ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);34-46

Basit ve Karmaşık Test Desenlerinde Çok Boyutlu Madde Seçme Yöntemlerinin Karşılaştırılması*

A Comparison of Multidimensional Item Selection Methods in Simple and Complex Test Designs

Eren Halil ÖZBERK ** Selahattin GELBAL ***

Öz
Bu araştırmada diğer araştırmaların aksine toplam yetenek puanları gerçek test koşullarına uygun olacak şekilde farklı test koşullarında karşılaştırılmıştır (basit ve karmaşık). Araştırmada test deseni, boyut başına düşen soru sayısı, boyutlar arası korelasyon ve madde seçme yöntemleri olmak üzere dört koşul manipüle edilmiştir. Veri setleri, üretilen madde ve yetenek parametreleri ve M3PL telafi edici çok boyutlu madde tepki kuramı modeli kullanılarak belirlenen korelasyonlara bağlı kalarak üretilmiştir. Çok boyutlu bireyselleştirilmiş bilgisayarlı test uygulamaları sonucu elde edilen toplam yetenek puanları mutlak yanlılık (ABSBIAS), korelasyon ve hata kareleri ortalamasının karekökü (RMSE) kullanılarak karşılaştırılmıştır. Sonuçlar incelendiğinde çok boyutlu test deseni, boyut başına düşen madde sayısı ve boyutlar arası korelasyon değişkenlerinin toplam puanları kestirmede madde seçme yöntemleri üzerinde etkilerinin olduğu belirlenmiştir. Basit yapıdaki bir test için Minimum Hata Varyansı madde seçme yönteminin hem uzun hem de kısa testler için en düşük mutlak yanlılık değerinin ürettiği belirlenmiştir. Model karmaşıklaştıkça Kullback-Leibler madde seçme yönteminin diğer iki yöntemden daha iyi performans gösterdiği belirlenmiştir.

Anahtar Kelimeler: Madde seçme yöntemi, çok boyutlu bireyselleştirilmiş bilgisayarlı test, çok boyutlu maddde tepki kuramı, toplam puan kestirimi

Abstract
In contrast with the previous studies, this study employed various test designs (simple and complex) which allow the evaluation of the overall ability score estimations across multiple real test conditions. In this study, four factors were manipulated, namely the test design, number of items per dimension, correlation between dimensions and item selection methods. Using the generated item and ability parameters, dichotomous item responses were generated in by using M3PL compensatory multidimensional IRT model with specified correlations. MCAT composite ability score accuracy was evaluated using absolute bias (ABSBIAS), correlation and the root mean square error (RMSE) between true and estimated ability scores. The results suggest that the multidimensional test structure, number of item per dimension and correlation between dimensions had significant effect on item selection methods for the overall score estimations. For simple structure test design it was found that V1 item selection has the lowest absolute bias estimations for both long and short tests while estimating overall scores. As the model gets complex KL item selection method performed better than other two item selection method.

Keywords: Item selection method, multidimensional computer adaptive testing, multidimensional item response theory, composite score estimation

GİRİŞ

Eğitim ve psikoloji alanında değerlendirme araçlarının başlıca amacı ölçülen özeliğin miktarını belirlemek ve elde edilen numerik puanları kullanarak bireyleri örtük özelliklerine göre sıralamaktır. Puanlar, sıralama amacıyla kullanıldığı durumlarda önemli bir değerlendirme ölçütü olarak olabilmektedir. Özellikle başarı düzeylerinin belirlenmesinde, sertifika ve lisanslama yetkilerinin

* Bu çalışma, birinci yazarın Prof. Dr. Selahattin GELBAL danışmanlığında tamamlanan doktora tezinden türetilmiştir.
** Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, eposta: erenozberk@gmail.com
*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, eposta: sgelbal@gmail.com

DOI: 10.21031/epod.286956
Geliş Tarihi: 22.01.2017
Kabul Tarihi: 06.03.2017

verilmesinde puanlar önemli bir sıralama ölçütüdür. Örneğin bir öğrencinin üniversiteye giriş sınavından elde ettiği puan öğrencinin bilgi ve becerisinin bir göstergesidir.

Toplam yetenek puanlarının rapor edildiği durumlarda, tek boyutlu test desenleri yerine kullanılan çok boyutlu test desenlerinin parametre kestirimlerinde önemli rol oynamaktadır. Çok boyutlu madde tepki kuramı (ÇBMTK) modelleri test yapıları bakımından basit ve karmaşık yapı olarak ikiye ayrılmıştır. Literatürde bunun için birçok farklı adlandırma mevcuttur. Bazı araştırmacılar basit ve karmaşık yapılı test deseni olarak adlandırırken (Luo, 2013; Yao, 2012; Zhang, 2012) bazıları maddeler arası model (multidimensional between-item model) ve maddeler içi model (multidimensional within-item model) olarak adlandırmıştır (Adams, Wilson ve Wang, 1997; Bulut, 2013; Wang, Chen ve Cheng, 2004). Maddelerin faktöriyel karmaşık bir yapıda olması durumunda belirli maddeler birden fazla boyuta yük verebilirler ve yine birden fazla boyuta bilgi sağlayabilirler. Testlerde yer alan maddeler çok boyutlu özelliğe sahip olduğunda çok boyutlu ölçme bilgisinin ortaya çıkması, alt boyut ve toplam yetenek puanlarının kestirimine etki edebilmektedir (Bulut, 2013; Liu, 2015; Luecht, Gierl, Tan ve Huff, 2006; Luo, 2013; Zhang, 2012; Finch, 2010). Yapılan araştırmalarda çok boyutlu test yapılarının çok boyutlu bilgi düzeylerini farklılaştırdığı, bu bakımdan elde edilen yetenek ve madde parametrelerini etkilediği görülmüştür. ÇBMTK modellerinin verdiği avantajlar göz önüne alındığında bazı araştırmacılar gerçek test koşullarına yakın olması nedeniyle karmaşık çok boyutlu yapıların kullanılmasını önerirken (Ackerman, 1994; Reckase, 2009) diğer araştırmacılar basit yapıdaki desenler kullanmanın karmaşık yapıda desenler kullanmadan daha avantajlı olduğunu belirtmişlerdir (Luecht & Miller, 1992; Yao & Boughton, 2009).

Çok boyutlu bireyselleştirilmiş bilgisayarlı testler (ÇBBBT), hem alt boyut hem de toplam puanlar rapor etmede tek boyutlu bireyselleştirilmiş bilgisayarlı test (TBBBT) uygulamalarına göre avantajlı durumdadır. ÇBBBT uygulamalarında her bir alt yeteneğe göre puanlar daha az madde ile kestirilebilmektedir ve her boyuta ait puanlar rapor edilebildiğinden bireylerin zayıf ve güçlü yanları boyutlara göre rahatlıkla belirlenebilmektedir (Wang ve Chang, 2011). Yakın zamanda yapılan araştırmalarda TBBBT uygulamaları için geliştirilen birçok madde seçme yöntemi ÇBBBT uygulamalarına göre tekrar geliştirilmiştir (Mulder ve van der Linden, 2010; Segall, 1996; van der Linden, 1999; Veldkamp ve van der Linden, 2002). TBBBT ve ÇBBBT madde seçme yöntemleri kestirdikleri yetenek sayılarının farklı olması bakımından ayrılmaktadır. Özellikle ÇBBBT madde seçme yöntemleri, çoklu yeteneklerin kestirimlerine getirdikleri farklı teknikler ile alt boyut puanlarının hesaplanmasında daha kararlı sonuçlar elde etmeye çalışmışlardır. Örneğin, Segall (1996) tarafından önerilen madde seçme yöntemi, genel varyansı azaltarak yeteneklere ait güven aralıklarını düşürmek isterken, van der Linden (1999) tarafından önerilen yöntem ise her bir yetenek kestirimine ait toplam varyansı azaltma yoluna gitmiştir.

Çok Boyutlu Bireyselleştirilmiş Bilgisayarlı Testler

Bireyselleştirilmiş testlerde çok boyutlulukla ilgili ilk çalışmalar Bloxom ve Vale (1987), Fan ve Shu (1996), Luecht (1996), Segall (1996) ile başlamış daha sonraları van der Linden (1999; 2005), Mudler ve van der Linden (2009) ile devam etmiştir. Yapılan ilk çalışmalar yetenek kestirimleri ve madde seçme yöntemleri üzerine yoğunlaşmıştır. Daha sonraları Wang ve Chen (2004), çok boyutlu test yapılarının yetenek kestirimleri üzerindeki etkilerini incelemişledir. Yakın zamanda yapılan çalışmalar, madde seçme yöntemlerinin yetenek kestirimlerine etkisi ile (Wang ve Chang, 2011; Yao, 2012, 2014), ÇBBBT uygulamalarında farklı durdurma kurallarının etkisi üzerinde yoğunlaşmaktadır (Wang, Chang ve Boughton, 2011; Yao, Pommerich, Segall, 2014).

Hem TBBBT hem de ÇBBBT uygulamalarında alt boyut ve toplam puanlarını kestirmede yetenek kestirim yöntemleri doğrudan sonuçları etkilese de, uygun maddelerin seçilmemesi durumunda hiçbir yetenek kestirim yöntemi fonksiyonel olmayacaktır (Reckase, 2009). Maddelerin çok zor, çok kolay ve düşük bilgi verici maddeler arasından seçilmesi yetenek kestirimlerini etkilemektedir. Maddelerin havuzdan seçme işlemindeki kurallar, ÇBBBT uygulamalarında önem kazanmaktadır. Literatürdeki madde seçme yöntemleri θ kestirimde kullanılacak bazı kritik değerleri maksimize ya da minimize

etme ilkesine dayanmaktadır. Madde seçme yöntemleri de kritik değerleri tanımlama konusunda birbirinden farklılaşmaktadır.

ÇBBBT Madde Seçme Yöntemleri

Fisher Bilgi Matrisinin Determinantının Artırtılması-Hacim Yöntemi (Vol)

Segall (1996) önceki araştırmalarda (Bloxom & Vale, 1987; Tam, 1992) çok boyutlu madde tepki kuramı çerçevesinde yeteneğin ortak dağılımına ait önsel bilgilerin kullanılmadığı için sonuçların geçerli sayılamayacağını belirtmiştir. Segall'e göre ÇBBBT her bir alt boyuttan belirli sayıda madde seçme yerine her bir alt boyutun özelliğini etkili şekilde ortaya çıkaracak madde seçme prosedürleri sağlayabilmektedir. ÇBBBT uygulamalarında ayrıca boyutlar arasındaki ilişkiler dikkate alındığından madde seçme prosedürlerinin etkililiği daha da artırılabilmektedir. Segall (1996) Bayes modellemesine dayalı, yeteneğin ortak dağılımına ait önsel bilgileri de dikkate alan bir madde seçme yöntemi önermiştir. Yöntem, bireylerin alt boyut yeteneklerini ÇBMTK modelleri yardımıyla seçilen maddelerden $(k-1)$ kestirmektedir. Bu maddelerden elde edilen bilgi $(I_{k-1}(\theta^{k-1})$ önsel dağılım olarak kullanılmakta ve bir sonraki maddenin (k) seçiminde kullanılmaktadır. Bu sayede yetenek kestirimlerinin $\left(\theta^{k-1}\right)$ doğruluğunun artırıldığı belirtilmektedir. Sonsal bilgi dağılımının determinantını maksimize eden eşitlik denklem 1'de gösterilmiştir.

$$W = \left|I_{k-1}\left(\theta^{k-1}\right) + I_k\left(\theta^{k-1}\right) + \sum^{-1}\right| \quad (1)$$

Madde havuzundaki her i maddesi için, hacim ya da bilgi fonksiyonunun determinantı denklem 2'deki eşitlik ile hesaplanabilmektedir (Yao, 2012).

$$W_m = \left|I_{k-1}\left(\theta^{k-1}\right) + \frac{(P_{i1}-\beta_{3i})^2(1-P_{i1})}{P_{i1}(1-\beta_{ik})^2}\beta_{2i}x\hat{\beta}_{2i} + \sum^{-1}\right| \quad (2)$$

Kullback-Leibler

Kullback-Leibler (KL) bilgisinin TBBBT uygulamalarında ilk olarak Chang ve Ying (1996) tarafından kullanılmıştır. Veldkamp ve van der Linden (2002) KL bilgisini gölge test yöntemi (shadow test method) kullanarak çok boyutlu yapıya uyarlamışlardır. KL madde seçme yöntemi gerçek yetenek (θ_0) ile kestirilen yetenek (θ) arasındaki iki olabilirlik arasındaki uzaklığı ölçmektedir. M3PL model için KL bilgisi denklem 3 ile gösterilmiştir.

$$K_i(\theta, \theta_0) = P_i(\theta_0)Ln\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + [1 - P_i(\theta_0)]Ln\left[\frac{1-P_i(\theta_0)}{1-P_i(\theta)}\right] \quad (3)$$

Denklem 3'te $i = (1,2,\ldots,N)$ madde havuzundaki N sayıdaki maddeyi belirtmektedir. θ_o değeri, θ_o bilinmediğinde ve θ tanımlanmadığı durumda, sonsal beklenen KL bilgisine göre kestirilmektedir. θ sonsal dağılımının yoğunluğu $f(\theta\,|u_{i_1},\ldots,u_{i_{k-1}})$ ile tanımlanmaktadır ve uygulanan $k-1$ sayıdaki maddenin bir fonksiyonudur. Sonsal beklenen KL bilgisi kullanılarak $\hat{\theta}^{k-1}$ kestirimi denklem 4 kullanılarak hesaplanmaktadır.

$$K_i^B(\hat{\theta}^{k-1}) \equiv \int_\theta K_i\left(\theta, \hat{\theta}^{k-1}\right) f(\theta\,|u_{i_1},\ldots,u_{i_{k-1}})\partial\theta \quad (4)$$

Çok boyutluluk açısından bakıldığında KL denklemindeki θ ve θ_0 değerleri ÇBMTK modelinde skaler yerine vektörel olarak ifade edilmektedir.

ÇBBBT uygulamalarında KL madde seçme yönteminin kullanılmasının iki temel nedeni bulunmaktadır. İlk olarak, tek boyutlu KL madde seçme yönteminde gerçek θ değerlerinin kestiriminde Fisher bilgisinden daha başarılıdır. Ayrıca KL bilgisi önsel dağılımları kullandığından gerçek ve kestirilen yetenekleri geçerli bir şekilde ayırt etmektedir. KL yöntemi, Fisher yönteminin aksine θ ve θ_0 değerlerinin birer fonksiyonu olarak ifade edilebilir ve yetenek seviyelerinin birbirine yakın olmasını gerektirmez (Chang ve Ying, 1996; Veldkamp ve van der Linden, 2002).

Minimum Hata Varyansı Kriteri

Hata varyanslarının doğrusal birleşimlerinin minimize eden (V1) madde seçme yöntemi, eşit ağırlıklandırılmış boyutlardan elde edilen toplam puanlara en düşük hata varyansı veren maddeyi seçmektedir. Bu yöntem toplam puanların doğruluğunu artırmak için van der Linden (1999) tarafından ortaya atılmıştır.

van der Linden (1999) çok değişkenli bilgi matrisinin yerine asimptotik varyans-kovaryans matrisinin kullanılmasındaki amacın madde seçme yöntemini çok boyutlu MTK'ya göre uyarlamak olduğunu belirtmiştir. Bireyselleştirilmiş test uygulamalarında kestirilen yetenek, her bir madde seçiminden sonra elde edilen yeteneklerin $(\lambda(\theta_1, \dots, \theta_m);\ \lambda = \lambda_j = (\lambda_1, \lambda_2, \dots, \lambda_m);\ \lambda_j \geq 0)$ doğrusal kombinasyonlarına eşittir. Ağırlıklandırmanın (λ) değeri testin amacına göre değişmekte ve bu değer, BBT prosedürlerini ve yetenek kestirimlerini değiştirebilmektedir.

İlk olarak işlem MAP yöntemi kullanarak yetenek kestirimi ile başlamaktadır. Yerel bağımsızlık varsayımından dolayı $\widehat{\lambda'\theta} = \lambda'\theta$ olarak ifade edilebilir ve θ olabilirlik fonksiyonu denklem 5 yardımı ile hesaplanabilmektedir.

$$g\left(\theta \middle| u_{i_1}, u_{i_2}, \dots, u_{i_{k-1}}\right) = \frac{L\left(\theta \middle| u_{i_1}, u_{i_2}, \dots, u_{i_{k-1}}\right) g(\theta)}{\int L\left(\theta \middle| u_{i_1}, u_{i_2}, \dots, u_{i_{k-1}}\right) g(\theta) d\theta} \qquad (5)$$

Bireyselleştirilmiş test algoritmasında herhangi iki yetenek değişkenine ait madde seçme prosedürü için senaryo belirtilen şekildedir: $k-1$ madde seçilmiş olsun. $S_k = (i_1, i_2, \dots, i_{k-1})$ seçilen maddeleri; $R_k = (1,2, \dots, i)/S_k$ ise reddedilen maddeleri göstermektedir. Bireyselleştirilmiş test algoritması kullanılarak $k-1$ sayıda madde uygulandıktan sonra k maddesi denklem 6'daki kritere göre seçilmektedir.

$$min_{R_k}\left[Var(\lambda\hat{\theta}_1^k + (1-\lambda)\hat{\theta}_2^k | \hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}\right] \qquad (6)$$

Denklem 6'da $\hat{\theta}_1^k$ ve $\hat{\theta}_2^k$, θ_1 ve θ_2 değerlerinin kestirimlerini göstermektedir. Denklem incelendiğinde madde $\lambda\hat{\theta}_1^k + (1-\lambda)\hat{\theta}_2^k$ varyansını minimize edecek şekilde seçilmektedir. R_k kümesindeki maddeleri seçmek için, final denklemi madde parametrelerini $(a_{i1}, a_{i2}\ ve\ d_i)$ ve olasılık $[P_i(\theta_1, \theta_2)]$ değerlerini içermektedir. Özetlenecek olursa, algoritma ilk olarak testin amacını yansıttığı düşünülen ağırlıklandırılmış değerlerini (λ) seçmektedir. Ağırlıklandırılmış deneysel veya seçilen problemden elde edilmiş olabilmektedir. En son adımda ise yetenek parametreleri MAP eşitlikleri kullanılarak kestirilmektedir.

V1 yöntemini kullanmanın bir takım avantajları vardır (van der Linden, 1999; Yao, 2012). Bazı durumlarda madde havuzu birden çok yeteneği ölçecek şekilde desenlenmiştir. Bu yetenekler arasında ilişkiler olabileceği gibi herhangi bir ilişkiye rastlanmıyor olabilir. Bu durumda bireysel yetenekler ağırlıklandırma seçimini etkileyebilmektedir. Bu bakımdan V1 yöntemi alt boyut puanlarının doğrusal kombinasyonlarını düzenleyerek toplam puanların daha doğru kestirilmesini sağlamaktadır

Literatürde toplam puanları rapor etmede çok boyutlu madde seçme yöntemlerinin karşılaştırılmasına ilişkin çalışmalar olsa da (Wang ve Chang, 2011; Yao, 2012, 2013), çok boyutlu test deseninin yapısına göre nasıl performans gösterdiğine ilişkin bir çalışmaya rastlanmamıştır.

Araştırmanın Amacı

Çok boyutlu bireyselleştirilmiş bilgisayarlı test (ÇBBBT) uygulamalarının tek boyutlu bireyselleştirilmiş bilgisayarlı test (TBBBT) uygulamalarına göre birtakım üstünlükleri bulunmaktadır. ÇBMTK test yapılarında boyutlar arasında ilişkiler mevcuttur ve bir madde birden çok boyuta yük verebilir. Bu sayede her bir maddenin bilgi vericiliği ÇBBBT uygulamalarında daha dikkatle ele alınır ve TBBBT'ye göre daha kararlı sonuçlar elde edilir. Maddelerden elde edilen bilgiler arttığından test uzunluğu da düşmektedir (Segall, 1996). TBBBT uygulamalarında kullanılan kapsam dengeleme kısıtlamaları sonucu bazı alt boyutlar bireyin genel yeteneğine daha az katkı sağlamaktadır. ÇBBBT uygulamaları, kapsam alanlarını korelasyon değerlerini de göz önüne alarak ayrı ayrı ele alır

ve farklı kapsamlardan elde edilen bilgiyi bütün boyutlarla beraber işleme koyar (Segall, 1996; Wang ve Chang, 2011).

İncelenen araştırmalar madde sayısının ve maddeler arası korelasyon değerlerinin alt boyut ve toplam puanları hesaplamada değişebildiğini göstermektedir (Segall, 1996; Wang ve Chang, 2011; Wang, Chang ve Boughton, 2013; Yao, 2012) . Bu sebeple boyutlar arası korelasyon ve boyut başına düşen madde sayısının madde seçme yöntemleri üzerindeki etkisinin nasıl değişeceğinin belirlenmesinin uygulayıcılara önemli bilgiler sağlayacağı düşünülmektedir.

ÇBBBT uygulamalarında toplam puanları hesaplamada madde seçme yöntemlerinin farklı kestirim değerleri sunduğu belirlenmiştir (Yao, 2012, 2013). Bu sebepten madde seçme yöntemlerinin karşılaştırılması ve toplam puanları kestirmede farklı koşullar için en az hata veren yöntemlerin belirlenmesi gerekmektedir. Geniş ölçekli sınavlarda puanların rapor edilmesinin her geçen gün arttığı, sınavı alan bireylere bu puanlar doğrultusunda geri bildirimler verildiği dikkate alındığında ÇBBBT uygulamalarında toplam puanları rapor etmede en az hata içeren koşulların belirlenmesinin önemli olduğu düşünülmektedir. Bu araştırmada, PISA 2012 Türkiye örnekleminden elde edilen veriler kullanılarak çok boyutlu test yapılarında toplam puanları belirlemede madde seçme yöntemlerinin farklı koşullar altındaki performanslarını ortaya çıkarmak amaçlanmıştır.

Problem Cümlesi

Basit, düşük ve yüksek karmaşık yapıdaki testlerde, boyutlar arası korelasyonun ve boyut başına düşen madde sayısının madde seçme yöntemlerinin hata, mutlak yanlılık ve korelasyon değerlerine etkisi nasıldır?

Alt Problemler

1. Basit Yapılı (BY) test deseninde çok boyutlu madde seçme yöntemlerinin hatası, mutlak yanlılığı ve korelasyon değerleri boyutlar arası korelasyona ve boyut başına düşen madde sayısına göre toplam yetenek puanları için nasıl değişmektedir?
2. Düşük Karmaşık Yapılı (DKY) test deseninde çok boyutlu madde seçme yöntemlerinin hatası, mutlak yanlılığı ve korelasyon değerleri boyutlar arası korelasyona ve boyut başına düşen madde sayısına göre toplam yetenek puanları için nasıl değişmektedir?
3. Yüksek Karmaşık Yapılı (YKY) test deseninde çok boyutlu madde seçme yöntemlerinin hatası, mutlak yanlılığı ve korelasyon değerleri boyutlar arası korelasyona ve boyut başına düşen madde sayısına göre toplam yetenek puanları için nasıl değişmektedir?

YÖNTEM

Araştırmada var olan yöntem ve tekniklerin gerçek veri üzerinden performanslarının karşılaştırılması amaçlandığından araştırma nicel karşılaştırma araştırmasıdır.

Araştırmanın Deseni

Araştırmada test deseni, boyutlar arası korelasyon değeri, boyut başına düşen madde sayısı ve madde seçme yöntemleri olmak üzere dört farklı durum manipüle edilmiştir. Manipülasyonların sonucunda 3x3x2x3 olmak üzere toplam 54 deneysel koşul çapraz olarak test edilmiştir.

Tablo 1. Araştırma Deseni

Test Deseni	Boyutlar arası korelasyon	Boyut Başına Düşen Madde Sayısı	Madde Seçme Yöntemi
Basit Yapı (BY) Düşük Karmaşık Yapı (DKY) Yüksek Karmaşık Yapı (YKY	ρ=0.2 ρ=0.5 ρ=0.8	Kısa Test (n=10) Uzun Test (n=15)	Kullback-Leibler (KL) Hacim (Vol) Minimum Hata Varyansı (V1)

Verilerin Üretilmesi

Araştırmada kullanılan madde parametrelerinin üretilmesinde PISA 2012 Türkiye verisinden yararlanılmıştır. Eğim, güçlük ve düşük asimptot parametreleri a, b ve c, 3 parametreli lojistik model kullanılarak a ~ LN{0, 0.2), b ~ N(0, 1), ve c ~ Beta{6,16) koşullarını sağlayacak şekilde üretilmiştir. Ampirik olarak elde edilen madde parametresi değerlerine bağlı kalmak koşuluyla a ve b parametreleri [0.5, 1.5] ve [-2, 2] arasında değerler alıncaya kadar yeniden üretilmiştir. Düşük asimptot değeri olan c-parametresi 0.15 değerine sabitlenmiştir.

Test Desenlerinin Oluşturulması

Araştırmada çok boyutlu test desenleri, tek boyutlu olarak üretilen *a*-parametrelerinden yararlanılarak ve

$$a_j = MDISC = \sqrt{a_1^2 + a_2^2 + a_3^2}$$

maddenin çok boyutlu ayırt ediciliği (ÇBAE-MDISC) formülü kullanılarak belirlenmiştir. MDISC değeri çok boyutlu ayırt ediciliğin tek boyutlu halidir (Ackerman, Gierl ve Walker, 2003).

Araştırmada çok boyutlu test yapılarından basit yapı, düşük karmaşık yapı ve yüksek karmaşık yapı olmak üzere üç farklı test deseni çok boyutlu ayırt edicilik değerlerinin sabit tutulması koşuluyla *a*-parametresi değerlerinin boyutlara dağıtılmasıyla belirlenmiştir.

Yetenek Parametrelerinin Üretilmesi

BY, DKY ve YKY test desenleri için 1000 birey ve 3 alt boyuttan oluşan 1000x3 gerçek yetenek parametreleri matrisi, çok değişkenli normal dağılıma göre $\theta_i = MVN(0, \sum)$) varyans-kovariyans matrisleri kullanılarak üretilmiştir. Simülasyon sonucu elde edilen madde ve yetenek parametreleri kullanılarak, boyutlar arası korelasyon değerleri ile birlikte, telafi edici çok boyutlu MTK modeline göre cevap matrisleri MIRTGEN 3.0 (Luecht, 2004) programı kullanılarak üretilmiştir. Cevap matrisleri üretilirken çok boyutlu 3 parametreli lojistik model kullanılmıştır (M3PL). Maddelerin kalibrasyonunda MML yöntemine dayanan Bock ve Aitkin Expectation-Maximization (BAEM) algoritması (Bock ve Aitkin, 1981) kullanılmıştır. BAEM algoritması EM algoritmasından farklı olarak her bir madde parametresinin log-olabilirlik değerlerinin türevlerini sadece o madde parametrelerine bağlı olarak hesaplanmaktadır.

Verilerin Analizi

ÇBBBT analizlerinde yetenek kestirimleri ve madde seçme yöntemleri her boyuta aynı anda uygulanmıştır. Başlangıç maddesi $\theta_{başlangıç} = \{\theta_1, \theta_2, \theta_3\} = \{0,0,0\}$ koşulunu sağlayacak şekilde seçilmiştir. Test uzunlukları madde havuzlarının uzunluklarına eşit olacak şekilde her bir boyut için toplamda 30 ve 45 olarak belirlenmiştir ve sabit uzunluklu sonlandırma kuralı uygulanmıştır.

Araştırmada ÇBBBT alt boyut ve toplam yetenek puanları MAP yöntemi kullanılarak 100 iterasyonun ortalaması hesaplanarak kestirilmiştir. ÇBBBT toplam yetenek puanlarının kestirimi SimuMCAT (Yao, 2011) programı kullanılarak hesaplanmıştır.

Değerlendirme Kriteri

Her bir ÇBBBT koşulu için, toplam yetenek parametrelerinin kesinliğini belirlemede gerçek ve kestirilen yetenek puanları arasındaki korelasyon katsayısı ($r = \frac{\sigma_{\hat{\theta}\theta}}{\sigma_{\hat{\theta}}\sigma_{\theta}}$), hata kareleri ortalamasının karekökü ($RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}$) ve mutlak yanlılık ($ABSBIAS = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right|$) kullanılmıştır.

BULGULAR

Gerçek yetenek parametreleri basit, düşük karmaşık ve yüksek karmaşık test desenlerine, boyutlar arasındaki korelasyona ve her bir boyuta düşen soru sayısına göre çok boyutlu normal dağılım kullanılarak üretilmiştir. Üretilen parametrelerin kesinliğini kontrol etmek amacıyla boyutlar arası korelasyon değerleri, ortalama ve standart sapma değerleri her bir koşul için hesaplanmıştır. Tablo 2 MIRTGEN (Luecht, 2004) programı kullanılarak üretilen gerçek yetenek parametrelerinin sonuçlarını özetlemektedir.

Tablo 2. Gerçek Yetenek Puanlarına Ait Ortalama, Standart Sapma ve Boyutlar arası Korelasyon Değerleri

			Altboyut 1		Altboyut 2		Altboyut 3				
Madde Sayısı*	Test Deseni	Korelasyon**	Ort	Ss	Ort	Ss	Ort	Ss	ρ'_{12}	ρ'_{13}	ρ'_{23}
10 Madde	BY	$\rho = 0.2$	.006	.99	-.067	.99	-.059	.99	0.191	0.229	0.247
		$\rho = 0.5$	.032	.97	.047	.96	-.012	.98	0.462	0.473	0.443
		$\rho = 0.8$	-.048	1.02	-.058	1.01	-.058	1.02	0.814	0.827	0.806
	DKY	$\rho = 0.2$	.013	.97	-.034	1.01	-.057	.99	0.268	0.192	0.276
		$\rho = 0.5$	.000	.96	.028	1.02	-.014	.98	0.508	0.548	0.528
		$\rho = 0.8$	-.005	.98	-.018	.98	.005	.98	0.809	0.803	0.792
	YKY	$\rho = 0.2$	.084	1.01	.032	1.01	.051	.98	0.148	0.221	0.227
		$\rho = 0.5$	.035	1.06	.034	1.01	-.001	.99	0.543	0.539	0.501
		$\rho = 0.8$	.034	1.00	.044	1.01	.005	1.01	0.803	0.802	0.781
15 Madde	BY	$\rho = 0.2$	.013	.99	-.009	1.01	.009	.99	0.185	0.185	0.231
		$\rho = 0.5$	.006	1.01	-.026	1.04	.012	1.02	0.526	0.490	0.495
		$\rho = 0.8$	-.004	.99	-.021	1.03	.002	1.01	0.798	0.789	0.814
	DKY	$\rho = 0.2$	-.008	.98	.000	1.01	-.014	.94	0.209	0.173	0.194
		$\rho = 0.5$	-.035	.96	-.012	1.01	-.048	1.04	0.505	0.517	0.502
		$\rho = 0.8$	-.020	1.03	-.013	1.02	.009	1.03	0.798	0.813	0.816
	YKY	$\rho = 0.2$	-.024	.99	-.034	1.01	-.052	.99	0.170	0.228	0.196
		$\rho = 0.5$	-.061	.98	-.027	.99	-.001	.96	0.492	0.516	0.498
		$\rho = 0.8$	-.015	1.01	-.021	.97	-.018	1.01	0.793	0.794	0.802

**Boyut başına düşen madde sayısı, **Boyutlar arası korelasyon*

Tablo 2 incelendiğinde üretilen yetenek parametrelerine ait korelasyon değerleri varsayılan (hipotetik) korelasyon değerlerine benzer olarak elde edilmiştir. Her bir alt boyuta ait üretilen gerçek yetenek puanlarına ait ortalamalar -0.067 ile 0.084 arasında, standart sapma değerleri ise 0.94 ile 1.06 arasında

değişmektedir. Alt boyutlara ait gerçek puanlar, ortalaması 1, standart sapması 0 olan çok boyutlu normal dağılıma yakın değerlerde elde edilmiştir.

Tablo 3. Test Desenlerine Ait Mutlak Yanlılık, Hata ve Korelasyon Değerleri

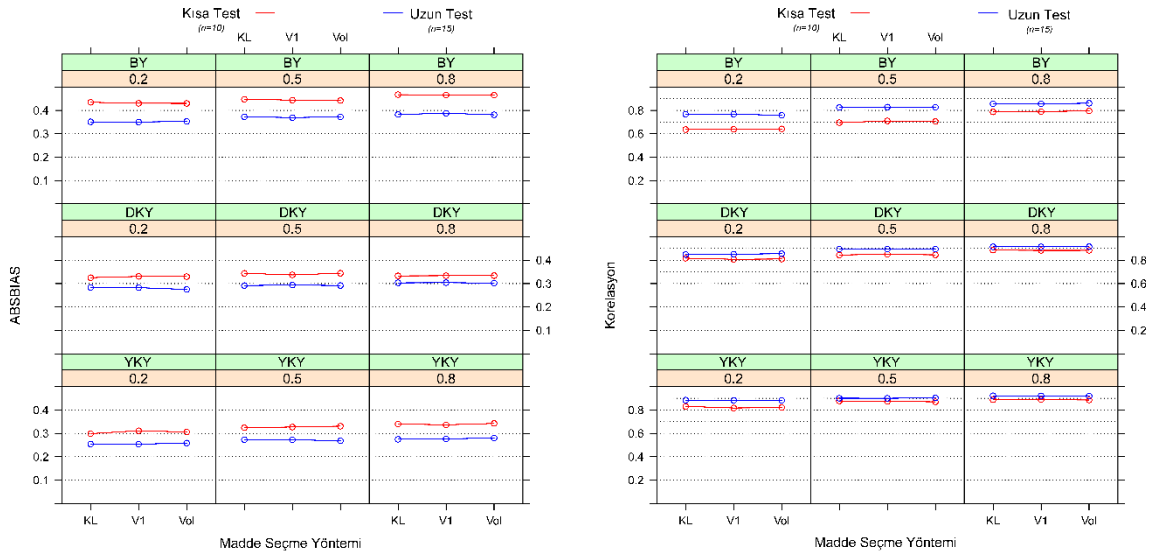
	BY			DKY			YKY		
	KL	V1	Vol	KL	V1	Vol	KL	V1	Vol
ABSBIAS									
$\rho = 0.2$, N=10	.328	.323	.327	.244	.251	.246	.232	.248	.240
$\rho = 0.5$, N=10	.358	.349	.355	.271	.268	.274	.250	.251	.259
$\rho = 0.8$, N=10	.359	.345	.354	.269	.258	.263	.263	.261	.270
$\rho = 0.2$, N=15	.307	.302	.313	.191	.190	.185	.185	.188	.189
$\rho = 0.5$, N=15	.290	.286	.289	.206	.202	.209	.204	.204	.200
$\rho = 0.8$, N=15	.302	.304	.306	.213	.213	.209	.207	.206	.210
RMSE									
$\rho = 0.2$, N=10	.014	.014	.014	.010	.010	.010	.009	.010	.010
$\rho = 0.5$, N=10	.014	.014	.014	.011	.011	.011	.010	.010	.010
$\rho = 0.8$, N=10	.015	.015	.015	.011	.011	.011	.011	.011	.011
$\rho = 0.2$, N=15	.011	.011	.011	.009	.009	.009	.008	.008	.008
$\rho = 0.5$, N=15	.012	.012	.012	.009	.009	.009	.009	.009	.008
$\rho = 0.8$, N=15	.012	.012	.012	.010	.010	.010	.009	.009	.009
Korelasyon									
$\rho = 0.2$, N=10	.64	.64	.64	.82	.81	.81	.83	.83	.82
$\rho = 0.5$, N=10	.70	.71	.70	.84	.85	.85	.88	.87	.87
$\rho = 0.8$, N=10	.79	.79	.79	.89	.88	.88	.89	.89	.88
$\rho = 0.2$, N=15	.77	.77	.76	.85	.85	.86	.88	.88	.88
$\rho = 0.5$, N=15	.82	.83	.83	.90	.90	.90	.90	.90	.90
$\rho = 0.8$, N=15	.86	.86	.86	.91	.91	.92	.92	.92	.92

Tablo 3 incelendiğinde basit yapılı bir testte testin katı çok boyutluluk (ρ=0.2) özeliği gösterdiği ve boyut başına düşen madde sayısının n=10 olduğu durumda toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.328, 0.323 ve 0.327 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri de 0.014; korelasyon değerleri ise 0.64 olarak hesaplanmıştır. Boyut başına düşen madde sayısının 10 olduğu kısa testlerde, toplam yetenek puanlarında çok boyutluluk değerlerine göre ABSBIAS ve RMSE değerlerinde kısmi artışlar gözlenmiştir. Boyut başına düşen madde sayısının 10 olduğu ve çok boyutluluğun etkisinin azaldığı durumda, yani boyutlar arası korelasyonun 0.2'den 0.8'e çıktığı durumda, korelasyon ile birlikte RMSE ve ABSBIAS değerleri de artmıştır. Testin çok boyutluluğunun azaldığı (ρ=0.8) ve boyut başına düşen madde sayısının n=10 olduğu durumda toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.359, 0.345 ve 0.354 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri 0.015; korelasyon değerleri ise 0.79 olarak hesaplanmıştır.

Tablo 3 incelendiğinde test yapısı karmaşıklaşmaya başladıkça mutlak yanlılık ve hata değerlerinde azalma olduğu görülmüştür. Test yüksek karmaşık yapıda olduğu durumda en düşük mutlak yanılık değerleri Tablo 3'te belirtilmiştir. Boyut başına düşen madde sayısının 10 olduğu durumda, düşük yapılı testteki mutlak yanlılık ve hata değerleri basit yapıdakiler ile benzerlik göstermiş, boyutlar arası korelasyon değerinin 0,2'den 0,5'e yükselmesi ile artmış; 0,5'ten 0,8'e çıkmasıyla azalmıştır. Yüksek karmaşık yapılı testlerde ise boyutlar arası korelasyon değeri arttıkça mutlak yanlılık ve hata değerleri sürekli olarak artmıştır. Boyut başına düşen madde sayısının 10 olduğu durumda en düşük mutlak yanlılık ve hata değerleri yüksek karmaşık yapıdaki test deseninde görülmüştür. Ayrıca gerçek puanlar ile kestirilmiş puanlar arasındaki korelasyon değerlerinin en yüksek değer aldığı test deseni de yüksek karmaşık yapıdaki test desenidir.

Testin uzunluğu artırıldığında (n=15) beklenildiği gibi yetenek puanları daha az hata ve mutlak yanlılık ile kestirilmiştir. Testin katı çok boyutluluk (ρ=0.2) özeliği gösterdiği ve boyut başına düşen madde sayısının n=15 olduğu durumda toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.307, 0.302 ve 0.313 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri 0.011; korelasyon değerleri ise sırasıyla 0.77, 0.77 ve 0.76 olarak hesaplanmıştır. Bu bulguların sonucuna bakıldığında tüm test desenleri için testin katı çok boyutluluk (ρ=0.2) özelliği gösterdiği durumda, boyut başına düşen madde sayısının artırılması, toplam yetenek puanlarının kestirimini olumlu olarak etkileyecektir denilebilir.

Basit yapıdaki bir testte kısa testlerin aksine uzun testlerde boyutlar arası korelasyon değeri arttığında RMSE ve ABSBIAS değerlerinde bir azalma görülmüştür. Testin çok boyutluluğunun azaldığı (ρ=0.8) ve boyut başına düşen madde sayısının n=15 olduğu durumda ise toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.302, 0.304 ve 0.306 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri de 0.012; korelasyon değerleri ise 0.86 olarak hesaplanmıştır. Test yapısı karmaşıklaşmaya başladıkça mutlak yanlılık ve hata değerleri boyutlar arası korelasyonun artışı ile doğru orantılı olarak artmıştır. Toplam test puanlarına ait en düşük mutlak yanlılık ve hata değerleri testin yüksek karmaşık yapıda olduğu durumda görülmüştür. Ayrıca gerçek puanlar ile kestirilmiş puanlar arasındaki korelasyon değerlerinin en yüksek değer aldığı test deseni de yüksek karmaşık yapıdaki test desenidir (r=0.92).



Şekil 1. Madde Seçme Yöntemlerinin Mutlak Yanlılık ve Korelasyon Değerleri

Şekil 1'de madde seçme yöntemlerinin test desenine, boyutlararası korelasyon değerine ve boyut başına düşen madde sayısına göre mutlak yanlılık ve korelasyon değerleri gösterilmiştir. Şekil 1 incelendiğinde, boyular arası korelasyonun düşük (ρ=0.2) ve boyut başına düşen madde sayısının n=10 olduğu durumda basit yapıdaki testlerde V1; test karmaşıklaştığı durumlarda ise KL madde seçme yöntemi en iyi performansı göstermiştir. Boyular arası korelasyonun yüksek (ρ=0.8) ve boyut başına düşen madde sayısının n=10 olduğu durumda tüm test desenlerinde V1 madde seçme yöntemi en iyi performansı göstermiştir.

Boyutlar arası korelasyonun düşük (ρ=0.2) ve boyut başına düşen madde sayısının n=15 olduğu durumda basit yapıda V1, düşük karmaşık yapıda Vol ve yüksek karmaşık yapıda ise KL madde seçme yöntemi en iyi performansı göstermiştir. Boyutlar arası korelasyonun yüksek (ρ=0.8) ve boyut başına düşen madde sayısının n=15 olduğu durumda basit yapıda KL, düşük karmaşık yapıda Vol ve yüksek karmaşık yapıda ise V1 madde seçme yöntemi en iyi performansı göstermiştir.

SONUÇLAR ve TARTIŞMA

Araştırmada üç farklı madde seçme yönteminin performansı karşılaştırılmıştır. Sonuçlar incelendiğinde önce V1 daha sonra da KL yönteminin daha iyi performans gösterdiği belirlenmiştir. Yao (2012) her bir boyuttan belirli sayıda madde seçilmesi durumunda KL ve V1 madde seçme yöntemlerinin alt boyut ve toplam puanları kestirmede daha iyi sonuçlar vereceğini belirtmiştir. Bu bakımdan sonuçlar literatürdeki araştırmalarla örtüşmektedir. Ayrıca V1 yönteminin daha iyi performans göstermesindeki en büyük neden ÇBBBT işlemlerinde ÇBMTK uygulamaları sonucu elde edilen boyutlara arası teorik ağırlıklandırmalar yerine Tablo 1'de belirtilen kestirilen ağırlıkların kullanılmasıdır.

Toplam puanların kestirilmesinde BY test deseninde V1 madde seçme yöntemi daha ağırlıklı olarak seçilmekte iken, test yapısının karmaşıklaştığı durumda KL ve Vol yöntemleri de iyi performans göstermiştir. Özellikle Vol yöntemi her bir boyuttan elde edilen bilgiyi eşit dağıtmaya çalışmaktadır (Yao, 2012). Test karmaşık yapıya doğru gittikçe bilginin boyutlara belirli oranlarda dağıldığı bilinmektedir. Bu bakımdan test karmaşıklaştıkça Vol madde seçme yönteminin belirli koşullarda daha iyi performans vermesi beklenen bir durumdur. KL diğer iki yöntemden farklıdır ve MDISC değerine göre madde seçmektedir. KL olabilirlik fonksiyonuna göre maddeleri seçtiğinden dolayı, olabilirlik fonksiyonları birbirinden uzak olan yetenek dağılımlarında çok az sayıda iyi maddeleri seçme eğilimindedir. Araştırmada sabit sayıda soru sorulduğundan ve boyutlardaki sorular önceden belirlendiğinden dolayı KL madde seçme yönteminin avantajlarının tam olarak yansıtılmadığı düşünülmektedir.

Genel çerçevede bakıldığında madde sayısı arttığında korelasyonların arttığı ve mutlak yanlılık ve hata değerlerinin ise azaldığı belirlenmiştir. Toplam yetenek puanlarında boyut başına düşen madde sayısının 15 olduğu bir teste ait mutlak yanlılık değeri ise tüm koşullarda boyut başına düşen madde sayısının 10 olduğu bir testin mutlak yanlılık değerlerinden düşük olarak kestirilmiştir. Elde edilen bu bulgular literatürdeki çalışmaları destekler niteliktedir (Lee, 2014; Su, 2016; Yao, 2014; Yao, Pommerich ve Segall, 2014).

Tüm test desenlerinde, madde seçme yöntemi fark etmeden, boyutlar arası korelasyon değeri arttığında boyut başına düşen madde sayısına göre toplam yetenek puanlarına ait mutlak yanlılık değerlerinde farklılaşmalar görülmüştür. Basit ve düşük karmaşık yapı test deseninde test katı çok boyutluluk özelliğinden tek boyutluluğa yaklaştığı durumlarda kısa test için mutlak yanlılık değerleri önce artmış daha sonra V1 ve Vol kestirimleri için azalmıştır. Uzun testlerde ise her üç madde seçme yönteminde önce bir azalış daha sonra da bir artış görülmüştür. Ancak uzun testlerde yapı karmaşıklaşmaya başlayınca boyutlar arası korelasyon arttığında her üç madde seçme yöntemine göre mutlak yanlılık değerleri düzgün şekilde artmıştır. Yüksek karmaşık yapılarda ise hem kısa hem uzun testler madde seçme yöntemleri farketmeksizin, boyutlar arası korelasyon değerleri arttıkça daha fazla mutlak yanlılık değeri üretmişlerdir.

Araştırmada ayrıca tüm madde seçme yöntemlerinde testin karmaşıklığı arttığında yanlılık ve hata değerlerinin azaldığı görülmüştür. Testin karmaşıklığı arttıkça, boyutlar arası korelasyon değerinin 0.2 olduğu durumda toplam puanların raporlanmasındaki korelasyon değerleri arasındaki fark fazla iken, boyutlar arasındaki korelasyonun 0.8 olması durumunda ise birbirine çok yakın değerler elde etmişlerdir. Bu bakımdan testin hem karmaşıklığının artırılması hem de boyutlar arası korelasyonun artırılması toplam puanları rapor etmede mutlak yanlılıklarda önemli bir durumdur. Bu durumun telafi edici modellerin etkisinden olduğu düşünülmektedir. Elde edilen bulgular literatürdeki çalışmaları destekler niteliktedir (Su, 2016).

Araştırma PISA 2012 Türkiye verilerine dayalı bir simülasyon çalışması olarak ele alınmış ve gerçek test koşullarına en yakın durumlar manipüle edilmiştir. Araştırma gerçek ÇBBBT uygulamaları üzerinden yapılarak madde ve yetenek dağılımına ilişkin sonuçlar karşılaştırılabilir. Ayrıca, ÇBBBT yöntemlerinin etkililiklerin ortaya konmasında madde havuzlarının önemi büyüktür. Bu bakımdan benzer koşullar daha büyük çok boyutlu madde havuzlarında test edilebilir ve madde seçme yöntemlerinin performansı karşılaştırılabilir.

KAYNAKÇA

Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255-278. doi: 10.1207/s15324818ame0704_1

Ackerman, T. A., Gierl, M. J., & Walker, C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37–53. doi: 10.1111/j.1745-3992.2003.tb00136.x

Adams R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. doi: 10.1177/0146621697211001

Bloxom, B., & Vale, C. D. (1987). *Multidimensional adaptive testing: An approximate procedure for updating*. Paper presented at the annual meeting of the psychometric society. Montreal, Canada.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459. doi: 10.1007/BF02293801

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Unpublished doctoral dissertation). University of Minnesota, USA.

Fan, M., & Hsu, Y. (1996). *Multidimensional computer adaptive testing*. Paper presented at the annual meeting of the American Educational Testing Association, New York City, NY.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement, 34*(1), 10–26. doi: 10.1177/0146621609336112

Lee, M. (2014). *Application of higher-order IRT models and hierarchical IRT models to computerized adaptive testing* (Unpublished doctoral dissertation). University of California, Los Angeles.

Liu, F. (2015). *Comparisons of subscoring methods in computerized adaptive testing: A simulation study* (Unpublished doctoral dissertation). University of North Carolina Greensboro, North Carolina, USA.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404. doi: 10.1177/014662169602000406

Luecht, R. M. (2004). *MIRTGEN 2.0 Manual*. Department of Educational Research Methodology, University of North Carolina at Greensboro, Greensboro, NC.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*(3), 279-293. doi: 10.1177/014662169201600308

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006). *Scalability and the development of useful diagnostic scales*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Luo, X. (2013). *The optimal design of the dual-purpose test* (Unpublished doctoral dissertation). University of North Carolina Greensboro, North Carolina, USA.

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp.77-101). New York: Springer.

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*(2), 273–296. doi: 10.1007/s11336-008-9097-5

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354. doi: 10.1007/BF02294343

Su, Y. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied Psychological Measurement, 40(*5), 346-360. doi: 10.1177/0146621616639305

Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait* (Unpublished doctoral dissertation). Columbia University.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398–412. doi: 10.3102/10769986024004398

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*(4), 575–588. doi: 10.1007/BF02295132

Wang, C., & Chang, HH. (2011). Item selection in multidimensional computerized adaptive tests: Gaining information from different angles. *Psychometrika, 76*(3), 363-384. doi: 10.1007/s11336-011-9215-7

Wang, C., Chang, HH., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika, 76*(1), 13-39. doi: 10.1007/s11336-010-9186-0

Wang, C., Chang, HH., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*, 99-122. doi: 10.1177/0146621612463422

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28,* 295-316. doi: 10.1177/0146621604265938

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9,* 116-136. doi: 10.1037/1082-989X.9.1.116

Yao, L. (2011). *simuMCAT: Simulation of multidimensional computer adaptive testing* [Computer software]. Monterey: Defense Manpower Data Center.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika, 77,* 495-523. doi: 10.1007/s11336-012-9265-5

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37,* 3-23. doi: 10.1177/0146621612455687

Yao, L. (2014). Multidimensional CAT item selection procedures with item exposure control and content constraints. *Journal of Educational Measurement, 51,* 18-38. doi: 10.1111/jedm.12032

Yao, L., & Boughton, K. A. (2009). Multidimensional linking for tests containing polytomous items. *Journal of Educational Measurement, 46,* 177–197. doi: 10.1111/j.1745-3984.2009.00076.x

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet precise, screening test. *Applied Psychological Measurement, 38*(8), 614-631. doi: 10.1177/0146621614541514

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36,* 375–398. doi: 10.1177/0146621612445904

EXTENDED ABSTRACT

Introduction

A test can be designed for many purposes, including the ranking of people along a continuum or providing diagnostic value about examinees. However, a very common problem that often arises is the reporting the overall scores when items are designed for multidimensional purposes. Multidimensional computer adaptive testing (MCAT) is capable of measuring multiple dimensions efficiently by using multidimensional IRT (MIRT) applications.

Maximizing the determinant of the fisher information matrix (Vol), Minimizing the error variance of the linear combination (V1) and Kullback-Leibler (KL) item selection methods allow sufficient items from each content by incorporating information from several dimensions simultaneously. Volume item estimation method includes prior knowledge of the joint distribution of ability in a Bayesian framework (Segall, 1996). Volume item selection estimates ability of each domain with using MIRT model by using selected items. This information is used as a prior distribution to select next item which is believed to contribute to the precision of ability estimates. KL information measures the distance between two likelihoods at true ability and current ability and it is concluded that KL information is a better indicator discriminating true and estimated ability based on posterior densities and doesn't require ability levels close to each other (Veldkamp and van der Linden, 2002). Also KL information overcomes the attenuation paradox which helps to estimate correct θ values rather than using Fisher information. Minimizing the error variance of the linear combination (V1, Variance) is more effective when item pool is designed to measure more than one abilities whether these domains are correlated or not. It is stated that V1increasses the precision for overall scores (van der Linden, 1999). Method simply selects the item which contributes minimum error variance for the composite score of equal weighted domains

There have been several research studies about MCAT item selection methods to improve the overall ability score estimations accuracy (Wang and Chang, 2011; Yao, 2012, 2013). According to the literature review it has been found that most studies focused on comparing item selection methods in many conditions except for the structure of test design. In contrast with the previous studies, this study employed various test designs (simple and complex) which allow the evaluation of the overall ability score estimations across multiple real test conditions. The purpose of this study is to compare MCAT item selection methods while estimating the overall ability scores in terms of test design, correlations between dimensions and number of items per dimension in MCAT framework. This study also aims

to find the better item selection procedure which produces higher precisions for the composite score estimations. The comparison is performed by examining the absolute bias (ABSBIAS), root mean square error (RMSE) and correlation between true and estimated ability scores.

Method

In this study, four factors were manipulated, namely the test design, number of items per dimension, correlation between dimensions and item selection methods. For each simple structure (SS), complex low structure (CLS) or complex high structure (CHS) design 1000x3 and 1000x45 matrix of true ability parameters were randomly generated from the multivariate normal distribution. Using the generated item and ability parameters, dichotomous item responses were generated by using M3PL compensatory multidimensional IRT model with specified correlations. A three-dimensional item pool was simulated with simple and complex structures. Dimensions correlated at $\rho = 0.2$, 0.5, and 0.8. For item calibration, multidimensional Bock and Aitkin's EM algorithm (BAEM) calibration method were employed. The multidimensional CAT item selection procedures: minimum angle, minimize the error variance of the composite score with the optimized weight, and Kullback–Leibler (KL) information were also examined. MCAT composite ability score accuracy was evaluated using absolute bias (ABSBIAS), correlation and the root mean square error (RMSE) between true and estimated ability scores.

Results and Discussion

The results suggest that the multidimensional test structure, number of item per dimension and correlation between dimensions had significant effect on item selection methods for the overall score estimations. For SS test design it was found that V1 item selection has the lowest absolute bias estimations for both long and short test while estimating overall scores. For CLS test design it was found that KL item selection has the lowest absolute bias estimations for short test and Vol item selection has the lowest absolute bias estimations for long test while estimating overall scores. For CHS test design it was found that KL item selection has the lowest absolute bias estimations for both long and short test while estimating overall scores.

As the model gets complex absolute biases have decreased significantly for overall scores. Based on the findings V1 item selection had the most accurate estimations for the overall scores. As the number of item increased, correlations tend to increase however, absolute bias and errors decreased. As expected, longer test provided more accurate scores. Correlations increased in overall ability score when the complexity of test increased. Results also suggest that, the overall scores were more sensitive to test complexity (trait contamination), multidimensionality and test length. In all conditions, longer test produced the lowest ABSBIAS and RMSE values and higher correlations.