya da AMO) ve 6 farklı test sonlandırma kriteri (10, 20 ya da 30 maddeden oluşan sabit uzunluktaki testler ile en fazla .20, .30 ya da .40 standart hata ile kestirilen değişken uzunluklu testler) karşılaştırılmıştır.

a) Değişken uzunluktaki testler kullanılarak gerçekleştirilen simülasyonlar

Belirlenen durumların değişken uzunluktaki testler üzerindeki etkilerini karşılaştırabilmek amacıyla gerçekleştirilen simülasyonlarda, ortalama test uzunlukları ile gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon katsayıları hesaplanmıştır. Elde edilen sonuçlar Tablo 4'te verilmektedir.

Tablo 4. Değişken uzunluktaki testlerin ortalama soru sayıları ile gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon katsayıları

Yöntem	İlk yetenek kestirimi	Test sonlandırma kuralı					
		SH < .20		SH < .30		SH < .40	
		test uzunluğu	r	test uzunluğu	r	test uzunluğu	r
BSD	ilk maddeden sonra	33	.978	12	.961	6	.931
	rastgele yöneltilen 3 maddeden sonra	35	.980	13	.955	7	.926
	rastgele yöneltilen 6 maddeden sonra	36	.978	15	.960	9	.922
AMO	ilk maddeden sonra	36	.979	13	.953	7	.929
	rastgele yöneltilen 3 maddeden sonra	36	.980	14	.957	9	.934
	rastgele yöneltilen 6 maddeden sonra	38	.980	16	.958	10	.933

Tablo 4'te görüldüğü üzere, standart hata referans değerinin daha düşük belirlendiği durumda daha yüksek korelasyon değerleri elde edilmiştir. Zira bu durum test puanlarının standart hatası ile güvenirlik düzeyi arasındaki ilişki ile açıklanabilir. Ayrıca, ortalama test uzunlukları da bu durum ile doğrudan ilişkilidir. Bir başka deyişle, yetenek kestiriminin .20'den daha düşük bir standart hata ile yapılabilmesi amacıyla daha fazla madde katılımcılara yöneltilmektedir. Yetenek kestiriminde standart hata referans değerinin .40'tan .30'a düşürülmesi sonucu ise yanıtlanan ortalama madde sayısı yaklaşık iki katına çıkmakta, standart hata referans değerinin .30'dan .20'ye düşürülmesi halinde ise yanıtlanan ortalama madde sayısı yaklaşık 3 katına çıkmaktadır. Değişken uzunluktaki testlerde, test başlangıcında daha fazla sayıda rastgele madde kullanımının test uzunluğunu (ya da yanıtlanan madde sayısını) artırdığı gözlenmiştir. Daha ayrıntılı belirtmek gerekirse, test başlangıcında kullanılan rastgele madde sayısının arttırılması yetenek düzeyi hakkında en fazla bilgiyi veren maddeyi seçmek yerine rastgele madde seçimine odaklandığından değişken uzunluktaki testlerde test uzunluğunun artmasına yol açmaktadır. Son olarak, değişken uzunluktaki testlerde BSD ve AMO yetenek kestirim yöntemlerinin etkisi incelendiğinde hem gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon değerleri bakımından hem de testlerdeki ortalama madde sayısı bakımından belirgin farklar bulunmadığı görülmektedir.

Değişken uzunluktaki test uygulamalarında farklı yetenek kestirim yöntemleri kullanılarak elde edilen madde kullanım sıklığı dağılımları, RMSE ve yanlılık değerleri Tablo 5'te verilmektedir.

Tablo 5. Değişken uzunluktaki testlerin madde kullanım sıklığı dağılımları, RMSE ve yanlılık değerleri

Yöntem	İlk yetenek kestirimi	Test sonlandırma kuralı	Madde kullanım sıklığı				RMSE	Yanlılık
			<.01	.01-.20	.21-.40	>.40		
BSD	ilk maddeden sonra	SH < .20	0	334	38	21	.210	-.008
		SH < .30	168	205	14	6	.286	-.012
		SH < .40	235	148	6	4	.359	.001
	rastgele yöneltilen 3 maddeden sonra	SH < .20	0	338	35	20	.209	.000
		SH < .30	163	211	13	6	.296	-.006
		SH < .40	203	181	7	2	.375	.014
	rastgele yöneltilen 6 maddeden sonra	SH < .20	0	339	34	20	.210	-.004
		SH < .30	150	226	14	3	.280	.017
		SH < .40	214	173	4	2	.382	-.021
AMO	ilk maddeden sonra	SH < .20	0	333	35	25	.202	-.013
		SH < .30	116	257	15	5	.303	-.017
		SH < .40	198	184	8	3	.383	-.039
	rastgele yöneltilen 3 maddeden sonra	SH < .20	1	334	35	23	.198	-.019
		SH < .30	138	236	15	4	.284	-.016
		SH < .40	168	215	7	3	.376	-.026
	rastgele yöneltilen 6 maddeden sonra	SH < .20	1	333	35	24	.202	-.007
		SH < .30	131	245	14	3	.292	-.022
		SH < .40	189	197	5	2	.365	-.015

Tablo 5'te görüldüğü üzere, değişken uzunluktaki testlerde madde kullanım sıklığı, RMSE ve yanlılık üzerine etkileri incelenmiş ve belirlenen standart hata referans değerinin azaltılması durumunda hiç kullanılmayan ya da çok az kullanılan (kullanım sıklığı .01'den küçük) madde sayılarında azalmalar meydana geldiği belirlenmiştir. Bu durum olumlu görünmesine rağmen standart hata referans değerini küçültmenin çok fazla kullanılan (kullanım sıklığı .40'tan büyük) madde sayılarını arttırması gibi olumsuz sonuçlara yol açmıştır. Test sonlandırma kuralının RMSE ve yanlılık üzerindeki etkileri incelendiğinde ise, hemen hemen tüm durumlarda daha katı değişken uzunluktaki test sonlandırma kuralı belirlenmesi halinde (standart hata referans değerinin daha küçük tutulması durumunda) daha küçük RMSE ve yanlılık değerleri elde edildiği gözlenmiştir.

Tablo 5'te verilen BSD ve AMO yetenek kestirim yöntemlerinin madde kullanım sıklığı, RMSE ve yanlılık üzerindeki etkileri karşılaştırıldığında ise genel anlamda bir örüntü elde edilememesine rağmen BSD yönteminin yanlılık değerlerinin sıfıra daha yakın değerler aldığı belirlenmiştir. Bununla birlikte, AMO yönteminin kullanıldığı durumlarda yanlılık değerlerinin tamamının negatif değerler aldığı gözlenmiştir. Bu durum, değişken uzunluklu testlerde AMO yöntemi kullanılması durumunda gerçek yetenek düzeylerine kıyasla daha yüksek kestirimlerde bulunduğu şeklinde yorumlanmaktadır.

Test başlangıcında kullanılan rastgele madde sayısının madde kullanım sıklığı, RMSE ve yanlılık üzerine etkisi değerlendirildiğinde ise daha fazla sayıda rastgele madde yöneltilmesinin yüksek kullanım sıklığı değerine sahip (kullanım sıklığı .40'tan büyük olan) madde sayısını azalttığı belirlenmiştir.

b) Sabit uzunluktaki testler kullanılarak gerçekleştirilen simülasyonlar

Belirlenen durumların sabit uzunluktaki testler üzerindeki etkilerini belirlemek amacıyla gerçekleştirilen simülasyonlarda ise 10, 20 ve 30 maddeden oluşan testlerde farklı yetenek kestirim yöntemleri ile test başlama kurallarının sonuçları incelenmiştir. Sabit uzunluktaki testlerden elde edilen ortalama standart hata değerleri ve gerçek ile kestirilen yetenek düzeyleri arasındaki korelasyon değerleri Tablo 6'da verilmektedir.

Tablo 6. Sabit uzunluktaki testlerin ortalama standart hata değerleri ile gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon değerleri

Yöntem	İlk yetenek kestirimi	Test sonlandırma kuralı					
		10 madde		20 madde		30 madde	
		ort. SH	r	ort. SH	r	ort. SH	r
BSD	ilk maddeden sonra	.310	.946	.231	.973	.194	.980
	rastgele yöneltilen 3 maddeden sonra	.328	.947	.237	.969	.198	.979
	rastgele yöneltilen 6 maddeden sonra	.375	.918	.249	.969	.205	.977
AMO	ilk maddeden sonra	.329	.943	.244	.971	.206	.978
	rastgele yöneltilen 3 maddeden sonra	.348	.935	.245	.969	.210	.980
	rastgele yöneltilen 6 maddeden sonra	.430	.912	.260	.966	.212	.976

Tablo 6 incelendiğinde test uzunluğundaki artışın ortalama standart hata değerlerinde azalmaya ve korelasyon değerlerinde artışa neden olduğu görülmektedir. Hemen hemen tüm durumlarda, testin başında kullanılan rastgele madde sayısındaki artışın gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon değerlerinde azalmaya ve ortalama standart hata değerlerinde artışa neden olmaktadır. Genel bir perspektiften bakıldığında ise, BOBUT uygulamalarında madde seçimine müdahale edildiğinde aynı güvenirlikte sonuç elde edilebilmesi için test uzunluğunun arttırılması gerekmektedir. Bu nedenle, testin başında 6 rastgele madde kullanımı sonucu gerçek yetenek düzeyi ile daha düşük korelasyon değerleri ve daha yüksek ortalama standart hatalar elde edilmiştir. Son olarak, yetenek kestirim yönteminin etkisi incelendiğinde, BSD yönteminin AMO yöntemine göre kısmen daha iyi sonuçlar verdiği görülmektedir.

Sabit test uzunluğunun 10, 20 ya da 30 olarak belirlendiği durumlarda madde kullanım sıklığı dağılımları, RMSE ve yanlılık değerleri incelenmiş ve elde edilen sonuçlar Tablo 7'de paylaşılmıştır.

Tablo 7. Sabit uzunluktaki testlerin madde kullanım sıklığı dağılımları, RMSE ve yanlılık değerleri

Yöntem	İlk yetenek kestirimi	Test sonlandırma kuralı	Madde kullanım sıklığı				RMSE	Yanlılık
			<.01	.01-.20	.21-.40	>.40		
BSD	ilk maddeden sonra	10 madde	247	129	11	6	.324	-.020
		20 madde	209	142	27	15	.229	.001
		30 madde	182	148	36	27	.202	.000
	rastgele yöneltilen 3 maddeden sonra	10 madde	224	154	11	4	.330	-.008
		20 madde	193	162	26	12	.246	.006
		30 madde	163	171	35	24	.207	.002
	rastgele yöneltilen 6 maddeden sonra	10 madde	233	152	6	2	.398	-.008
		20 madde	201	160	24	8	.256	-.012
		30 madde	170	170	31	22	.213	-.013
AMO	ilk maddeden sonra	10 madde	236	141	10	6	.334	.005
		20 madde	205	146	29	13	.252	.006
		30 madde	174	158	34	27	.211	.007
	rastgele yöneltilen 3 maddeden sonra	10 madde	219	160	10	4	.360	-.006
		20 madde	189	167	26	11	.246	-.004
		30 madde	154	176	37	26	.209	-.001
	rastgele yöneltilen 6 maddeden sonra	10 madde	229	157	4	3	.453	-.007
		20 madde	196	164	23	10	.272	-.006
		30 madde	165	175	31	22	.223	-.013

Sabit uzunluktaki testlerde daha uzun testlerin kullanılması durumunda hiç kullanılmayan ya da çok az kullanılan (kullanım sıklığı .01'den küçük) madde sayılarında ciddi azalmalar meydana getirmesi olumlu görünmesine rağmen çok fazla kullanılan (kullanım sıklığı .40'tan büyük) madde sayılarını arttırdığından olumsuz sonuçlara da yol açtığı görülmektedir. Ayrıca, incelenen tüm durumlarda test uzunluğu arttığında RMSE değerleri azalmıştır. Sabit uzunluktaki testlerde yetenek kestiriminin rastgele seçilen 6 maddeden sonra yapılmasının madde kullanım sıklığı değerlerini olumlu yönde etkilediği belirlenmiştir.

Her ne kadar 10, 20 ve 30 madde içeren sabit uzunluktaki testlerde madde kullanım sıklığı değişimleri farklılaşsa da bu durum BSD ve AMO yetenek kestirim yöntemlerinin birbirlerine yönelik üstünlüğünü belirleme açısından yeterli görünmemektedir. Diğer bir deyişle, yetenek kestirim yöntemleri farklılaşsa da madde kullanım sıklığı değişimleri benzer bir dağılım göstermektedir. Ancak, RMSE değerleri irdelendiğinde BSD yönteminde daha düşük değerler aldıklarından BSD yönteminin AMO yöntemine kıyasla daha iyi sonuçlar verdiği söylenebilir.

Çalışmada şu ana kadar yürütülen simülasyon çalışmalarında karşılaştırılan durumlar değerlendirildiğinde sabit uzunluktaki testlerin değişken uzunluktaki testler ile kıyaslandığında daha uygulanabilir sonuçlar verdiği belirlenmiştir. Şöyle ki, değişken uzunluktaki testlerde algoritma belli bir standart hata ile yetenek kestirimini yapmaya çalıştığından bazı katılımcılar için testler, 100'ün üzerinde madde yanıtlanmasıyla sonlanmıştır. Hatta bazı çok düşük ya da çok yüksek yetenek düzeyindeki kullanıcılara madde havuzundaki 393 maddenin tamamı yöneltildiğinde dahi standart hata belirlenen değerin altına düşememiştir. Öte yandan, bazı kullanıcıların yetenek kestirimlerindeki standart hatanın aynı referans değerin altına düşebilmesi için 4 maddenin uygulanması yeterli olmaktadır. Değişken uzunluktaki testlerin kullanıldığı BOBUT uygulamalarında, alt ve üst değerler tanımlanarak test uzunluğu kontrol altına alınmayacaksa bu test sonlandırma kontrolünün TIMSS'in bireye uyarlanmış testlerinde kullanımı uygun görünmemektedir. Peki, sabit uzunluktaki testler daha uygun ise test uzunluğu ne olmalıdır: 10 mu, 20 mi ya da 30 mu? Çalışmadaki 10 maddenin yöneltildiği simülasyon sonuçları incelendiğinde korelasyon katsayılarının farklı yetenek kestirim yöntemleri ve farklı test başlama kurallarında değişkenlik gösterdiği belirlenmiştir. Bu nedenle, kullanılacak sabit uzunluktaki testin incelenen 20 ya da 30 maddeden oluşması daha mantıklı görünmektedir. Test uzunluğunun 20 ve 30 madde olduğu analiz sonuçları incelendiğinde ise 20 maddeden oluşan testten elde edilen gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon katsayısının .97 olduğu aynı değerin 30 maddeden oluşan testlerde .98 olduğu gözlemlenmiştir. Bununla birlikte, testlerden elde edilen ortalama standant hata değerleri incelendiğinde ise 20 madde uzunluğundaki testin .23 ile .26 arasında olduğu ve 30 maddeden oluşan testin ise .19 ile .21 arasında değerler aldığı belirlenmiştir. Buna göre, 20 maddeden oluşan testlerin .93 güvenirlikle ve 30 maddeden oluşan testlerin de .96 güvenirlikle kestirim yaptıkları ortaya çıkmaktadır. RMSE değerleri karşılaştırıldığında ise 20 maddeden oluşan testlerde .229 ile .272 arasında, 30 maddeden oluşan testlerde ise .202 ile .223 arasında değerler aldığı belirlenmiştir. Her iki test uzunluğunun yanlılık düzeyleri karşılaştırıldığında ise gerek 20 gerekse 30 maddeden oluşan testlerin gerçek yetenek düzeylerine yakın yetenek kestirimlerinde bulunduğu görülmüştür.

Tüm bu bilgiler değerlendirildiğinde, 20 maddeden oluşan sabit uzunluktaki testlerin gerek korelasyon katsayılarının gerekse güvenirlik düzeylerinin yüksek olması nedeniyle TIMSS sekizinci sınıf matematik değerlendirmelerinde kullanılabileceği sonucuna varılmıştır.

c) Madde kullanım sıklığı kontrolüne yönelik gerçekleştirilen simülasyonlar

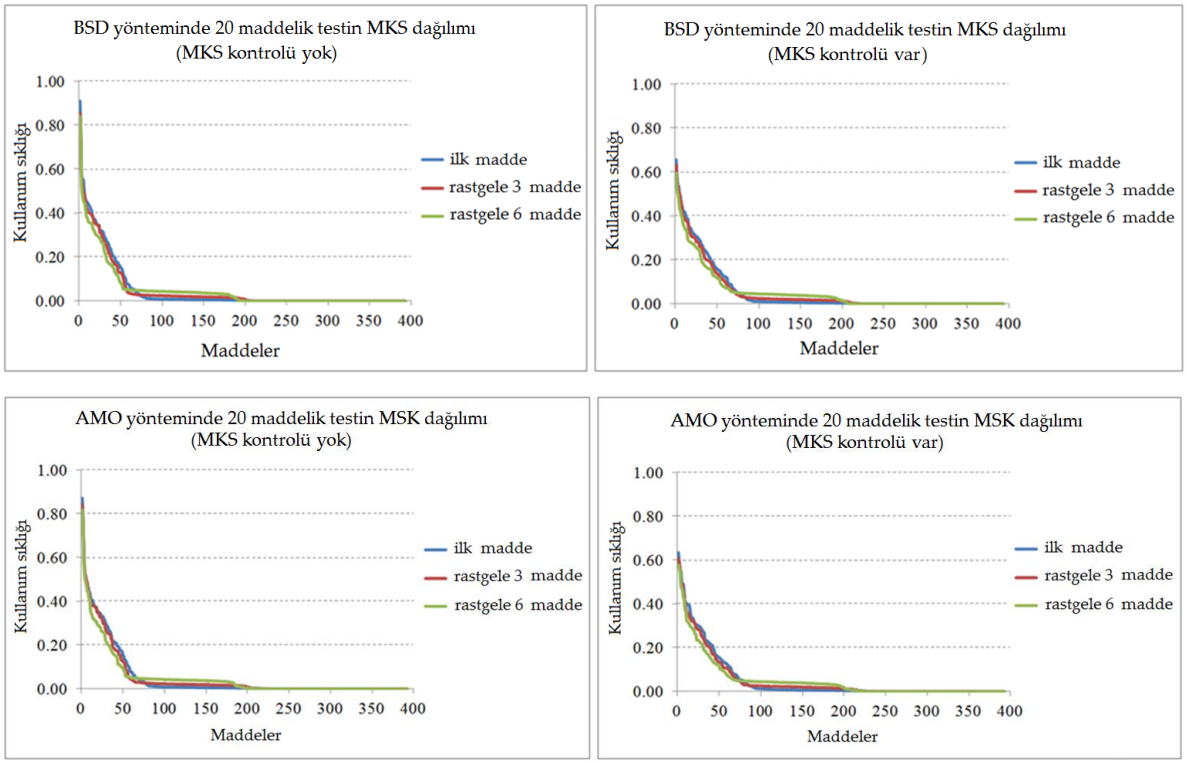
Madde kullanım sıklığı kontrolüne yönelik gerçekleştirilen simülasyonlarda 20 maddeden oluşan sabit uzunluktaki testlerde Randomesque yönteminin etkililiği araştırılmıştır. Gerçekleştirilen simülasyonlardan elde edilen veriler kullanılarak her madde için seçilme olasılığı tanımlanmıştır. Bu değerlerin madde seçiminde kullanılması sonucu düşük kullanım sıklığı değerlerine sahip maddeler daha fazla, yüksek kullanım sıklığı değerlerine sahip maddeler ise daha az sayıda kullanılmıştır. Elde edilen bulgular Tablo 8'de verilmektedir.

Tablo 8. Test uzunluğunun 20 olarak sabitlendiği durumda madde kullanım sıklığı kontrolünün madde kontrol sıklığı dağılımına, RMSE ve yanlılığa etkisi

Yöntem	İlk yetenek kestirimi	MKS kontrolü	Madde kullanım sıklığı				RMSE	Yanlılık
			< .01	.01 - .20	.21 - .40	> .40		
BSD	ilk maddeden sonra	yok	209	142	27	15	.229	.001
		var	197	156	27	13	.245	-.003
	rastgele yöneltilen 3 maddeden sonra	yok	193	162	26	12	.246	.006
		var	178	175	32	9	.242	.005
	rastgele yöneltilen 6 maddeden sonra	yok	201	160	24	8	.256	-.012
		var	189	172	25	9	.266	-.008
AMO	ilk maddeden sonra	yok	205	146	29	13	.252	.006
		var	179	174	31	12	.244	-.006
	rastgele yöneltilen 3 maddeden sonra	yok	189	167	26	11	.246	-.004
		var	170	189	25	13	.267	-.010
	rastgele yöneltilen 6 maddeden sonra	yok	196	164	23	10	.272	-.006
		var	183	179	25	11	.274	-.008

Tablo 8'deki değerler incelendiğinde madde kullanım sıklığı kontrolüne bağlı olarak madde havuzunda hiç kullanılmayan ya da çok az kullanılan maddelerin (kullanım sıklığı .01'den küçük) sayısını azalttığı görülmektedir. RMSE ve yanlılık değerlerinde gerek yetenek kestirim yöntemleri gerekse yetenek kestirim zamanları açısından farklılaşmalar olsa da genellemeye imkan veren bir örüntü gözlenememiştir.

BSD ve AMO yetenek kestirim yöntemlerinde yeteneğin hemen ilk maddeden sonra mı, rastgele 3 maddeden sonra mı yoksa rastgele 6 maddeden sonra mı kestirilmesinin daha iyi olduğunu belirlemek için analizler yapılmıştır. Söz konusu tüm durumlar için havuzda yer alan maddelerin kullanım sıklığı değerleri hesaplanmıştır. Madde kullanım sıklığı değerlerindeki farklılaşmaların daha net biçimde görülebilmesi için değerler büyükten küçüğe sıralanmıştır. Farklı yetenek kestirim yöntemlerine ve test başlangıç kurallarına göre madde kullanım sıklığı kontrolünün etkisini gösteren grafikler Şekil 3'te paylaşılmaktadır. Şekilde düşey eksen maddelerin kullanım sıklığını ifade etmektedir. Yatay eksen ise kullanım sıklığı en yüksek maddeden en düşük maddeye göre soldan sağa sıralanmış havuzdaki maddeleri göstermektedir.



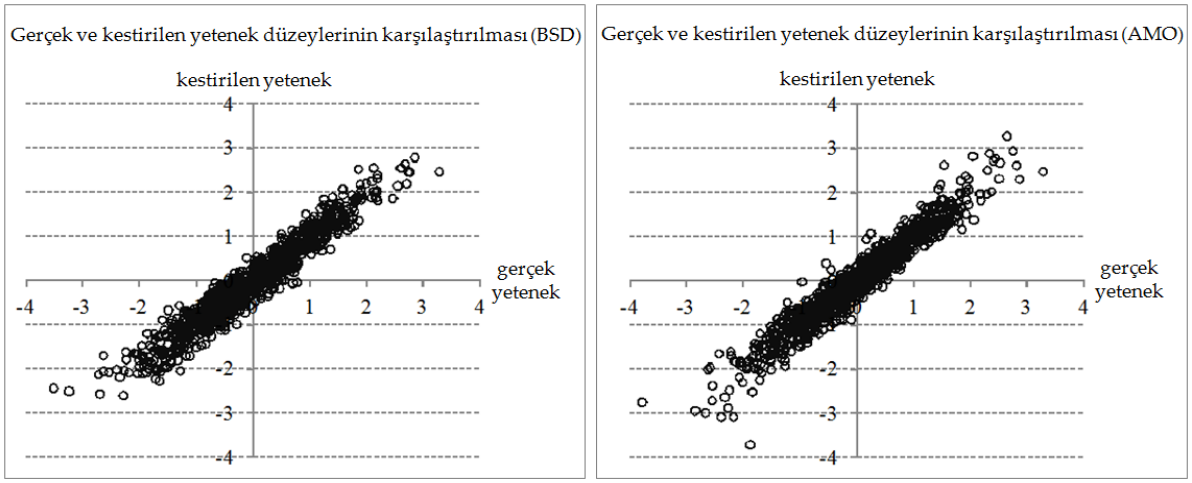
Şekil 3. Farklı yetenek kestirim yöntemlerine ve test başlangıç kurallarına göre madde kullanım sıklığı kontrolünün madde havuzunun kullanımına etkisi

Şekil 3 incelendiğinde madde kullanım sıklığı kontrolünün havuzdaki maddelerin kullanım sıklığı değerleri üzerinde olumlu bir etkisinin olduğu gözlenmesine rağmen maddelerin yaklaşık yarısının hiçbir uygulamada yer almaması önemli bir sorun olarak karşımıza çıkmaktadır. Test başlangıç kuralları karşılaştırıldığında ise yetenek kestiriminin rastgele 6 maddeden sonra yapıldığı durumda yüksek kullanım sıklığı değerleri gözle görülür biçimde azalmıştır. Bunun temel sebebi, testin başlangıcında rastgele madde seçiminin havuzdaki kullanılmayan maddelerin de uygulamalarda yer alabilmesini sağlamasıdır. Dolayısıyla, TIMSS sekizinci sınıf matematik değerlendirmeleri için geliştirilen en uygun BOBUT algoritmasında test başlangıcında rastgele 6 madde yöneltilmesinin havuzdaki maddelerin daha etkili biçimde kullanılabilmesini sağlayacağı düşünülmektedir.

Yetenek kestirim yöntemleriyle ilgili genel bir değerlendirme yapmak gerekirse bu aşamaya kadar gerçekleştirilen simülasyonlardan elde edilen bulgular BSD ve AMO yetenek kestirim yöntemlerinin hemen hemen benzer sonuçlar verdiğini göstermektedir.

d) Yetenek kestirim yöntemini belirlemeye yönelik gerçekleştirilen simülasyonlar

BSD ve AMO yöntemlerinin etkililiğini belirleyebilmek amacıyla gerçekleştirilen simülasyonlarda test başlangıç kuralı olarak rastgele 6 maddenin yöneltildiği, test sonlandırma kuralı olarak sabit uzunluktaki 20 maddeden oluşan testlerin kullanıldığı ve kullanım sıklığı kontrolünün yer aldığı BOBUT algoritması kullanılmıştır. Gerçek ve kestirilen yetenek düzeylerinin ilişkisi Şekil 4'te verilmektedir.



Şekil 4. BSD ve AMO yöntemlerinde gerçek ve kestirilen yetenek düzeylerinin karşılaştırılması

Şekil 4'te, BSD yöntemi ile kıyaslandığında AMO yönteminin çok düşük ve çok yüksek yetenek düzeylerinde daha fazla farklılaşmaya yol açtığı görülmektedir. Bu nedenle, gerçek yetenek düzeyleri ile kestirilen yetenek düzeylerinin genel dağılımına bakıldığında BSD yönteminin AMO yöntemine kıyasla daha iyi kestirimde bulunduğu ve daha tutarlı sonuç verdiği ifade edilebilir.

Simülasyon sonuçlarını özetlemek gerekirse, teste yetenek kestiriminin rastgele seçilen 6 madde sonunda başlandığı, BSD yetenek kestirim yönteminin kullanıldığı, 20 madde uygulandıktan sonra testin sonlandığı ve madde kullanım sıklığı kontrolünün kullanıldığı algoritmanın TIMSS sekizinci sınıf matematik uygulamaları için çalışma kapsamındaki en uygun algoritma olduğu belirlenmiştir. Bu durumda, ortalama standart hata .253 (en küçük değer .135 ve en büyük değer .468) olarak ve gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon katsayısı ise .964 olarak hesaplanmıştır.

TARTIŞMA ve SONUÇ

Bu araştırmada, kağıt kalem testlerinin kullanıldığı TIMSS sekizinci sınıf matematik değerlendirmesine alternatif olabilecek en uygun BOBUT algoritmasının belirlenebilmesi amaçlanmıştır. Yürütülen simülasyonlarda, farklı test başlama kuralları, yetenek kestirim yöntemleri, test sonlandırma kuralları karşılaştırılmış ve madde kullanım sıklığı kontrolünün etkisi incelenmiştir.

Çalışma kapsamında test başlama kuralı olarak tanımlanan ve testin hemen başında rastgele sayıda madde kullanımının incelendiği senaryolarda yetenek kestiriminin ilk maddeden sonra, rastgele yöneltilen 3 maddeden sonra ve rastgele yöneltilen 6 maddeden sonra yapılması durumları karşılaştırılmıştır. Her ne kadar, testin başında rastgele seçilerek uygulanan madde sayısındaki artışın RMSE değerleri üzerinde az da olumsuz etkileri görülse de madde kullanım sıklığı değerleri üzerindeki olumlu etkileri düşünüldüğünde testin başında rastgele 6 madde yöneltmenin en uygun algoritmanın belirlenmesine önemli bir yere sahip olduğu sonucuna varılmıştır. Ancak, bu durumun yeterli uzunluktaki testler için geçerli olduğu belirtilmelidir. Diğer bir deyişle, 10 maddeden oluşan sabit uzunluktaki testlerde ya da standart hata referans değerinin .40 olarak tanımlandığı değişken uzunluktaki testlerde bu durum geçerli değildir çünkü 6 madde, bu testlerde toplam test uzunluğunun büyük kısmını oluşturmaktadır.

Çalışmada karşılaştırılan yetenek kestirim yöntemlerinin simülasyon sonuçları incelendiğinde BSD ve AMO yöntemlerinin gerçek yetenek düzeylerini belirlemede benzer sonuçlar verdikleri görülmüştür. Ancak, AMO yöntemi ile kıyaslandığında BSD yönteminin özellikle çok düşük ve çok yüksek yetenek düzeylerindeki bireylerin yetenek düzeylerini daha düşük standart hatalar ile kestirdiği belirlenmiştir. Bu sonuç Gu ve Reckase (2007) tarafından yapılan araştırma bulgularıyla benzerlikler taşımaktadır.

En uygun test sonlandırma kriterinin belirlenmesi amacıyla değişken uzunluktaki testler ile sabit uzunluktaki testler karşılaştırılmıştır. Değişken uzunluktaki testlerin kullanıldığı simülasyonlarda .20 standart hatanın kullanıldığı algoritmanın daha yüksek korelasyon değerlerine sahip olduğu belirlenmiştir. Fakat, standart hatanın .20 olarak belirlendiği bazı simülasyonlarda havuzdaki tüm maddeler kullanılmasına rağmen standart hatanın .20'nin altına inemediği görülmüştür. Sonuç olarak, bu durum gerçek hayatta uygulanabilir değildir. Daha ayrıntılı bir biçimde açıklamak gerekirse, çok düşük ya da çok yüksek başarı düzeyine sahip bireylerin yetenek kestirimlerinin en fazla .20 standart hata ile yapılmak istenmesi durumunda, bu bireylere havuzdaki 393 maddenin tamamının uygulanması durumunda bile bu amaca ulaşılamayabilir. Benzer sonuçlar Gökçe ve Berberoğlu (2015) tarafından gerçekleştirilen çalışmada da vurgulanmaktadır. Bu nedenle, TIMSS sekizinci sınıf uygulanmalarında sabit uzunluktaki test kullanımının daha uygun olacağı sonucuna varılmıştır. Sabit uzunluktaki testlerin karşılaştırıldığı simülasyonlarda 10, 20 ve 30 maddeden oluşan testler kullanılmıştır. Elde edilen yüksek korelasyon değerleri ve diğer bulgular 20 maddeden oluşan testin TIMSS sekizinci sınıf matematik değerlendirmelerinde tercih edilebilecek bir uzunluk olduğunu göstermektedir. Çalışma kapsamında kullanılan Randomesque kontrol yönteminin özellikle yüksek kullanım sıklığı değerlerine sahip maddelere uygulamalarda daha az yer vererek, hiç kullanılmayan maddelere ise uygulamalarda az da olsa yer vererek madde havuzunun daha dengeli biçimde kullanılmasına katkı sağlamıştır. Ancak yine de havuzda yer alan TIMSS maddelerinin yaklaşık yarısı hiçbir uygulamada kullanılmamıştır. BOBUT uygulamalarında karşılaşılabilecek en büyük sorunlardan biri madde havuzunda hiç kullanılmayan ya da çok fazla kullanılan maddelerdir (Eggen, 2001; Eggen & Straetmans, 2000). Madde kullanım sıklığı oranlarını dengelemek için daha farklı madde kullanım sıklığı kontrollerinin karşılaştırılması önerilmektedir.

TIMSS sekizinci sınıf matematik değerlendirmelerinde kullanılan başarı testlerinde ortalama test uzunluğu yaklaşık 30'dur. Bu testlerde EAP yöntemi kullanılarak ortalama .328 standart hata ile yetenek kestirimi yapılmaktadır. Öte yandan, BOBUT uygulamaları EAP yöntemi kullanarak 20 madde ile .253 ortalama standart hata ile kestirim yapmaktadır. Bu durumda, TIMSS uygulamalarında BOBUT'un standart kağıt kalem uygulamaları ile karşılaştırıldığında yaklaşık %35 daha az madde ile daha güvenilir kestirimde bulunduğu söylenebilir. Bu sonuç aslında BOBUT uygulamalarının en önemli avantajlarından biri olarak karşımıza çıkmaktadır. Alanyazında, kağıt kalem testleriyle karşılaştırıldığında bireye uyarlanmış testlerin daha az madde ile daha güvenilir kestirimlerde bulunduğu ve test uzunluğu ile uygulama süresini kısalttığı (Eggen, 2007; Hambleton vd., 1991; Meijer & Nering, 1999; Mills & Stocking, 1996; Verschoor & Straetmans, 2010) ifade edilmektedir.

Çalışmada yer alan tüm durumlar için gerçek ve kestirilen yetenek parametreleri arasında pozitif yönde yüksek korelasyon katsayıları elde edilmiştir. Yapılan araştırmalar farklı test başlama kuralları, yetenek kestirim yöntemleri ve test sonlandırma kuralları kullanılarak kestirildiğinde de benzer sonuçlar alınabileceğini ifade etmektedir (Kalender, 2011; Kezer & Koç, 2014).

TIMSS sekizinci sınıf uygulamasının bilgisayar ortamında test olarak uygulanabilirliğini inceleyen ve en uygun test algoritmasının belirlenmesini amaçlayan bu çalışmanın bazı sınırlılıkları bulunmaktadır. Öncelikle, test başlangıç kuralı olarak farklı güçlük düzeyindeki maddeler ile teste başlanması durumu bu çalışmada iredelenmemiştir. Bunun yerine, teste farklı sayıda rastgele madde ile başlamanın etkisi araştırılmıştır. Ayrıca, madde havuzunda iki ve çok kategorili olmak üzere iki tip soru yer almaktadır. Kağıt kalem kullanılarak gerçekleştirilen TIMSS uygulamalarında kontrol altında tutulabilen soru tipleri bireye uyarlanmış test algoritmasının madde seçiminde dikkate alınmamıştır. Benzer durum içerik kontrolünde de söz konusudur. Uygulanan TIMSS sekizinci sınıf matematik başarı testlerinde 4 farklı öğrenme alanından (sayılar, geometri, cebir ile veri ve olasılık) sorular yer almakta ve böylece her öğrenme alanından istenen oranda soru seçilip uygulanabilmektedir. Bireye uyarlanmış test algoritmasının madde seçiminde içerik kontrolü ile ilgili de bir simülasyon çalışması gerçekleştirilmemiştir. Son olarak, çalışma kapsamında kullanılan madde havuzunda açık uçlu maddeler de yer almaktadır. Simülasyonlarda bu maddeler kullanılarak yetenek kestirimi gerçekleştirilmesine rağmen gerçek yaşamda kullanılacak BOBUT uygulamalarında bu tür sorular için anında yetenek kestirimi yapılması zor görünmektedir. Bu durum da çalışmanın bir sınırlılığı olarak karşımıza çıkmaktadır.

Araştırmada TIMSS'in 2007, 2011 ve 2015 yıllarındaki uygulamalarına katılan Türkiye ve ABD'ye ait veriler kullanılmıştır. Gelecekte yürütülecek çalışmalarda TIMSS uygulamalarına 1995, 1999, 2003, 2007, 2011 ve 2015 yıllarında katılan farklı ülkelere ait veriler kullanılarak çalışmalardan elde edilen bulgular karşılaştırılabilir. Sekizinci sınıf matematik verilerinin kullanıldığı bu çalışma, dördüncü sınıf matematik verileri kullanılarak incelenebilir ve her iki uygulamanın karşılaştırılabilir sonuçlar verip vermediğine bakılabilir.

KAYNAKÇA

Davey, T., & Pitoniak, M. J. (2006). Designing computerized adaptive tests. *Handbook of Test Development*, 543-574. Routledge.

Eggen, T. J. H. M. (2001). Overexposure and underexposure of items in computerized adaptive testing. *Measurement and Research Department Reports, 1*.

Eggen, T. J. H. M. (2004). Contributions to the Theory and Practice of Computerized Adaptive Testing. Dissertation. Print Partners Ipskamp B.V., Enschede.

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.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 60(5), 713-734.

Glas, C. A. W. (2010) MIRT: Multidimensional Item Response Theory. (Computer Software). University of Twente. Retrieved from https://www.utwente.nl/nl/bms/omd/Medewerkers/medewerkers/glas/#software

Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, 35, 83-88.

Gökçe S. & Berberoğlu G. (2015) Comparison of Linear, Computerized Adaptive and Multi Stage Adaptive Versions of the Mathematics Assessment of Turkish Pupil Monitoring System. In: Millsap R., Bolt D., van der Ark L., Wang WC. (Eds) Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics, vol 89. Springer, Cham.

Gu, L., Reckase M. D. (2007). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

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

Kalender, I. (2011). Effects of different computerized adaptive testing strategies on recovery of ability. Yayınlanmamış Doktora Tezi. Middle East Technical University, Ankara.

Kezer, F. & Koç, N. (2014). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [A comparison of computerized adaptive testing strategies]. *Eğitim Bilimleri Araştırmaları Dergisi - Journal of Educational Sciences Research*, 4 (1), 145-174.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359-375.

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.

Masters, G. N. (2016). Partial credit model. In *Handbook of Item Response Theory, Volume One* (pp. 137-154). Chapman and Hall/CRC.

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.

Mullis, I. V. S., Martin, V. & Loveless, T. (2016). 20 years of TIMSS, international trends in mathematics and science achievement, curriculum, and instruction. IEA, TIMSS&PIRLS International Study Center Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). Methods and Procedures in TIMSS 2015. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/publications/timss/2015-methods.html

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*.

Smits, N., van Straten, A., & Cuijpers, P. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188(1), 147-155.

van der Linden, W. J. (1995). Advances in computer applications. In T. Oakland & R. K. Hambleton (Eds.), *International Perspectives on Academic Assessment*, (pp. 105-124). Kluwer Academic Publishers.

van der Linden, W. J. (2001). Computerized test construction. *Research Report*. Twente University, Enschede (Netherlands).

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.

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

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