Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) ??: SS-SS [201?]
doi: 10.16986/HUJE.2016024284

Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinde Madde Havuzu Özelliklerinin Test Uzunluğu ve Sınıflama Doğruluğu Üzerindeki Etkisi*

The Effects of Item Pool Characteristics on Test Length and Classification Accuracy in Computerized Adaptive Classification Testings

Ceylan GÜNDEĞER**, Nuri DOĞAN***

• ***Geliş Tarihi:*** *26.10.2016* • ***Kabul Tarihi:*** *17.12.2016* • ***Yayın Tarihi:*** *28.12.2016*

ÖZ: Bu çalışmada bireyselleştirilmiş bilgisayarlı sınıflama testlerinde (BBST) madde havuzu özelliklerinden dağılım ve büyüklüklerin ortalama test uzunluğu ve ortalama sınıflama doğruluğu üzerindeki etkisi incelenmiştir. Bu amaçla, sivri ve basık dağılımlı 50, 100, 200 ve 300 maddelik madde havuzlarında; tesadüfi madde seçme yöntemi (TMSY), Maksimum Fisher Bilgisi (MFB) ve Kullback-Leibler Bilgisi (KLB) yöntemleri incelenmiştir. 1000 bireye ait yetenek parametreleri -3,3 aralığında N(0,1) olacak şekilde türetilmiştir. Sivri dağılıma sahip madde havuzlarındaki maddelerin a parametresi U[0,5; 2,0] aralığından; b parametresi N(1, 0,4) ve c parametresi N(0,15, 0,05) şeklinde; basık dağılıma sahip madde havuzlarındaki maddeler ise a parametresi U[0,5; 2,0] aralığından; b parametresi N(1, 1,5) ve c parametresi N(0,15, 0,05) şeklinde türetilmiştir. R'da gerçekleştirilen simülasyon sonucunda tüm madde havuzlarında ortalama test uzunluğu bakımından en yüksek değerin TMSY'ye ait olduğu; MFB ve KLB yöntemlerinin birbirine oldukça benzer çalıştıkları söylenebilir. Madde havuzu büyüklüğü arttıkça test uzunluklarının kısaldığı; sınıflama doğruluklarının azaldığı ancak tüm koşullarda 0,90 üstünde yüksek sınıflama doğruluğu elde edildiği görülmüştür. Ayrıca sivri dağılıma sahip madde havuzlarında test uzunluğunun kısaldığı ve test etkililiğinin arttığı; sınıflama doğruluklarının ise değişmediği görülmüştür. Bu sonuçlar dikkate alındığında, BBST'de çok sayıda maddeden oluşan sivri dağılıma sahip madde havuzları ile yüksek sınıflama doğruluğuna sahip daha kısa testlerin oluşturulabileceği söylenebilir.
Anahtar sözcükler: bireyselleştirilmiş bilgisayarlı sınıflama testleri, madde havuzu dağılımı, madde havuzu büyüklüğü, test uzunluğu, sınıflama doğruluğu

ABSTRACT: In this study it was investigated that the effects of distributions and sizes on average test length and average classification accuracy in computerized adaptive classification testings (CACT). For that purpose random item selection method (RISM), Maximum Fisher Information (MFI) and Kullback-Leibler Information (KLI) were studied in broad and peaked item pools with 50 items, 100 items, 200 items and 300 items. Thetas are derived from N(0,1). In peaked item pools items are simulated from U[0,5; 2,0] for a parameters, N(1, 0,4) for b parameters and N(0,15, 0,05) for c parameters; and in broad item pools items are simulated from U[0,5; 2,0] for a parameters, N(1, 1,5) for b parameters and N(0,15, 0,05) for c parameters. The simulation study was performed in R results show that RISM has the maximum value with respect to average test length; and MFI and KLI perform similar. The more items in the pool, the shorter test length and fewer the classification accuracy but in all conditions classification accuracy has high rate above 90%. In addition, in peaked item pools it is seen that the average test lengths are getting shorter and the test effectiveness is getting higher; but the classification accuracies are not changing. In conclusion it can be said that with the peaked item pools with more items, CACT provides shorter tests and high classification accuracy.

Keywords: computerized adaptive classification testing, item pool distribition, item pool size, test length, classification accuracy

* Bu çalışma 1-3 Eylül 2016 tarihleri arasında Antalya'da düzenlenmiş olan 5. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.

** Arş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Ankara-Türkiye, cgundeger@gmail.com

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Ankara-Türkiye, nuridogan2004@gmail.com

1. GİRİŞ

Bireyselleştirilmiş bilgisayarlı testler (BBT; Computerized Adaptive Testing: CAT) günümüzde birçok alanda popülerleşmektedir. Bilgisayarların ulaşılabilirliği ve Madde Tepki Kuramı'nın (MTK) sunduğu avantajlar sayesinde BBT çok sayıda ortamda kullanılmaktadır (Babcock ve Weiss, 2009). BBT'de test, güvenirliği yüksek ölçme sonuçları elde edebilmek amacıyla bireylerin yetenek düzeyine göre ayarlanmaktadır. Bu testlerin sınıflama amacına sahip olması durumunda ise bireyselleştirilmiş bilgisayarlı sınıflama testleri (BBST; Computerized Adaptive Classification Testing: CACT) gündeme gelmektedir. BBST'nin temel amacı bireyleri, önceden belirlenmiş bir (ya da daha fazla) kesme noktasına göre en az sayıda maddeyle ve yüksek sınıflama doğruluğunda sınıflara ayırmaktır.

BBT ve BBST'nin dayandığı temel kuram MTK'dır. MTK, bireyin test performansı altında yatan gözlenemeyen yetenek düzeyi ile gözlenen performansı arasındaki ilişkiyi matematiksel modellerle belirleyen bir kuramdır. MTK'da, madde karakteristik eğrisi (MKE) yardımıyla belli bir yetenek düzeyindeki öğrencinin maddeyi doğru cevaplama olasılığı kestirilir. MKE, yetenek ölçeğindeki farklı noktalar için maddenin doğru cevaplanma olasılığını verir. Bu eğri bireyin örtük özelliği ile madde performansı arasında monoton artan bir fonksiyondur. Buna göre yüksek yetenek düzeyindeki bireyin maddeyi doğru cevaplama olasılığı, düşük yetenek düzeyindeki bireyin maddeyi doğru cevaplama olasılığından daha yüksektir (Hambleton ve Swaminathan, 1985).

MTK'da uygun modelden beklenen iki özellik madde ve birey parametrelerinin değişmezliğidir. MTK'nın Klasik Tesk Kuramı'ndan belki de en üstün özelliği madde parametrelerinin gruptan ve birey parametrelerinin maddelerden bağımsızlığıdır ve bu iki özellik BBT ve BBST uygulamaları için önemli bir temel teşkil etmektedir. MTK sayesinde BBT ve BBST'de bireyin yeteneği cevaplamış olduğu maddelerden bağımsız kestirilebilmektedir. Böylece bireyleri, farklı maddeleri cevaplamış olsalar bile, karşılaştırabilmek mümkündür (Hambleton ve Swaminathan, 1985).

1.1. Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST)

BBT uygulamaları genel olarak, (i) Tepki modeli; (ii) Madde havuzu; (iii) Başlama kuralı; (iv) Madde seçme yöntemi; (v) Yetenek kestirim yöntemi ve (vi) Sonlandırma kuralı olmak üzere altı ana bileşenden oluşmaktadır (Weiss ve Kinsbury, 1984). BBST'de ise sonlandırma kuralı yerine sonlandırma kriterleri kullanılmaktadır. Bu bileşenlerden tepki modeli MTK kapsamında hangi modelin kullanılacağının belirlenmesi amacını taşımaktadır. Bu noktada tepki modeli, çoklu puanlanan maddelere dayanan Ardışık Tepki Modeli, Kısmi Kredi Modeli vb. olabildiği gibi; ikili puanlanan maddeleri temel alan 1 Parametreli Lojistik Model (1 PLM), 2 PLM veya 3 PLM olabilmektedir. Bu çalışmada, bireyleri başarılı-başarısız, geçti-kaldı vb. gruplara sınıflama amacıyla oluşturulan testlerin yüksek risk (high-stakes testing) içermesine bağlı olarak simülasyonda şans parametresini (c) de dikkate alabilmek amacıyla 3 PLM temel alınmıştır. Bu modelde maddelerin ayırıcılık (a) ve güçlük (b) parametreleri değişkenlik gösterdiği gibi maddelere ait şans parametresi (c) de söz konusudur. 3 PLM tahmin faktörünün bulunduğu çoktan seçmeli testlerde sıklıkla tercih edilen bir modeldir. 3 PLM'ye ait MKE formülü aşağıdaki gibi ifade edilmektedir (Hambleton ve Swaminathan, 1985):

$$Pi(\theta) = ci + (1 - ci)\frac{e^{Dai(\theta - bi)}}{1 + e^{Dai(\theta - bi)}} \qquad (1)$$

BBST'nin en önemli bileşenlerinden biri de kalibre edilmiş madde havuzudur. Madde havuzu büyüklüğü ve madde parametrelerinin dağılımı gibi madde havuzu özellikleri, BBST'de ölçme ve sınıflama etkililiğini etkileyen faktörler arasındadır. Madde havuzu geliştirilirken ne kadar sayıda maddeye ihtiyaç duyulacağı ve madde parametrelerinin dağılımının nasıl olacağı

cevaplanması gereken sorulardır. Bu soruların cevabı birçok koşula bağlıdır. Örneğin yüksek risk içeren (high-stakes) ve hatanın çok düşük bir miktarı tolere edilebilen testlerde, diğer testlere kıyasla daha çok sayıda maddeye gereksinim duyulacaktır. Eğer testte MTK temel alınacaksa yine fazla sayıda madde gerekecektir. Test bilgi fonksiyonunun dağılımı da çalışmanın amacıyla örtüşmelidir. Örneğin BBST, kesme noktasına göre bireyleri sınıflama amacı içermektedir. Bu durumda kesme noktası etrafında yüksek bilgi veren test bilgi fonksiyonuna ihtiyaç duyulacaktır. Test bilgi fonksiyonunun dağılımının sivri veya basık olması durumunda veya havuz büyüklüğünün farklı koşullarında bireyselleştirilmiş test algoritmasının nasıl performans göstereceği simülasyon çalışmaları yardımıyla önceden belirlenebilmektedir (Thompson ve Weiss, 2011). Flaugher'a (2000) göre madde havuzunun kalitesi ne kadar iyiyse bireyselleştirilmiş test algoritması da o kadar başarılı performans gösterecektir (Flaugher, 2000). Madde havuzu özelliklerinin BBST'de bu denli önemli olmasına karşın alanyazın incelendiğinde havuz büyüklüğü ve test bilgi fonksiyonu dağılımı gibi madde havuzu özelliklerinin çoğu çalışmada raporlanmamış olduğu göze çarpmaktadır. Bu çalışmada sivri ve basık dağılıma sahip 50, 100, 200 ve 300 maddeden oluşan toplam 8 madde havuzu ele alınmıştır.

BBST'nin üçüncü bileşeni olan başlama kuralı teste hangi düzeyden başlanacağını ifade etmektedir. Testin tekrarlı olarak alınabildiği durumlarda testi ikinci kez alanların başlama noktası bir önceki testten kestirilen yetenek düzeyleri olabilir. Bunun dışında testi alan popülasyonun ortalaması başlama noktası olarak atanabilir (Thompson, 2007). Bu çalışmada başlama noktası tüm koşullar için $\theta = 0$ olarak belirlenmiştir.

BBST'nin önemli bir diğer bileşeni ise madde seçme yöntemidir. Madde seçme yöntemi, testin etkililiğini (madde sayısını) ve doğruluğunu belirler (Thompson, 2009). Bu çalışmada incelenen madde seçme yöntemleri Tesadüfi Madde Sesim Yöntemi (TMSY), Maksimum Fisher Bilgisi (MFB) ve Kullback-Leibler Bilgisi (KLB)'dir. TMSY, maddenin havuzdan testin her noktasında tesadüfi seçilmesine dayanan basit bir yaklaşımdır (Kingsbury & Weiss, 1983; Akt: Thompson, 2007). MFB bilginin tek bir noktada maksimize edilmesini sağlarken (Embretson ve Reise, 2000); KLB θ_0'dan θ_1'e kadar olan bölgedeki bilgiyi değerlendirir (Eggen, 1999; Akt: Thompson, 2007). Kestirim temeli madde seçim yöntemleri bireyselleştirilmiş (bireye uyarlanmış) madde seçimleri olarak düşünülebilir çünkü bireysel olarak öğrencinin kestirilen yetenek düzeyini dikkate almakta ve öğrenci cevabının bir vektörü olmaktadır. Böylece bireysel olarak öğrencinin yetenek düzeyine uygun madde seçilmektedir. Ancak TMSY'nin, bireylere ait kestirilen geçici yetenekler veya madde parametrelerini dikkate almaması, sadece tesadüfi olarak havuzdan madde seçmesi sebebiyle bireyselleştirilmiş madde seçme yöntemleri grubuna dâhil edilemeyeceği ve onlar kadar etkili olmayacağı düşünülebilir.

BBST'nin beşinci bileşeni yeteneğin kestirilmesidir ve bu bileşen son sınıflama kararlarının etkililiği ve uygunluğu bakımından oldukça önemli bir değişkendir (Yang, Poggio ve Glasnapp, 2006). Bu değişken, raporlanan son yetenek kestirimini etkilediği gibi madde seçimi ve test sonlanmasını da etkilemektedir (Wang, 2011). Alanyazında birçok yetenek kestirim yöntemi tanımlanmış olsa da bu çalışmada ortalama madde sayısını azaltması ve hızlı kestirim yapabilmesi bakımından Beklenen Sonsal Dağılım (BSD; Expected a Posteriori) yöntemiyle yetenek kestirimleri tamamlanmıştır. BSD iteratif olmayan birikimli bir süreç içermesi ve pozitif sonsuz veya negatif sonsuz cevap örüntülerinde de sonuç verebilmesi bakımından güçlü ve tercih edilebilir bir yöntemdir. (Embretson ve Reise, 2000). BSD temelde sonsal dağılımın ortalamasını ve varyansını bulmaya odaklanır ve normallik varsayımı gerektirmez (Wang ve Vispoel, 1998). BSD'nin önsel dağılıma dayanması sebebiyle önsel dağılımın tanımlanmasının doğruluğu önem arz etmektedir. Önsel dağılım ne kadar doğru tanımlanırsa BSD kestirimleri hatadan o kadar arınık olacaktır. Simülasyon çalışmalarında önsel dağılım araştırmacı tarafından belirlendiği için BSD kestirimleri daha az hataya sahip olmaktadır.

BBST'nin son bileşeni sonlandırma kriteridir. Alanyazın incelendiğinde çalışmalarda birçok sonlandırma kriterinin çalışıldığı görülmektedir. Bu çalışmada sonlandırma kriterlerinden

Genelleştirilmiş Olabilirlik Oranı (GOO; Generalized Likelihood Ratio: GLR) ele alınmıştır. Bunun sebebi yöntemin diğerlerine kıyasla simülasyonda daha hızlı sonuç vermesidir. GOO temelde iki hipotezi test etme sürecini içermektedir. GOO sonucuna göre öğrencinin testi almaya devam edip etmeyeceğine ve etmeyecekse hangi sınıfa yerleştirileceğine karar verilir.

1.2. Araştırmanın Amacı ve Önemi

BBST'nin yukarıda açıklanan altı bileşeninden en önemlisi madde havuzu ve madde seçme yöntemidir. Alanyazın incelendiğinde çalışmalarda genellikle madde havuzunun dağılımı hakkında bilgi yer almazken sadece madde havuzu büyüklükleri tanımlanmaktadır. Ancak madde havuzunun şekli de madde sayısı kadar önemli bir bileşendir. Bu çalışmada BBST'de madde havuzu dağılımının ve madde havuzu büyüklüklerinin ortalama test uzunluğu ve ortalama sınıflama doğruluğu üzerindeki etkisi incelenmiştir.

Son yıllarda yurt içi ve yurt dışında yoğun olarak çalışılan BBT çalışmalarının bir alt dalı olan BBST çalışmaları incelendiğinde, konunun özellikle yurt dışı alanyazında çalışılmış olduğu ve Türkiye'de hiç çalışılmamış olduğu görülmektedir. Bu çalışma ülkemizde yapılan *ilk* BBST çalışması olması bakımından önem taşımaktadır. Ayrıca teknolojinin gelişmesi ve eğitimin çağa ayak uydurma çabasının bir sonucu olarak ülkemizde de bilgisayarlı sınavlara doğru bir yönelim olduğu söylenebilir. Buna göre yakın zamanda bilgisayarlı sınıflama testleri uygulamaya koyulabilir. Bu noktada çalışmanın uygulayıcılara, madde havuzlarının dağılımı ve büyüklüğü; madde seçme yöntemlerinin etkililiği hakkında bilgi sağlaması beklenmektedir.

2. YÖNTEM

2.1. Araştırmanın Türü

Bu çalışma, "… olsa ne olurdu?" sorusuna cevap arayan bir Monte Carlo simülasyon çalışmasıdır (Dooley, 2002). Çalışmada hem bireylere ait yetenek parametreleri hem de oluşturulan madde havuzlarının parametreleri R'da araştırmacı tarafından türetilmiştir. Bireylerin yetenek parametreleri (θ) , (-3,+3) θ düzeyleri aralığında, ortalaması 0 ve standart sapması 1 olacak şekilde normal dağılım yardımıyla toplam 1000 kişi üzerinden random türetilmiştir.

2.2. Veri Üretimi

Bu simülasyon çalışmasında birey parametreleri gibi madde parametreleri de simülatif veriden oluşmaktadır. Madde havuzları 3 PLM temel alınarak yapılandırılmıştır. Bu amaçla sivri ve basık dağılıma sahip 50, 100, 200 ve 300 maddelik madde havuzları oluşturulmuştur. Sivri dağılıma sahip madde havuzlarındaki maddelerin a parametresi tekbiçimli (uniform) dağılım yardımıyla U[0,5; 2,0] aralığından; b parametresi normal dağılım yardımıyla N(1, 0,4) ve c parametresi de yine normal dağılım yardımıyla N(0,15, 0,05) şeklinde; büyüklükleri 50, 100, 200 ve 300 madde olacak şekilde türetilmiştir. Basık dağılıma sahip madde havuzlarındaki maddeler ise a parametresi U[0,5; 2,0] aralığından; b parametresi N(1, 1,5) ve c parametresi N(0,15, 0,05) şeklinde ve yine büyüklükleri 50, 100, 200 ve 300 madde olacak şekilde türetilmiştir. Oluşturulan madde havuzlarının test bilgi fonksiyonu grafikleri Ek 1 ve Ek 2'de yer almaktadır.

2.3. Verilerin Analizi

Bu araştırmanın bağımsız değişkenlerini, tesadüfi madde seçme yöntemi (TMSY), Maksimum Fisher Bilgisi (MFB) ve Kullback-Leibler Bilgisi (KLB) olmak üzere üç madde

seçme yöntemi ve sivri ile basık dağılıma sahip 50, 100, 200 ve 300 maddeden oluşmak üzere sekiz adet madde havuzu oluşturmaktadır. Çalışmanın bağımlı değişkenleri ise ortalama test uzunluğu ve Cohen'in Kappası'yla hesaplanan ortalama sınıflama doğruluğudur. Buna göre çalışmada ***üç madde seçme yöntemi x dört madde havuzu büyüklüğü x iki madde havuzu dağılımı = 24 koşul*** oluşturulmuştur.

Yetenek parametrelerinin türetilmesi ve madde parametrelerinin türetilerek madde havuzlarının oluşturulması aşamalarından sonra, BBST simülasyonu her bir koşul için yazılan döngülerle 25 tekrarla R'da tamamlanmıştır (R Core Team, 2013). Veri analizinde bağımlı değişkenler olan ortalama test uzunluğu ve ortalama sınıflama doğruluğuna ilişkin değerler 25 tekrarın ortalaması olacak şekilde araştırmacı tarafından yazılan fonksiyonlarla R'dan çekilmiştir. Ortalama sınıflama doğruluğunda Cohen'in Kappa istatistiğinden yararlanılmıştır. Cohen (1960) tarafından geliştirilen Kappa istatistiği, iki veya daha fazla gözlemcinin yaptığı değerlendirmeler arasındaki uyuşmayı belirlemek için kullanılır. Bu uyum -1 ile +1 arasında değer alır. Sıfır değeri tesadüfî uyuşmayı, negatif değerler tesadüfî olmaktan daha kötü bir uyuşmayı ve +1 değeri ise mükemmel uyuşmayı temsil eder (Şencan, 2005).

3. BULGULAR

BBST simülasyon sonuçlarına göre basık ve sivri dağılımlı farklı büyüklükteki madde havuzları üzerine oluşturulan koşulların 25 tekrara dayanan ortalama test uzunlukları ve ortalama sınıflama doğrulukları Tablo 1'de yer almaktadır. Tablo 1'e göre oluşturulan tüm madde havuzlarında ortalama test uzunluğu bakımından en yüksek değerin TMSY'ye ait olduğu; bunu takiben MFB ve KLB madde seçme yöntemlerinin geldiği ve bu iki yöntemin birbirine oldukça benzer performans gösterdiği görülmektedir. Bu bulguya dayanarak BBST'de madde havuzu büyüklüğü fark etmeksizin, MFB veya KLB madde seçme yöntemlerinin TMSY'ye kıyasla test uzunluğunu kısaltarak test etkililiğini artırdığı söylenebilir.

Tablo 1: Koşullara ait ortalama test uzunlukları ve ortalama sınıflama doğrulukları

Madde Havuzu Büyüklüğü	Madde Seçme Yöntemi	Ortalama Test Uzunluğu (Basık)	Ortalama Sınıflama Doğruluğu (Basık)	Ortalama Test Uzunluğu (Sivri)	Ortalama Sınıflama Doğruluğu (Sivri)
50 madde	TMSY	27,00	0,99	23,78	0,99
	MFB	22,96	0,99	19,87	0,99
	KLB	23,05	0,99	20,01	0,99
100 madde	TMSY	26,53	0,95	22,87	0,96
	MFB	19,53	0,98	16,90	0,98
	KLB	19,58	0,98	16,90	0,98
200 madde	TMSY	25,79	0,94	22.20	0,95
	MFB	17,25	0,97	15.67	0,97
	KLB	17,27	0,97	15.88	0,97
300 madde	TMSY	25,10	0,94	21,99	0,95
	MFB	16,61	0,97	15,38	0,97
	KLB	16,57	0,97	15,33	0,97

Tablo 1'de yer alan ortalama sınıflama doğrulukları bakımından ise tüm madde havuzlarının ve tüm madde seçme yöntemlerinin 0,94 ile 0,99 arasında olmak üzere oldukça yüksek bir sınıflama doğruluğu oranı sağladıkları görülmektedir. Hesaplanan sınıflama doğrulukları madde havuzu büyüdükçe azalmış olsa da tüm koşullarda %90'dan yüksek bir tutarlılığa işaret etmektedir. Tablo 1'de ortalama sınıflama doğruluğu bakımından tüm madde havuzlarında MFB ve KLB yöntemlerinin birbirine benzer çalıştıkları ve bu iki yöntemin TMSY'ye kıyasla daha iyi bir sınıflama doğruluğu verdikleri görülmektedir.

Tablo 1'deki sivri ve basık dağılımlı madde havuzları, ortalama test uzunluğu ve ortalama sınıflama doğruluğu bakımlarından karşılaştırıldığında, madde havuzu dağılımının sivri olması durumunda, basık madde havuzlarına kıyasla ortalama test uzunluğu değerlerinin düştüğü, bir başka deyişle test etkililiğinin arttığı; ortalama sınıflama doğrulukları oranlarının ise pek değişmediği görülmektedir.

4. TARTIŞMA ve SONUÇ

Bu çalışmada BBST'de madde havuzu dağılımının ve madde havuzu büyüklüklerinin ortalama test uzunluğu ve ortalama sınıflama doğruluğu üzerindeki etkisi incelenmiştir. Bu amaçla oluşturulan sekiz farklı madde havuzu üzerinde üç farklı madde seçme yöntemi incelenmiştir. Araştırma sonucuna göre sivri veya basık dağılıma sahip madde havuzlarında MFB ve KLB madde seçme yöntemleri ortalama test uzunluğu bakımından benzer performans göstermiş ve bu iki yöntemin TMSY'ye kıyasla test etkililiğini artıran yöntemler oldukları görülmüştür. Kestirilen yetenek düzeyinde maksimum bilgi sağlayan maddeyle (MFB) kestirilen yetenek düzeyinin çevresindeki bölgede maksimum bilgi sağlayan maddenin (KLB) seçimi temelde benzer mantığa, bilgiyi maksimize etmeye dayanmaktadır (Spray ve Reckase, 1994). Dolayısıyla bu iki yöntemin benzer sonuçlar vermesi bu duruma bağlı olabilir. Ayrıca bu iki yöntemin TMSY'den daha iyi performans göstermiş olması beklenen bir durumdur. TMSY'de maddeye veya bireye ait herhangi bir bilgi dikkate alınmamakta; maddeler havuzdan tesadüfi seçilmektedir. Bu nedenle yöntemin etkililiği diğer yöntemlere kıyasla daha düşük olmaktadır (Thompson, 2007). Bu bulguya dayanarak araştırmacılara ve uygulayıcılara BBST çalışmalarında veya uygulamalarında MFB veya KLB madde seçme yöntemlerinin kullanılması önerilebilir.

Araştırma sonuçlarından bir diğeri madde havuzu büyüklüğü arttıkça tüm madde seçme yöntemlerinin daha az sayıda maddeyle sınıflama yapabildiğidir. Madde havuzu büyüdükçe tüm madde seçme yöntemlerinden hesaplanan ortalama sınıflama doğruluklarının azaldığı ancak yine de yüksek bir oran sağladıkları araştırmanın bir diğer sonucudur. Bu sonuca dayanarak araştırmacılara ve uygulayıcılara test etkililiğini artırabilmek amacıyla çok sayıda madde içeren büyük madde havuzları oluşturup bu havuzlarda çalışmaları önerilebilir. Basık ve sivri dağılımlı madde havuzlarına ait sonuçlar karşılaştırıldığında, madde havuzu dağılımının sivri olması durumunda, basık madde havuzlarına kıyasla ortalama test uzunluğu değerlerinin düştüğü, bir başka deyişle test etkililiğinin arttığı; ortalama sınıflama doğrulukları oranlarının ise pek değişmediği görülmektedir. Thompson'ın (2009) çalışmasıyla da örtüşen bu bulguya dayanarak, araştırmacılara veya uygulayıcılara BBST uygulama veya çalışmalarında bireyi sınıflamak için gerekli madde sayısını azaltmak ve test etkililiğini artırmak amacıyla sivri dağılıma sahip madde havuzlarının kullanılması önerilebilir.

5. KAYNAKLAR

Babcock, B. & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [15.1.2015] from www.psych.umn.edu/psylabs/CATCentral/

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

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologist. London: Lawrence Erlbaum Associates Publishers.

Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.) *Computerized adaptive testing: A Primer*. Mahwah, NJ: Erlbaum.

Hambleton, R. K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer Nijhoff Publishing.

R Core Team. (2013). R: A language and environment for statistical computing (Version 3.0.1) Vienna, Austria: R Foundation for Statistical Computing.

Spray, J. A. & Reckase, M. D. (1994). The Selection of Test Items for Decision Making with a Computer Adaptive Test. Paper presented at the Annual Meeting of the National Council on Measurement in Education. NewOrleans, LA, April 5-7, 1994.

Şencan, H. (2005). *Sosyal ve Davranışsal Ölçümlerde Güvenirlilik ve Geçerlilik*. Ankara: Seçkin Yayıncılık.

Thompson, N. A. (2007). A Practitioner's Guide for Variable-length Computerized Classification Testing. Practical Assessment Research & Evaluation, 12(1). Available online: http://pareonline.net/getvn.asp?v=12&n=1

Thompson, N. A. (2009). Item selection in computerized classification testing. Educational and Psychological Measurement, 69(5), pp. 778-793.

Thompson, Nathan A., & Weiss, David A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, 16(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1.

Wang, T. & Vispoel, W. P. (1998). Properties of Ability Estimation Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 35 (2), pp. 109-135.*

Wang, T. (2011). Essentially unbiased EAP estimates in computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association Conference, Chicago, USA.

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

Yang, X, Poggio, J. C. & Glasnapp, D. R. (2006). Effects of Estimation Bias on Multiple-Category Classification with an IRT-Based Adaptive Classification Procedure. *Educational and Psychological Measurement*, 66 (4), pp. 545-564.

Extended Abstract

Through the advantages of Item Response Theory (IRT) such as invariance of item parameteres and person parameters Computerized Adaptive Testing (CAT) is getting more attention in last years. When the CAT applications are aimed to classification the students into two or several gruops according to one or more cut-point Computerized Adaptive Classification Testing (CACT), which is a sub field of CAT becomes a current issue. CACT has six components: *(i)* Response model; *(ii)* Item pool; *(iii)* Starting rule; *(iv)* Item selection methods; *(v)* Ability estimation method and *(vi)* Classification rule. First of all, for CACT simulation or application one must determine appropriate IRT model as 1 Parameter Logistic Model (1 PLM), 2 PLM or 3 PLM for binary scored items. In this study 3 PLM was taken into account. One of the important component of the CACT is the item pool characteristic that based on item distribition and size. In this study it was investigated that the effects of item distribitions and sizes on average test length and average classification accuracy in CACT. For representing the distribution types boared and peaked item pools and for the item pool sizes 50 items, 100 items, 200 items and 300 items were compared. In this research for all CACT conditions the starting rule was set to $\theta = 0$ like many other studies. In the CAT and CACT literature there were lots of item selection methods such as Maximum Fisher Information (MFI), Kullback-Leibler Information (KLI) and Random Item Selection Method (RISM). In this study MFI, KLI and RISM were investigated and compared to each-others with respect to the item pools which have different distributions and sizes. Ability estimation methods like Expected a Posteriori (EAP) or Maximum Likelihood Estimation (MLE) aim to predict the appropriate ability level with low estimation error for individuals. In many researches EAP has much better performed than the other estimaiton methods with regard to total simulation time, errors and full true-false responses. That's why EAP was selected in this study for all CACT conditions. Finally, classification rule is a hypothesis for the individuals to be assigned to the ability groups as succesfull and not succesfull. In this research Generalized Likelihood Ratio (GLR) was taken into account to rule the classifications.

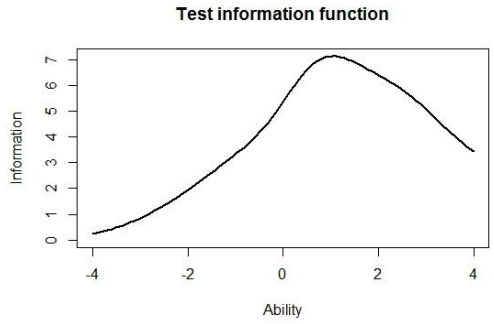
In this study it was investigated that the effects of item distributions and sizes on average test length and average classification accuracy in CACT. For that purpose RISM, MFI and KLI were

studied in boared and peaked item pools with 50 items, 100 items, 200 items and 300 items. This study gains importance because it is the first study about CACT in Turkey which runs to the computerized testing nowadays and it has 24 conditions which were never studied in the literature. Therefore it is expected to provide some information about item pool characteristic and item selection methods to the researchers or practitioners.

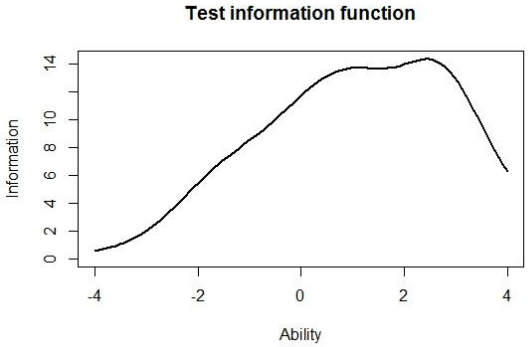
In the simulation study, individuals' thetas were derived from normal distribution as N(0,1) between (-3,+3) ability levels. In the peaked item pools items were simulated from uniform distribution as U[0,5; 2,0] for a parameters; normal distribution as N(1, 0,4) for b parameters and as N(0,15, 0,05) for c parameters; and in the broad item pools items were simulated from uniform distribution as U[0,5; 2,0] for a parameters; normal distribution as N(1, 1,5) for b parameters and as N(0,15, 0,05) for c parameters (See the appendix 1 and 2 for the graphics of item pools). The simulation study was performed in R with the codes for the *for cycle* (for each conditions 25 cycle) written by the researcher. The results were obtained from the each 25 cycles as average test length and average classification accuracy which are the dependent variables of the study.

Results show that RISM has the maximum value with respect to average test length; and MFI and KLI perform similar. The more items in the pool, the shorter test length and fewer the classification accuracy but in all conditions classification accuracy has high rate above 90%. In addition, in peaked item pools it is seen that the average test lengths are getting shorter and the test effectiveness is getting higher; but the classification accuracies are not changing. In conclusion it can be said that with the peaked item pools with more items, CACT provides shorter tests and high classification accuracy.

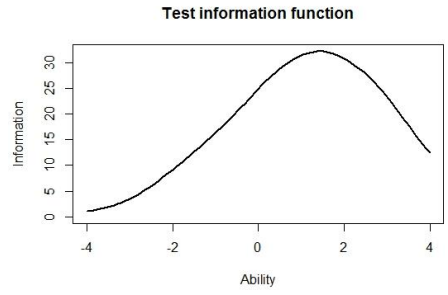
EK 1: Basık Madde Havuzlarına Ait Test Bilgi Fonksiyonları



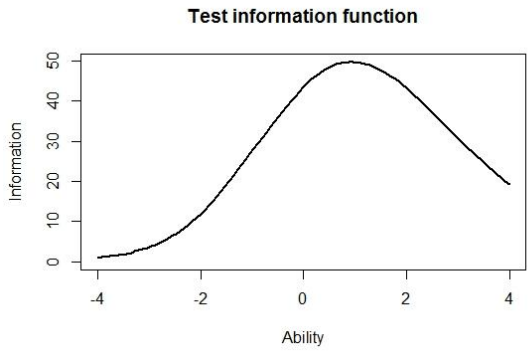
Basık Madde Havuzu 50 Madde



Basık Madde Havuzu 100 Madde

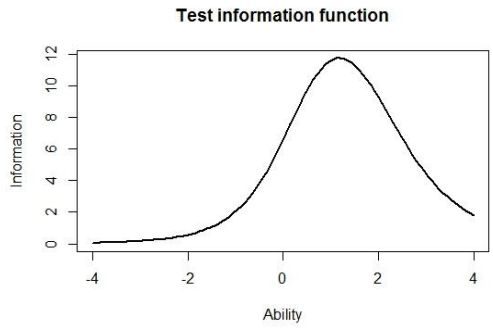


Basık Madde Havuzu 200 Madde

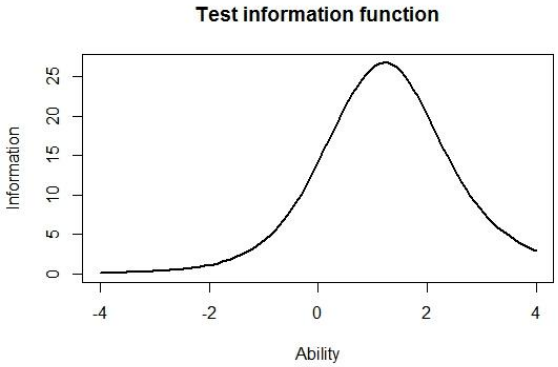


Basık Madde Havuzu 300 Madde

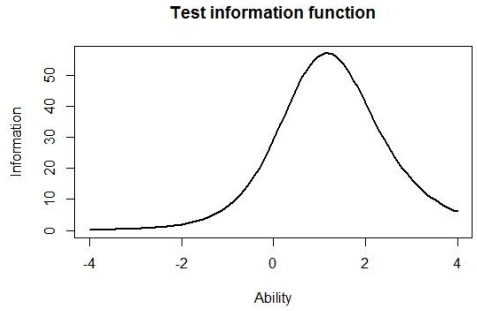
EK 2: Sivri Madde Havuzlarına Ait Test Bilgi Fonksiyonları



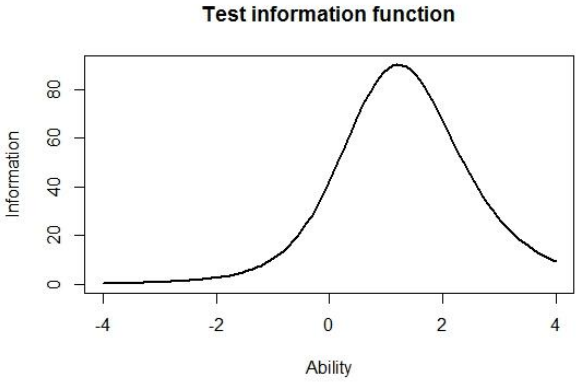
Sivri Madde Havuzu 50 Madde



Sivri Madde Havuzu 100 Madde



Sivri Madde Havuzu 200 Madde



Sivri Madde Havuzu 300 Madde