ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 365-384

İçerik Ağırlıklandırmasının Maddeler-Arası Boyutluluk Modeline Dayalı Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş Test Yöntemleri Üzerindeki Etkisinin İncelenmesi

Examining the Effect of Content Balancing on Multidimensional Computerized Adaptive Testing Based on Between-Item Dimensionality Model

Burhanettin ÖZDEMİR ** Selahattin GELBAL ***

Öz

Bu çalışmanın amacı, maddeler-arası boyutluluk modeline dayalı Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş (BOB) Test Yöntemlerinin performanslarının karşılaştırılması ve içerik ağırlıklandırmasının (content balancing) çok boyutlu BOB testi yöntemleri üzerindeki etkisinin incelenmesidir. Bu amaç doğrultusunda, 2009-2013 eğitim ve öğretim yıllarında Hacettepe Üniversitesi tarafından uygulanan İngilizce Yeterlik Sınavına (İYS) ait gerçek veri seti kullanılmıştır. Her bir testte yer alan dinleme, dilbilgisi ve okuduğunu anlamaya ilişkin maddeler ile üç boyutlu gerçek madde havuzu oluşturulmuştur. Maddeler-arası boyutluluk modeli ile kalibre edilerek oluşturulan madde havuzu toplamda 555 maddeden oluşmaktadır. En uygun çok boyutlu BOB testini belirlemek amacıyla; iki farklı yetenek kestirim yöntemi (Bayesyen MAP ve Fisher'in puanlama yöntemi), üç farklı madde seçim yöntemi (A-optimality, D-optimality ve seçkisiz) ve hata varyansı durdurma kuralına dayalı üç farklı ölçüt kullanılmıştır. Ayrıca içerik ağırlıklandırmasının çok boyutlu BOB testi yöntemleri üzerindeki etkisini incelemek amacıyla, içerik ağırlıklandırmasının yapıldığı ve içerik ağırlıklandırmasının yapılmadığı koşullara ilişkin bulgular karşılaştırılmıştır. Her bir koşula ilişkin çok boyutlu BOB testi bulguları, boyutlara ilişkin güvenirlik katsayıları, ölçmenin standart hatası ve RMSD değerlerine bakılarak karşılaştırılmıştır. Analiz sonuçlarına göre, A-Optimality madde seçim yöntemi kullanıldığında, hem Bayesyen MAP hem de Fisher'in Puanlama yöntemlerinin benzer sonuçlar verdiği bulgusuna ulaşılmıştır. Buna karşın, Fisher'in puanlama yönteminin hem madde seçim yöntemlerinden hem de içerik ağırlıklandırmasından etkilendiği söylenebilir. Ayrıca içerik ağırlıklandırması uygulandığında her bir koşul için testteki ortalama madde sayısı artarken, güvenirlik katsayılarının azaldığı, buna karşın RMSD ve standart hataların azaldığı bulgusuna ulaşılmıştır.

Anahtar Kelimeler: Çok boyutlu bireyselleştirilmiş testler, maddeler-arası boyutluluk modeli, içerik ağırlıklandırması,

Abstract

The purpose of this study is to compare the performance of Between-item dimensionality-based Multidimensional CAT designs and to examine the effect of content balancing on different MCAT designs. For this purpose, real data set obtained from English Proficiency Test (EPT), which was administered by Hacettepe University between 2009 and 2013 academic years, was used. The three dimensional item pool consisted of real items measuring students' listening, grammar and reading abilities. Item pool consisted of 555 items

* Bu çalışma, birinci yazarın Prof. Dr. Selahattin GELBAL danışmanlığında tamamlanan doktora tezinden türetilmiştir

** Arş. Grv. Dr., Siirt Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, b.ozdemir025@gmail.com
*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara-Türkiye, s.gelbal@hacettepe.edu.tr

Geliş Tarihi: 07.12.2015
Kabul Tarihi: 25.12.2015

which was calibrated with 2PL between-item MIRT model. In this study, two different theta estimation (Fisher scoring and Bayesian MAP) methods, three different fisher information based item selection methods (A-optimality, D-optimality and Random) and three different precision based termination methods were used in order to determine the best MCAT design. In addition, results of MCAT algorithms with content distribution and without content distribution were compared so as to examine the effect of content balancing in the context of MCAT. The results of each MCAT condition were compared with respect to, reliability index, SEM, RMSD values associated with each dimension. According to results, both Bayesian MAP and Fisher's scoring methods yielded similar results when A-Optimality item selection method was used. However, Fisher's scoring method appeared to be affected from item selection methods and content balancing. Moreover, average number of items tended to increase and reliability coefficients tended to decrease somewhat, while standard error and RMSD values tended to decrease when content balancing was applied in MCAT.

Keywords: Multidimensional adaptive testing, between-item dimensionality model, content balancing

GİRİŞ

Bilgisayar teknolojilerindeki gelişmeler sosyal hayatımızı, çevremizi ve yaşam tarzımızı etkilediği gibi eğitimde kullanılan ölçme ve değerlendirme araçlarını, yöntem ve tekniklerini de aşamalı olarak etkilemektedir. Bu gelişmelere bağlı olarak günümüzde daha nitelikli ve verimli eğitim vermek için bilgisayar teknolojilerinden önemli ölçüde faydalanılmaya çalışılmaktadır. Bu amaç doğrultusunda bireylerin özelliklerini veya yeteneklerini bilgisayar teknolojilerini kullanarak ölçmeyi amaçlayan ölçme yöntemleri geliştirilmiştir. Bunlardan en temel olanı testlerin kâğıt-kalem yerine bilgisayar ortamında sorulmasıdır. Bu yöntem genellikle bilgisayar destekli testler (Computer-Based Tests-CBT) olarak adlandırılmaktadır.

Diğer bir alternatif ölçme yöntemi, bireyin yetenek düzeyi ile maddelerin özelliklerinin bilgisayar ortamında eşleştirildiği bilgisayar ortamında bireyselleştirilmiş test (Computerized Adaptive Testing-CAT) yöntemleridir (McBride ve Martin, 1983; Weiss ve Kingsbury, 1984). Bilgisayar ortamında bireyselleştirilmiş (BOB) testler, psikolojik ve eğitimsel testlerin daha etkili ve verimli bir şekilde uygulanması için yeniden düzenlenerek interaktif bilgisayarlar ile uygulanmasıdır (Van der Linden ve Glas, 2000; Wainer et al., 2000). Bu yöntemin temel amacı, her bireyin ölçülen özelliğini en etkili ve verimli bir şekilde ölçecek bireyin yetenek düzeyine uygun maddeleri seçmektir.

Bilgisayar ortamında bireyselleştirilmiş testler, Binet'in geliştirmiş olduğu uyarlanmış test yönteminin uygulayıcı birey yerinde bilgisayar programı kullanılarak uygulanmasıdır. BOB testi yönteminin uygulanma sürecinde test maddeleri kullanılacak bilgisayar programına yüklenir ve bilgisayar ekranında görünmesi sağlanır. Daha sonra birey klavyeyi veya fareyi kullanarak maddelere cevap verir.

Binet testini uygulayan araştırmacı gibi, bilgisayar programı bireyin teste nasıl başlayacağına karar verir; bireyin önceki maddelere vermiş olduğu cevaplara bağlı olarak maddeleri seçer ve bir ya da birkaç kural belirleyerek testi sonlandırır. İlk BOB uygulamaları Binet'in kullanmış olduğu yöntemin farklı versiyonları iken (Weiss, 1973), sonraki uygulamalar ise madde havuzunun oluşturulma biçimine göre farklılık göstermektedir (Lord, 1971a, 1971b, 1971c).

Madde tepki tepki kuramının geliştirilmesine paralel olarak, bilgisayar teknolojisindeki ilerlemeler, bilgisayar ortamında bireyselleştirilmiş testlerin uygulanabilirliğini arttırmıştır. Madde seçim ve yetenek kestirim yöntemleri için tek boyutlu madde tepki kuramının kullanıldığı bireyselleştirilmiş testler tek boyutlu bilgisayar ortamında bireyselleştirilmiş test (Tek boyutlu BOB testleri-Unidimensional CAT) olarak adlandırılmaktadır (Wang ve Chen, 2004).

BOB testlerinde tek boyutlu MTK yaygın olarak kullanılmasına rağmen, gerçek test uygulamaları için uygun olmayabilir. Özellikle bilişsel özelliklerin ölçülmesi ve değerlendirilmesinin gerektiği portfolyo değerlendirmeleri, performans görevleri, klinik yeteneklerinin ölçülmesi ve değerlendirilmesi, yazma ve konuşma becerilerinin ölçülmesi ve projelerin değerlendirilmesi söz konusu olduğunda bireylerin yeteneklerinin doğru bir şekilde ölçülmesi için çok boyutlu madde tepki

kuramına ihtiyaç duyulmaktadır (van der Linden ve Hambleton, 1997, s. 221). Çok boyutlu madde tepki kuramının uygulama alanlarının yaygınlaşması ve bilgisayar ortamında bireyselleştirilmiş testlerin kâğıt-kalem testlerine alternatif olarak görülmesi, her iki yöntemin birleşimi olan *çok boyutlu bireyselleştirilmiş testlerin* (Segall, 1996, 2001) geliştirilmesine olanak sağlamaktadır.

Bireyselleştirilmiş testlerin geleneksel kâğıt-kalem testlerine göre avantajlı yönleri olduğunu belirtilmesine karşın, bu durum her bireyselleştirilmiş testin geleneksel kağıt-kalem testlerinden üstün olduğu anlamına gelmeyebilir. Ayrıca en uygun BOB testine karar vermek için testin içeriğine ve ölçtüğü özelliğin yapısına uygun bir BOB testi algoritmasının geliştirilmesi gerekir. Dolayısıyla, her hangi bir BOB testi geliştirme aşamasında cevaplanması gereken bazı temel sorular vardır. Bu temel sorular; hangi modelin kullanılacağı, ilk maddenin nasıl seçileceği, test sürecinde bireyin yetenek parametresinin nasıl hesaplanacağı, bir sonraki maddenin nasıl seçileceği ve testin nasıl sonlandırılacağıdır (Diao ve Reckase, 2009). Bu sorular uygun bir şekilde cevaplandırıldığında geliştirilmesi amaçlanan en uygun BOB testi belirlenmiş olur. Ayrıca, bu temel sorular, BOB testi geliştirme sürecinin aşamalarını oluşturmaktadır.

Madde Seçim Yöntemleri

Çok boyutlu bilgisayar ortamında bireye uyarlanmış testlerde kullanılan madde seçim yöntemleri ile ilgili ilk çalışmalar Bloxom ve Vale (1987) tarafından yürütülmüştür. Bloxom ve Vale (1987)'in çok boyutlu BOB testleri ile ilgili yapmış olduğu çalışmayı Fan ve Hsu (1996), Luecht (1996), Segall (1996, 2000), van der Linden (1996, 1999), ve Veldkamp ve van der Linden'in (2002) yapmış olduğu çalışmalar takip etmiştir.

Çok boyutlu BOB testlerinde kullanılan madde seçim yöntemlerinden bazıları optimal desenlere bağlı madde seçim yöntemleridir. Optimal desenler bilgi matrisinin veya kovaryans matrisinin determinantı veya izinin istatistiksel çıkarımları optimize etmede kullanıldığı yöntemlerdir. Bu yöntemler sırasıyla D-optimality ve A-optimality olarak adlandırılmaktadır (Silvey, 1980, s. 10). Fisher'in bilgi matrisi bu yöntemlerde önemli bir rol oynamaktadır. Çünkü Fisher'in bilgi matrisi gözlenen değişkenleri açıklayan gizil değişkenlere ait bilgiyi ölçmekte kullanılır.

Optimal desenler kullanılarak çok boyutlu BOB testleri için geliştirilmiş birçok madde seçim yöntemleri vardır. Bunlar; D-optimality (determinant of the posterior information matrix) (Luecht, 1996; Segall, 1996), A-optimality (minimizing the error variance of the composite measure) (van der Linden, 1999), C-optimality ve önsel dağılımın maksimize edildiği Kullback–Leibler bilgi fonksiyonun (Kullback–Leibler information- KLI) (Veldkamp ve van der Linden, 2002) yöntemleridir. Aşağıda çok boyutlu BOB testlerinde kullanılan madde seçim yöntemleri hakkında bilgi verilmiştir.

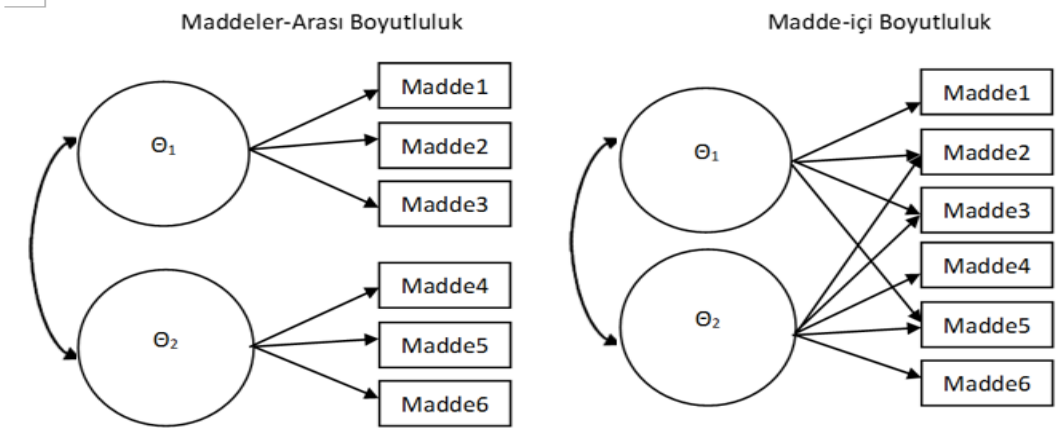
Geleneksel kâğıt-kalem testlerinde, test geliştirme sürecinde testin uygulanacağı alanın içeriğinin özellikleri göz önünde bulundurularak testi oluşturacak maddeler belirlenir. BOB testlerinde bireyin kestirilen yetenek parametresine ait en yüksek bilgiyi veren maddeler seçildiğinden test ile ölçülmek istenen farklı içeriklere ilişkin madde dağılımı farklılık gösterebilir. Dolayısıyla, bireyler ölçülen her bir içeriğe ya da konu alanına ait farklı sayıda maddelere cevap verir. Bu durum BOB testinin geçerliğini, puanların karşılaştırılabilirliğini tehlikeye sokmakla birlikte, testi alan ve testi uygulayan bireyler açısından sorun oluşturabilir. Bilgisayar ortamında bireyselleştirilmiş testlerde, yukarıda belirtilen problemleri ortadan kaldırmak için ilk olarak Green ve arkadaşları (1984) tarafından içerik ağırlıklandırması (content balancing) görüşü ortaya atılmış ve ilerleyen yıllarda farklı içerik ağırlıklandırması yöntemleri geliştirilmiştir.

Çok boyutlu MTK modelleri

Çok boyutlu BOB testlerinin temelini oluşturan çok boyutlu madde tepki kuramları Telafi-edici (compensatory) ve telafi-edici (noncompensatory) olmayan çok boyutlu modeller olarak ikiye ayrılır. Çok boyutlu modellerin dışında dikkat edilmesi gereken bir diğer nokta ise madde düzeyinde boyutluluktur. Madde düzeyindeki boyutluluk ise maddeler-arası boyutluluk (between-item dimensionality) ve madde-içi boyutluluk (within-item dimensionality) olmak üzere ikiye ayrılır

(Wang, Chen ve Cheng, 2004; Wang, Wilson ve Adams, 1997). Maddeler-arası boyutlulukta her bir madde sadece bir boyutta yük verir. Bu modelde maddelere ait ayırt edicilik parametreleri bir boyuttan sıfırdan farklı değer alırken diğer boyutlara ait ayırt edicilik parametresi sıfıra eşittir. Buna karşın, Madde-içi boyutluluk modelinde maddeler, birden fazla boyutta yük verir. Bu modelde maddelere ait madde ayırt edicilik parametresi ve madde yükleri diğer boyutlar içinde sıfırdan farklı değerler alabilir.

Şekil 1'de iki farklı boyutu ölçen bir test için telafi-edici çok boyutlu modellere ait madde düzeyinde boyutluluk modellerinden maddeler-arası ve madde-içi boyutluluk modelleri gösterilmiştir.



Şekil 1. Maddeler-arası ve madde-içi boyutluluk modeli (Wang ve Chen, 2004)

Bu çalışmada kullanılan telafi-edici çok boyutlu MTK modeli, maddelere ait şans başarının sıfıra eşit olduğu kabul edilen iki-parametreli çok boyutlu MTK modelidir (2PL-ÇBMTK/2PL MIRT). Bu modele ait formül aşağıda verilmiştir:

$$P(Uij = 1|\theta_j, a_i, d_i) = \frac{e^{(a_i \theta_j' + d_i)}}{1+e^{(a_i \theta_j' + d_i)}}$$

Burada θ_j, m yetenek düzeyinin ölçüldüğü 1xm şeklinde bir vektördür. Benzer şekilde a_i 1xm şeklinde ayrıt edicilik vektörü ve *d* kesişim parametresi ya da madde kolaylığı parametresi olarak adlandırılan değerdir.

Araştırmanın Amacı ve Önemi

Bu çalışmanın amacı, bireylerin yabancı dil yeteneklerinin maddeler-arası boyutluluk modeline dayalı Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş Test Yöntemleri ile ölçülmesi (Multidimensional Computerized Adaptive Testing-MCAT) ve içerik ağırlıklandırmasının (content balancing) çok boyutlu BOB testi yöntemlerinin performansları üzerindeki etkisinin incelenmesidir.

Alan yazınına bakıldığında, özellikle ülkemizde çok boyutlu BOB testi yöntemleri ile ilgili yapılan çalışmaların sınırlı sayıda olduğu görülmektedir. Bu araştırma çok boyutlu BOB testi yöntemlerinin performanslarının karşılaştırılmasına olanak sağladığından önemli görülmektedir. Ayrıca bu çalışmada testin yapısına bağlı olarak yapılan içerik ağırlıklandırmasının yapıldığı ve içerik ağırlıklandırmasının yapılmadığı duruma ilişkin analiz bulguları karşılaştırmaya olanak sağladığından önemli görülmektedir.

Bu çalışmayı önemli kılan bir diğer özelliği ise İYS sınavına ait gerçek verilerin kullanılması ve gerçek verilere dayalı simulasyon (post-hoc simulation) yönteminin kullanılmasıdır Monte Carlo simülasyon yönteminden temel farkı ise madde havuzunun gerçek maddelerden oluşmasıdır. Bu simülasyon yöntemi, genellikle, geleneksel kağıt kalem formatında kullanılan testin psikometrik özelliklerinde anlamlı bir değişim yapmadan BOB testi yöntemleri ile uygulandığında testteki

ortalama madde sayısında ne kadar azalma olacağını tespit etmeyi amaçlar (IACAT, 2015). Post-hoc simülasyon yönteminin diğer aşamaları, BOB testi yöntemi ile aynıdır.

Problem Cümlesi

İçerik ağırlıklandırmasının maddeler-arası boyutluluk modeline dayalı farklı yetenek kestirimi yöntemleri, madde seçim yöntemleri ve test sonlandırma kurallarının kullanıldığı çok boyutlu BOB testi yöntemlerinin performansları üzerindeki etkisi nasıldır?

Alt Problemler

1. A-optimality madde seçim yönteminin kullanıldığı çok-boyutlu BOB testi analizlerine ilişkin farklı yetenek kestirim yöntemleri ve durdurma kuralları altında her bir koşula ait güvenirlik katsayısı, standart hata, testin uzunluğu ve RMSD değerleri nasıldır?

2. D-optimality madde seçim yönteminin kullanıldığı çok-boyutlu BOB testi analizlerine ilişkin farklı yetenek kestirim yöntemleri ve durdurma kuralları altında her bir koşula ait güvenirlik katsayısı, standart hata, testin uzunluğu ve RMSD değerleri nasıldır?

3. Seçkisiz (random) madde seçim yönteminin kullanıldığı çok-boyutlu BOB test analizlerine ilişkin farklı yetenek kestirim yöntemleri ve durdurma kuralları altında her bir koşula ait güvenirlik katsayısı, standart hata, testin uzunluğu ve RMSD değerleri nasıldır?

4. İçerik ağırlıklandırmasının kullanılmadığı çok boyutlu BOB testleri ile kaşılaştırıldığında, İçerik ağırlıklandırmasının farklı madde seçim, yetenek kestirim yöntemleri ve durdurma kurallarının kullanıldığı çok-boyutlu BOB testi yöntemleri üzerindeki etkisi nasıldır?

Sayıltılar

Genellikle çok-boyutlu BOB testlerinde her bir boyuta ilişkin yetenek parametreleri arasındaki korelasyonun veya varyans-kovaryans matrisi önsel dağılımının bilindiği varsayılmaktadır (Yoo, 2011). Bu çalışmada bireylerin kestirilen yetenek parametreleri arasındaki korelasyona bakılarak varyans-kovaryans matrisine ait önsel dağılımının bilindiği varsayılmaktadır (vcv=c[0.9, 0.8, 0,8])

Ayrıca, 2009-2013 yıllarında uygulanan İYS testlerinin bireyin ölçülen özelliklerini doğru ve güvenilir bir şekilde ölçtüğü ve bireylerin test maddelerine verdiği cevapların gerçek yetenek düzeylerini yansıttığı varsayılmaktadır.

Sınırlılıklar

Bu çalışmada, gerçek verilere dayalı simülasyon yapılmasına olanak sağlayan ve çok-boyutlu BOB testleri için geliştirilmiş "MAT " (multidimensional adaptive testing, Choi ve King, 2011) R paket programı kullanılmıştır. Dolayısıyla, bu çalışmada kullanılan madde seçim yöntemleri, yetenek kestirim yöntemleri, test sonlandırma kuralı ve madde kullanım sıklığı kontrol yöntemleri paket programda tanımlı yöntemlerle sınırlıdır.

Telafi edici Çok boyutlu madde tepki kuramına dayalı madde düzeyindeki boyutluluk modelleri *madde-içi* ve *maddeler-arası boyutluluk* modelleri olmak üzere ikiye ayrılmaktadır. Bu çalışma maddeler-arası boyutluluk modeline dayalı çok boyutlu BOB testi yöntemleri ile sınırlı tutulmuştur.

YÖNTEM

Araştırma Yöntemi

Bu araştırmada, yabancı dil sınavına giren bireylerin yabancı dil yeteneklerinin maddeler-arası boyutluluk modeline dayalı Çok Boyutlu Bilgisayara ortamında bireyselleştirilmiş (Çok-Boyutlu BOB) test yöntemleri ile kestirilmesi ve çok boyutlu BOB testi yöntemlerinin performanslarının karşılaştırılması amaçlanmaktadır. Araştırmada var olan yöntem ve tekniklerin gerçek veri üzerinden performanslarının karşılaştırılması amaçlandığından araştırma nicel karşılaştırma araştırmasıdır.

Çalışma grubu

Araştırmanın çalışma grubunu Hacettepe Üniversitesinde İngilizce Yeterlik Sınavı (İYS)'na giren bireyler oluşturmaktadır. Her bir dönem sonunda ve dönem içerisinde İYS'ye giren öğrenci sayısı

1200 ile 2000 arasında değişmektedir. Araştırmanın örneklemini ise 2008-2013 eğitim-öğretim yıllarında ilkbahar ve sonbahar dönemlerinde Hacettepe Üniversitesinde İYS'ye giren bireyler oluşturmaktadır. Analiz sürecinde gerçek madde parametrelerine bağlı olarak bireylere ilişkin yetenek parametreleri türetilmiş ve her bir analiz için aynı yetenek parametreleri kullanılmıştır. Çok-boyutlu BOB testi sürecinde her bir koşul için yapılan analizlerde testi alan birey sayısı 500 ile sınırlandırılmıştır. Bireylerin her bir boyuta ilişkin yetenek parametreleri [-3, +3] aralığında çok değişkenli normal dağılım göstermektedir.

Araştırma Verileri

Araştırmanın verilerini, 2008-2013 eğitim-öğretim yıllarında ilkbahar ve sonbahar dönemlerinde Hacettepe Üniversitesinde İYS'ye ait testlerdeki maddeler ve sınava giren bireylere ait cevap örüntülerinin yer aldığı veri seti oluşturmaktadır. Sınav dinleme (listening), okuduğunu anlama (reading) ve dilbilgisi (grammar) olmak üzere üç temel bölümden oluşup her bir testteki ortalama madde sayısı ise toplam 65'dir.

Verilerin Analizi

Verilerin analizi 2 aşamadan oluşmaktadır. İlk aşamada çok boyutlu bilgisayar ortamında bireyselleştirilmiş testlerde kullanılacak maddelere ait madde parametrelerinin maddeler-arası boyutluluk modeline dayalı tanımlayıcı çok boyutlu MTK modelleri ile kestirilmiş ve her bir modele ait madde havuzu oluşturulmuştur. Bu aşamada, testlere ait maddeler telafi-edici çok boyutlu MTK modellerinden maddeler-arası boyutluluk modeli ile kalibre edilerek maddeler-arası boyutluluk modeline ait madde havuzu oluşturulmuştur. 2009-2013 yılları arasında uygulanan ve üç boyuttan oluşan İYS sınavına ait 10 testteki toplam 628 madde, telafi-edici çok boyutlu madde tepki kuramına (Compensatory 2PL-MIRT) dayalı maddeler-arası boyutluluk modeli ile kalibre edilmiştir. Boyutlara ait madde ayırt edicilik parametresi 0,5'in altında olan ve d-parametresi [-4,4] aralığının dışında olan toplam 73 madde havuzunda çıkartılmıştır. Sonuç olarak, telafi-edici çok boyutlu MTK'ya dayalı maddeler-arası boyutluluk modeline ait madde havuzu, dinleme boyutunda 115, dilbilgisi boyutunda 240 ve okuma boyutunda 200 madde olmak üzere toplam 555 maddeden oluşmuştur.

İkinci aşamada madde havuzu oluşturulduktan sonra post-hoc simülasyon yöntemi uygulanarak simülasyon veri seti yerine İYS'ye ait gerçek veri seti kullanılarak bireylere ait yetenek kestirilmiştir. Bireylere ait yetenek kestirimi yapılırken çok boyutlu BOB testi analizi için R-yazılımında tanımlı "MAT" paket programı (Choi ve King, 2011) kullanılmıştır. Çok boyutlu BOB testi yöntemleri ile analiz yapılırken, farklı yetenek kestirme yöntemleri, madde seçim yöntemleri ve durdurma kuralları kullanılarak her bir boyuta ilişkin yetenek kestirimi yapılmıştır.

Maddeler-arası boyutluluk modeline dayalı çok boyutlu BOB testi için en uygun madde seçim yöntemini belirlemek amacıyla Fisher'in bilgi matrisine dayalı *A-optimality*, *D-optimality* ve her hangi bir kuralın kullanılmadığı *seçkisiz (Random)* madde seçim yöntemleri kullanılmıştır. Çok-boyutlu BOB testi için en uygun yetenek kestirim yöntemini belirlemek için maksimum likelihood estimation (MLE) yetenek kestirim yöntemlerinden Fisher'in puanlama yetenek kestirim yöntemi ve Bayesyen maksimum a posteriori (MAP) yetenek kestirim yöntemi kullanılmıştır. Ayrıca, çok-boyutlu BOB testinde en uygun test sonlandırma kuralını belirlemek için ise farklı hata varyansı durdurma kuralı kullanılmıştır.

Bu çalışmada, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelemek amacıyla farklı madde seçim yöntemleri, yetenek kestirim yöntemleri ve durdurma kurallarının kullanıldığı çok-boyutlu BOB testi algoritmalarına ait her bir boyuta ilişkin *güvenirlik katsayıları*, yetenek parametrelerine ilişkin *RMSD katsayıları* ve *ölçmenin standart hatası* değerleri hesaplanarak karşılaştırılmıştır. Her bir koşul için maddenin kullanım sıklığını (item exposure) kontrol etmeye olanak sağlayan randomesque yöntemi kullanılmıştır. Bu yöntem ile analizlerde madde havuzundan madde seçilirken test bilgisini maksimum yapan ilk on madde arasından birinin seçkisiz olarak atanması kuralı uygulanmıştır.

BULGULAR

Tablo 1'de maddeler-arası boyutluluk modeli için madde seçim yöntemlerinden A-optimality, durdurma kurallarından hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait analiz bulgularına yer verilmiştir.
Her bir model için standart hata durdurma kuralı kullanılarak analizler bireylerin kestirilen yetenek parametrelerine ait hata varyansının sırasıyla 0,20, 0,25 ve 0,30'un altına düştüğünde testler sonlandırılmıştır. Ayrıca her bir koşul için içerik ağırlıklandırmasının kullanıldığı ve kullanılmadığı durumlara ait sonuçlar karşılaştırılarak, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelenmiştir

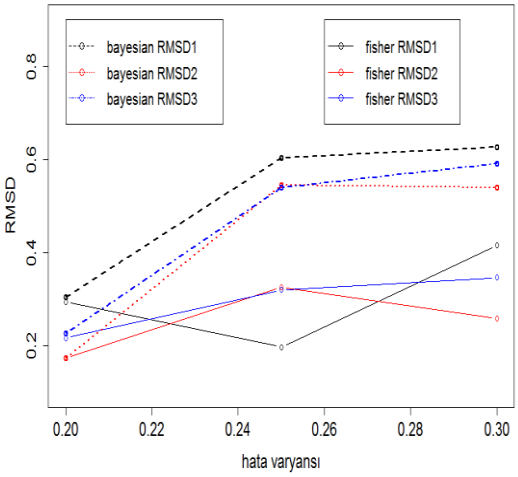
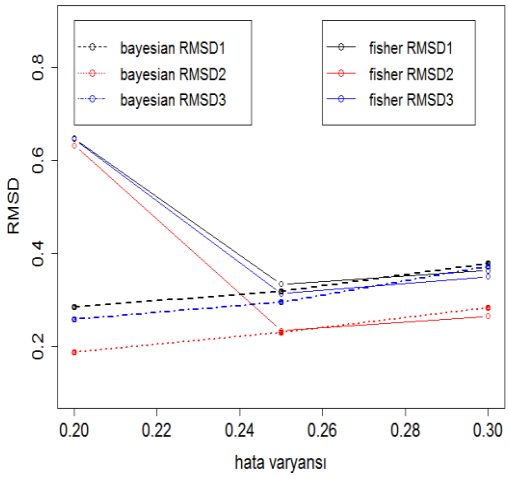
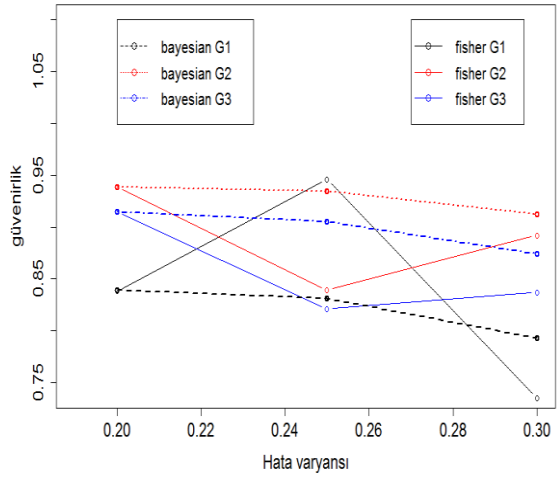
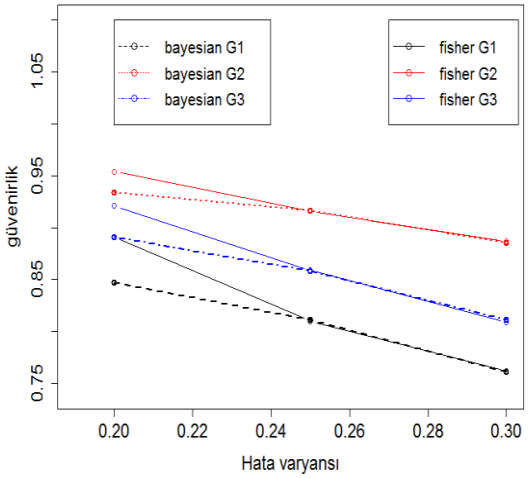
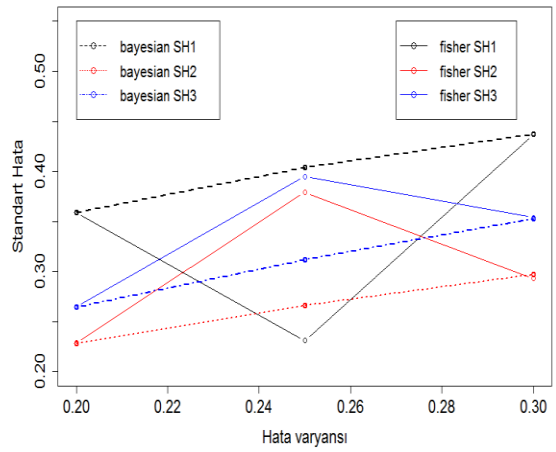
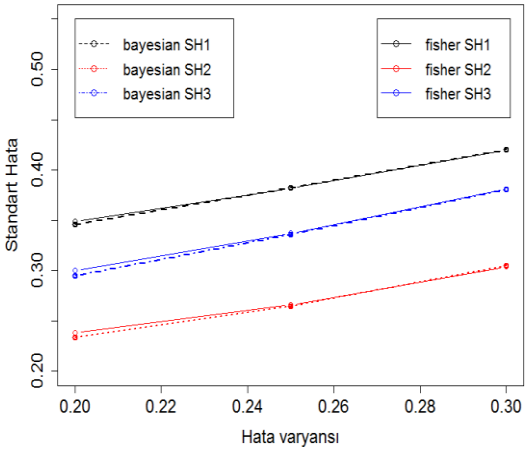
Tablo 1. A-Optimality madde seçim yöntemine ait Çok boyutlu BOB testi Bulguları

	Yetenek Kestirim yöntemi	Test sonlandırma kuralı	Test uzunluğu	güvenirlik			ÖSH*			RMSD		
		S. Hata	K	boy1	boy2	boy3	boy1	boy2	boy3	boy1	boy2	boy3
Eşit madde dağılımlı	Fisher	0,20	39,6	0,84	0,94	0,92	0,359	0,229	0,265	0,294	0,173	0,216
		0,25	31,4	0,95	0,84	0,82	0,231	0,379	0,395	0,197	0,326	0,320
		0,30	15,1	0,74	0,89	0,84	0,437	0,293	0,354	0,415	0,259	0,346
	Bayesian	0,20	39,7	0,84	0,94	0,92	0,359	0,228	0,265	0,305	0,174	0,226
		0,25	22,8	0,83	0,94	0,91	0,404	0,266	0,312	0,604	0,545	0,540
		0,30	15,2	0,79	0,91	0,87	0,437	0,297	0,353	0,626	0,540	0,591
İçerik ğırlıklandırmalı	Fisher	0,20	45,5	0,89	0,95	0,92	0,349	0,238	0,300	0,647	0,632	0,645
		0,25	29,7	0,81	0,92	0,86	0,382	0,266	0,337	0,334	0,234	0,314
		0,30	19,8	0,76	0,89	0,81	0,420	0,304	0,381	0,363	0,265	0,350
	Bayesian	0,20	44,9	0,85	0,93	0,89	0,346	0,234	0,295	0,285	0,187	0,259
		0,25	29,7	0,81	0,92	0,86	0,382	0,265	0,336	0,318	0,230	0,296
		0,30	20,1	0,76	0,89	0,81	0,420	0,305	0,381	0,379	0,283	0,372

*OSH= Ölçmenin standart Hatası

Tablo 1'deki çok-boyutlu BOB testi analiz bulgularına bakıldığında, genel olarak, madde seçim yöntemlerinden A-optimality, yetenek kestirim yöntemlerinden Fisher'in puanlama yöntemi ve standart hata durdurma kuralının 0,25 olduğu koşula ait testteki ortalama madde sayısının 31,4 olup daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir. Hata varyansı durdurma kuralı 0,30 olarak belirlendiğinde, testteki ortalama madde sayısı 15,1 e düşmesine karşın boyutlara ilişkin güvenirlik katsayılarının düştüğü, standart hata ve RMSD değerlerinin arttığı görülmektedir. Diğer taraftan, yetenek kestirim yöntemlerinden Bayesyen MAP yetenek kestirim yöntemi kullanıldığında standart hatanın 0,25'e eşitlendiği durumda ortalama madde sayısının 22,8'e düştüğü ve diğer koşullara göre güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

Tablo 1'deki içerik ağırlıklandırmasının yapıldığı duruma ilişkin çok boyutlu BOB testi sonuçlarına bakıldığında ise standart hata durdurma kuralının 0,25 olduğu koşula ait testteki ortalama madde sayısının 29,7 olup hem Fisher'in puanlama yöntemi hem de Bayesyen MAP yetenek kestirim yöntemlerinin benzer sonuçlar verdiği görülmektedir. Ancak içerik ağırlıklandırmasının yapılmadığı ve Fisher'in puanlama yetenek kestirim yöntemi kullanıldığı durumda birinci boyuta ilişkin güvenirlik katsayısı en yüksek değere sahipken diğer koşullarda ise, ikinci boyuta ilişkin güvenirlik katsayısı en yüksek değere sahiptir. Dolayısıyla, içerik ağırlıklandırması yapıldığında Fisher'in puanlama yönteminin daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

Eşit madde dağılımlı	İçerik Ağırlıklandırmalı
a.1 RMSD- Hata varyansı ilişkisi	a.2 RMSD- Hata varyansı ilişkisi
	
b.1 Güvenirlik- Hata varyansı ilişkisi	b.2 Güvenirlik- Hata varyansı ilişkisi
	
c.1 Ölçmenin standart hatası- Hata varyansı	c.2 Ölçmenin standart hatası- Hata varyansı
	

Şekil 2. A-Optimality Madde Seçim Yöntemine İlişkin Grafikler

Şekil 2'de içerik ağırlıklandırmasının yapılmadığı eşit madde dağılımlı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin madde seçim yöntemlerinden A-optimality, testi durdurma kurallarından boyutlara ilişkin hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerinin kullanıldığı çok-boyutlu BOB testi analizlerine ait grafikler verilmiştir.

Şekil 2'de yer alan a.1 ve a.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait her bir boyuta ilişkin RMSD değerleri ile yetenek parametrelerine ilişkin hata varyansları arasındaki ilişkiyi vermektedir. Şekil 2 a.1'e bakıldığında eşit madde dağılımlı durumda, yetenek parametrelerine ilişkin hata varyansı arttıkça Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinin arttığı görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça RMSD değerlerinin önce artıp sonra azaldığı görülmektedir. Dolayısıyla Bayesyen yetenek kestirimi yönteminin daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin RMSD değerleri ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 2 a.2'ye bakıldığında, kestirilen yetenek parametrelerine ilişkin hata varyansı arttıkça Fisher'in puanlama yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinin önce azaldığı, daha sonra ise artma eğilimi gösterdiği görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirim yönteminde hata varyansı arttıkça boyutlara ilişkin RMSD değerlerinin düzenli olarak artma eğilimi gösterdiği görülmektedir. Ayrıca içerik ağırlıklandırması uygulandığında, Bayesyen yetenek kestirim yöntemine ait RMSD değerlerinin daha düşük olduğu görülmektedir.

Şekil 2'de yer alan b.1 ve b.2 grafikleri içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin yetenek kestirim yöntemlerine ait her bir boyuta ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansları arasındaki ilişkiyi vermektedir. Şekil 2 b.1'e bakıldığında yetenek parametrelerine ait hata varyansları arttıkça, Bayesyen MAP yetenek kestirimine yöntemine ait her bir boyuta ilişkin güvenirlik katsayılarının düzenli bir şekilde azaldığı görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça boyutlara ilişkin güvenirlik katsayılarının artıp azaldığı görülmektedir.

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 2 b.2'ye bakıldığında, her bir boyuta ilişkin hata varyansı arttıkça hem Bayesyen MAP hem de Fisher'in puanlama yetenek kestirim yöntemine ait güvenirlik katsayılarının azaldığı görülmektedir. Hata varyansı durdurma kuralı arttıkça testteki ortalama madde sayısı arttığından her iki yetenek kestirim yöntemine ait güvenirlik katsayıları birbirine yakın değerler aldığı görülmektedir. Dolayısıyla içerik ağırlıklandırması yapıldığı durumda hem Bayesyen hem de Fisher'in puanlama yöntemi benzer sonuçlar verdiği söylenebilir.

Şekil 2'de yer alan c.1 ve c.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait her bir boyuta ilişkin ölçmenin standart hatası ile hata varyansı arasındaki ilişkiyi vermektedir. Şekil 2 c.1'e bakıldığında, Bayesyen MAP yetenek kestirimi yöntemi kullanıldığında, yetenek parametrelerine ilişkin hata varyansı arttıkça, her bir boyuta ilişkin ölçmenin standart hatasının da arttığı görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirimi yöntemi yerine Fisher'in puanlama yöntemi kullanıldığında, boyutlara ilişkin hata varyansı arttıkça boyutlara ilişkin standart hatanın düzenli bir artış göstermediği görülmektedir.

Şekil 2 c.2'ye bakıldığında, boyutlara ilişkin hata varyansı arttıkça hem Fisher'in puanlama hem de Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin ölçmenin standart hatasının da arttığı görülmektedir. Ayrıca içerik ağırlıklandırmasının yapıldığı durumda, Fisher'in puanlama ve Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin ölçmenin standart hatasının birbirine yakın değerler aldığı görülmektedir.

Genel olarak hata varyansı durdurma kuralının kullanıldığı maddeler-arası çok boyutlu BOB testi analiz yöntemlerine ilişkin bulgular karşılaştırıldığında, hem Bayesyen MAP hem de Fisher'in puanlama yöntemlerine ait en uygun hata varyansı durdurma kuralının 0,25 olduğu görülmektedir. Ayrıca Fisher'in puanlama yöntemine ait çok boyutlu BOB testi sonuçlarının içerik ağırlıklandırmasından etkilendiği ve içerik ağırlıklandırmasının uygulandığı çok-boyutlu BOB testi uygulamalarında daha tutarlı sonuçlar verdiği görülmektedir. Diğer taraftan, Bayesyen yetenek kestirim yönteminin içerik ağırlıklandırmasından çok az etkilendiği ve özellikle içerik ağırlıklandırması uygulandığı durumda testteki ortalama madde sayısında az bir artış olmasına rağmen daha düşük standart hata ve RMSD değerlerine sahip olduğu yorumu yapılabilir.

Tablo 3'te maddeler-arası boyutluluk modeli için madde seçim yöntemlerinden D-optimality, durdurma kurallarından hata varyansı, yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait analiz bulgularına yer verilmiştir. Ayrıca her bir koşul için içerik ağırlıklandırmasının kullanıldığı ve kullanılmadığı durumlara ait sonuçlar karşılaştırılarak, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelenmiştir.

Tablo 2. D-Optimality madde seçim yöntemine ait Çok boyutlu BOB testi Bulguları

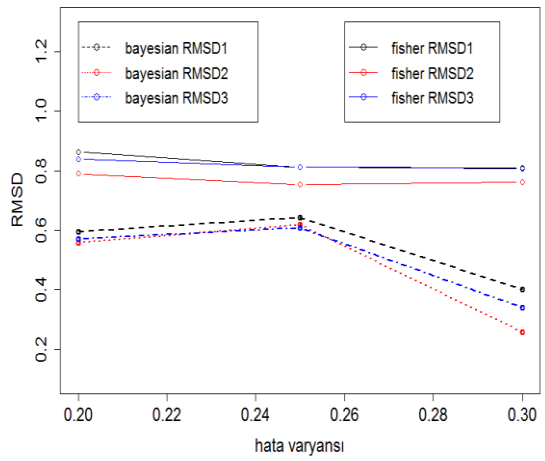
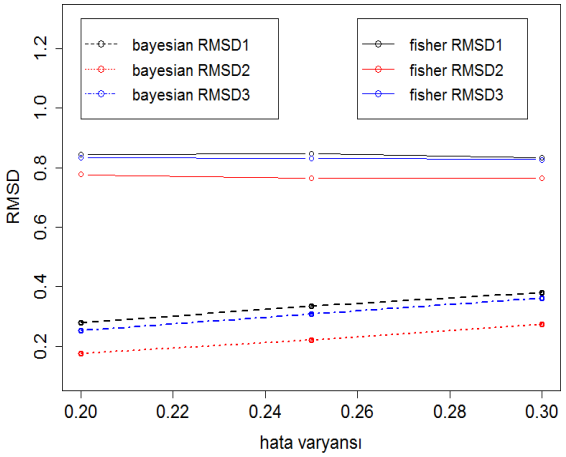
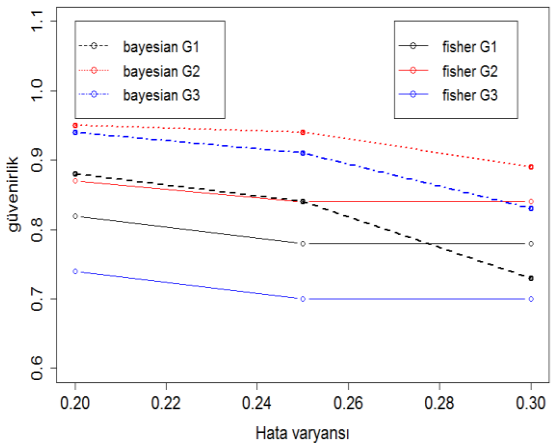
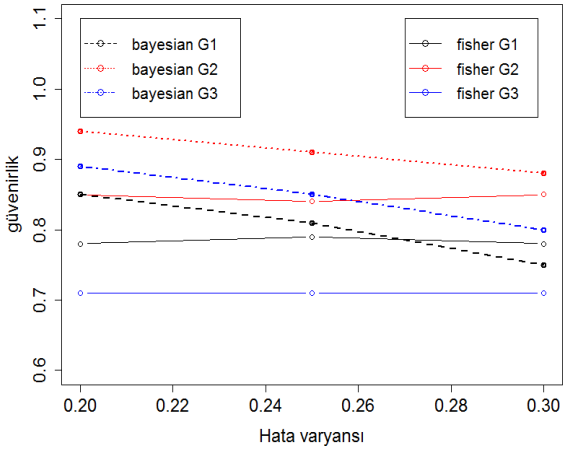
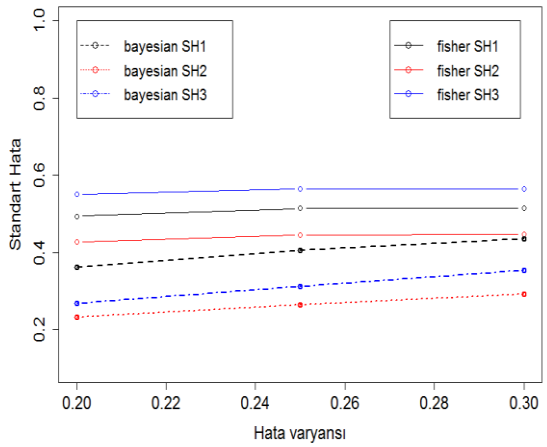
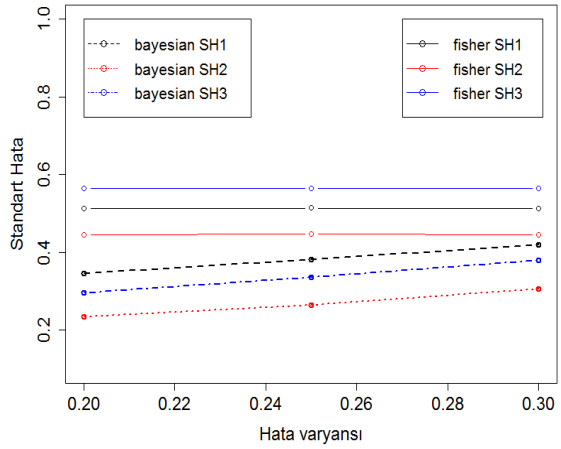
	Yetenek Kestirim yöntemi	Test Sonlandırma kuralı	Test uzunluğu	güvenirlik			ÖSH*			RMSD		
		S. Hata	K	boy1	boy2	boy3	boy1	boy2	boy3	boy1	boy2	boy3
madde		0,20	50,00	0,82	0,87	0,74	0,494	0,427	0,551	0,863	0,790	0,840
		0,25	50,00	0,78	0,84	0,70	0,514	0,446	0,565	0,812	0,753	0,812
	Fisher	0,30	49,43	0,78	0,84	0,70	0,514	0,447	0,565	0,806	0,763	0,811
Eşit dağılımlı		0,20	39,49	0,88	0,95	0,94	0,361	0,232	0,269	0,596	0,558	0,571
		0,25	22,83	0,84	0,94	0,91	0,406	0,265	0,311	0,644	0,620	0,608
	Bayesian	0,30	15,06	0,73	0,89	0,83	0,436	0,293	0,353	0,400	0,257	0,339
İçerik Ağırlıklandırmalı		0,20	50,00	0,78	0,85	0,71	0,513	0,445	0,564	0,844	0,777	0,833
		0,25	50,00	0,79	0,84	0,71	0,514	0,448	0,565	0,847	0,765	0,831
	Fisher	0,30	49,34	0,78	0,85	0,71	0,512	0,445	0,564	0,834	0,765	0,827
		0,20	45,31	0,85	0,94	0,89	0,346	0,235	0,297	0,278	0,175	0,253
		0,25	29,61	0,81	0,91	0,85	0,382	0,265	0,336	0,336	0,222	0,307
	Bayesian	0,30	19,59	0,75	0,88	0,80	0,420	0,306	0,380	0,381	0,275	0,362

*OSH= Ölçmenin standart Hatası

Tablo 2'deki çok-boyutlu BOB testi analiz bulgularına bakıldığında, yetenek kestirim yöntemlerinden Fisher'in puanlama yöntemi kullanıldığında, hata durdurma kuralının 0,30 olduğu koşulda bile testteki ortalama madde sayısının 49,43 olduğu görülmektedir. İngilizce yeterlik sınavının kâğıt-kalem testi formatındaki ortalama madde sayısı göz önünde bulundurularak çok-boyutlu BOB testi analizlerinde testteki maksimum madde sayısı 50 ile sınırlandırılmıştır. Böylece, standart hata durdurma kuralının kullanıldığı çok-boyutlu BOB testi analizlerinde çok uzun testler ile yetenek kestirimi önlenmiştir. Testteki ortalama madde sayısının diğer koşullarla karşılaştırıldığında oldukça yüksek olmasına karşın boyutlara ilişkin standart hata ve RMDS değerlerinin de yüksek olduğu görülmektedir. Diğer taraftan, yetenek kestirim yöntemlerinden Bayesyen MAP yetenek kestirim yöntemi kullanıldığında standart hatanın 0,25'e eşitlendiği durumda ortalama madde sayısının 22,8'e düştüğü ve A-Optimality madde seçim yönteminin kullanıldığı durum ile benzer sonuçlar verdiği söylenebilir.

Tablo 2'deki içerik ağırlıklandırmasının yapıldığı duruma ilişkin çok boyutlu BOB testi sonuçlarına bakıldığında ise Fisher'in puanlama yöntemi için testteki ortalama madde sayısında değişme olmazken güvenirlik katsayılarında az da olsa azalma olduğu görülmektedir. Genel olarak içerik ağırlıklandırmasının uygulandığı, her bir koşul için Bayesyen MAP yetenek kestirim yönteminin

daha az madde ile daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir. Fisher'in puanlama yönteminin kullanıldığı durumda boyutlara ilişkin hata varyanslarının azalması testteki ortalama madde sayısını ve her boyuta ilişkin güvenirlik katsayılarını etkilemediği görülmektedir.

Eşit Madde Dağılımlı	*İçerik Ağırlıklandırmalı*
a.1 RMSD- Hata varyansı ilişkisi	a.2 RMSD- Hata varyansı ilişkisi
	
b.1 Güvenirlik- Hata varyansı ilişkisi	b.2 Güvenirlik- Hata varyansı ilişkisi
	
c.1 Ölçmenin standart hatası- Hata varyansı	c.2 Ölçmenin standart hatası- Hata varyansı
	

Şekil 3. D-optimality Madde Seçim Yöntemine İlişkin Grafikler

Şekil 3'de eşit madde dağılımlı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin madde seçim yöntemlerinden D-optimality, testi durdurma kurallarından boyutlara ilişkin hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerinin kullanıldığı çok-boyutlu BOB testi analizlerine ait grafikler verilmiştir.

Şekil 3'de yer alan a.1 ve a.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait her bir boyuta ilişkin RMSD değerleri ile yetenek parametrelerine ilişkin hata varyansları arasındaki ilişkiyi vermektedir. Şekil 3 a.1'e bakıldığında eşit madde dağılımlı durumda, yetenek parametrelerine ilişkin hata varyansı arttıkça Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinin düzenli bir dağılım göstermediği görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça RMSD değerlerinin artması beklenirken boyutla ilişkin RMSD değerlerinin azaldığı görülmektedir. Dolayısıyla hem Bayesyen hem de Fisher'in puanlama yetenek kestirimi yöntemlerinin tutarlı sonuçlar vermediği söylenebilir.

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin RMSD değerleri ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 2 a.2'ye bakıldığında, kestirilen yetenek parametrelerine ilişkin hata varyansı arttıkça Fisher'in puanlama yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinde artma ya da azalma olmadığı görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirim yönteminde hata varyansı arttıkça boyutlara ilişkin RMSD değerlerinin düzenli olarak artma eğilimi gösterdiği görülmektedir. Ayrıca içerik ağırlıklandırması uygulandığında, Bayesyen yetenek kestirim yöntemine ait RMSD değerlerinin daha düşük olduğu görülmektedir. Dolayısıyla, çok boyutlu BOB testlerinde içerik ağırlıklandırması yapıldığında hem Bayesyen MAP hem de Fisher'in puanlama yönteminin daha güvenilir ve tutarlı sonuçlar verdiği yorumu yapılabilir.

Şekil 3'te yer alan b.1 ve b.2 grafikleri içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ait her bir boyuta ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansları arasındaki ilişkiyi vermektedir. Şekil 3 b.1'e bakıldığında yetenek parametrelerine ait hata varyansları arttıkça, Bayesyen MAP yetenek kestirimine yöntemine ait her bir boyuta ilişkin güvenirlik katsayılarının düzenli bir şekilde azaldığı görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça boyutlara ilişkin güvenirlik katsayılarında anlamlı bir değişim gözlenmemektedir

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 3 b.2'ye bakıldığında, her bir boyuta ilişkin hata varyansı arttıkça Bayesyen MAP yetenek kestirim yöntemine ait güvenirlik katsayılarının azaldığı, Fisher'in puanlama yöntemine ait güvenirlik katsayılarının ise değişmediği görülmektedir. Bunun temel sebebi, Fisher'in puanlama yöntemi için hata varyansı durdurma kuralı 0,20'den 0,30 çıkmasına rağmen testteki ortalama madde sayısı değişmemesidir. Dolayısıyla, D-Optimality madde seçim yöntemi için içerik ağırlıklandırması uygulandığında Bayesyen MAP yetenek kestirim yöntemi daha az madde ile daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

Şekil 3'de yer alan c.1 ve c.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin her bir boyuta ait ölçmenin standart hatası ile hata varyansı arasındaki ilişkiyi vermektedir. Şekil 3 c.1'e bakıldığında, Bayesian MAP yetenek kestirimi yöntemi kullanıldığında, yetenek parametrelerine ilişkin hata varyansı arttıkça, her bir boyuta ilişkin ölçmenin standart hatasının da arttığı görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirimi yöntemi yerine Fisher'in puanlama yöntemi kullanıldığında, boyutlara ilişkin hata varyansı arttıkça boyutlara ilişkin standart hatanın değişmediği görülmektedir.

Şekil 3 c.2'ye bakıldığında, içerik ağırlıklandırmasının yapılmadığı durum ile benzer sonuçlar verdiği görülmektedir. Ayrıca içerik ağırlıklandırmasının yapıldığı durumda, Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin standart hatanın Fisher'in puanlama yöntemine ait hata varyansı değerlerinden daha düşük olduğu görülmektedir.

Genel olarak hata varyansı durdurma kuralının kullanıldığı maddeler-arası çok boyutlu BOB testi analiz yöntemlerine ilişkin bulgular karşılaştırıldığında, Bayesyen MAP yetenek kestirim yöntemine

ait en uygun hata varyansı durdurma kuralının 0,25 olduğu görülmektedir. Ayrıca, D-Optimality madde seçim yöntemi ve içerik ağırlıklandırmasının uygulandığı çok-boyutlu BOB testi uygulamalarında Bayesyen MAP yetenek kestirim yönteminin daha tutarlı sonuçlar verdiği görülmektedir. Diğer taraftan, Fisher'in puanlama madde seçim yöntemlerinden etkilendiği ve D-Optimality madde seçim yöntemi kullanıldığında testteki ortalama madde sayısı 50 olduğu durumda bile güvenilir ve turtarlı sonuçlar vermediği söylenebilir.

Tablo 3'te madde seçim yöntemlerinden seçkisiz (Random) madde seçim yöntemi, durdurma kurallarından hata varyansı, yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait analiz bulgularına yer verilmiştir. Ayrıca her bir koşul için içerik ağırlıklandırmasının kullanıldığı ve kullanılmadığı durumlara ait sonuçlar karşılaştırılarak, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelenmiştir.

Tablo 3. Seçkisiz (Random) madde seçim yöntemine ait Çok boyutlu BOB testi Bulguları

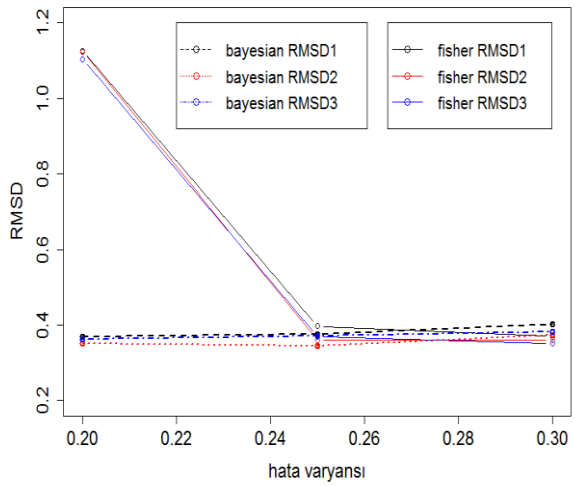
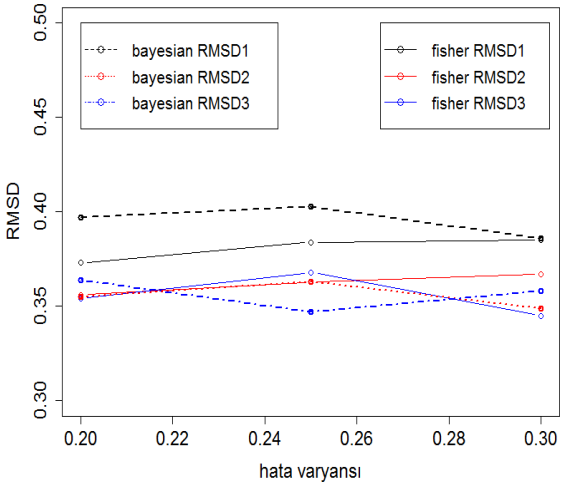
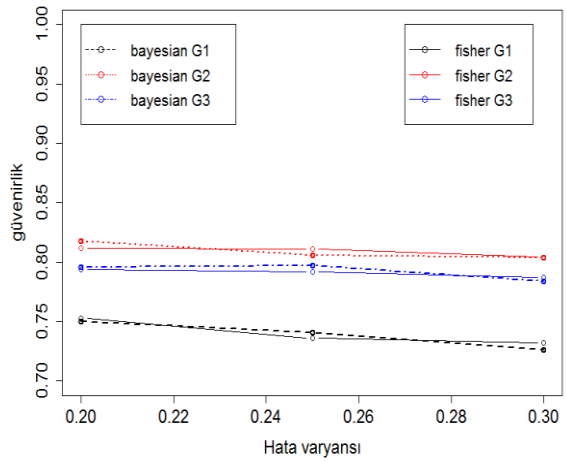
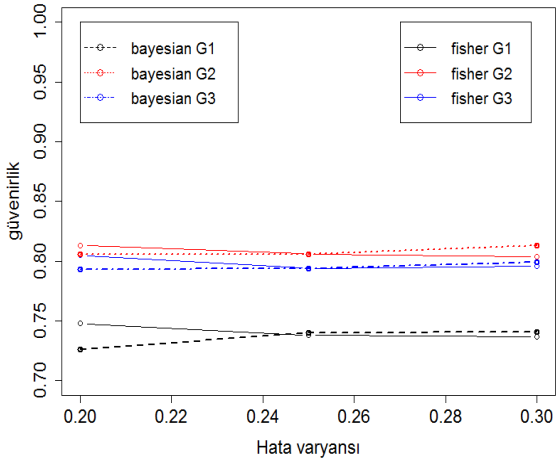
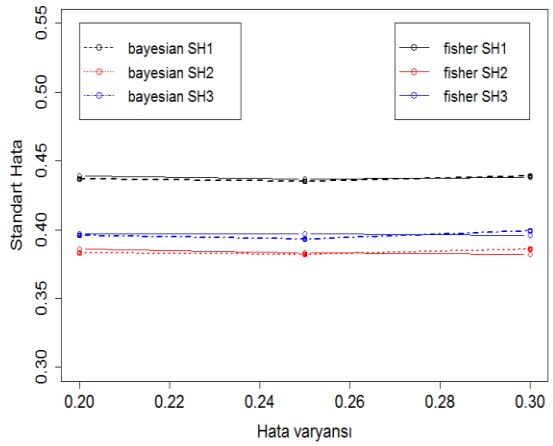
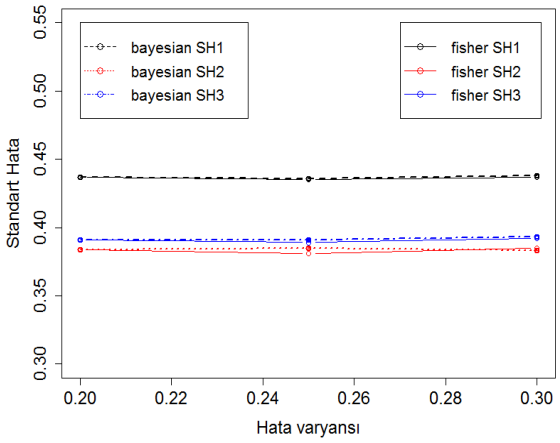
	Yetenek Kestirim yöntemi	Test Sonlandırma kuralı	Test uzunluğu	Güvenirlik			ÖSH			RMSD		
		S. Hata	K	boy1	boy2	boy3	boy1	boy2	boy3	boy1	boy2	boy3
Eşit madde dağılımlı	Fisher	0,20	50,00	0,75	0,81	0,79	0,439	0,386	0,397	1,126	1,124	1,104
		0,25	49,99	0,74	0,81	0,79	0,437	0,383	0,397	0,397	0,360	0,369
		0,30	48,69	0,73	0,80	0,79	0,438	0,382	0,396	0,370	0,361	0,352
	Bayesian	0,20	50,00	0,75	0,82	0,80	0,437	0,383	0,396	0,370	0,352	0,362
		0,25	50,00	0,74	0,81	0,80	0,435	0,382	0,393	0,376	0,344	0,371
		0,30	48,79	0,73	0,80	0,78	0,439	0,386	0,399	0,403	0,373	0,382
İçerik Ağırlıklandırmalı	Fisher	**0,20**	50,00	0,75	0,813	0,805	0,437	0,384	0,391	0,373	0,356	0,354
		0,25	49,97	0,74	0,806	0,794	0,435	0,381	0,389	0,384	0,363	0,368
		0,30	48,83	0,74	0,804	0,796	0,437	0,385	0,392	0,385	0,367	0,345
	Bayesian	0,20	50,00	0,73	0,806	0,793	0,437	0,384	0,391	0,397	0,355	0,364
		0,25	49,98	0,74	0,806	0,794	0,436	0,385	0,391	0,403	0,363	0,347
		0,30	48,30	0,74	0,813	0,799	0,438	0,383	0,393	0,386	0,349	0,358

*OSH= Ölçmenin standart Hatası

Tablo 3'teki analiz bulgularına bakıldığında, çok boyutlu BOB testi sürecinde maddeler her hangi bir kurala bağlı kalmadan seçkisiz (random) olarak seçildiğinde her bir yetenek kestirim yöntemi için hata varyansı durdurma kuralı 0,20'den 0,30'a çıkarıldığında testteki ortalama madde sayısı 48 ile 50 arasında değişkenlik gösterdiği görülmektedir. Ayrıca, diğer madde seçim yöntemleri ile karşılaştırıldığında her bir koşul için güvenirlik katsayılarının düşük, standart hata ve RMSD değerlerinin ise yüksek olduğu görülmektedir. Diğer taraftan, içerik ağırlıklandırmasının yapılmadığı ve yapıldığı duruma ait her bir yetenek kestirimin yönteminin benzer sonuçlar verdiği görülmektedir.

Şekil 4'te içerik ağırlıklandırmasının yapılmadığı eşit madde dağılımlı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin madde seçim yöntemlerinden seçkisiz madde seçim yöntemi, testi durdurma kurallarından boyutlara ilişkin hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerinin kullanıldığı çok-boyutlu BOB testi analizlerine ait grafikler verilmiştir

Şekil 4'te yer alan seçkisiz madde seçim yönteminin kullanıldığı çok boyutlu BOB testlerine ait RMSD değerlerini veren a.1 ve a.2 grafiklerine bakıldığında, Bayesyen MAP yetenek kestirim yöntemi için hata varyansı artmasına rağmen RMSD değerlerinde önemli bir artış gözlenmemektedir. Diğer taraftan, Fisher'in puanlama yetenek kestirim yöntemi kullanıldığında ise boyutlara ilişkin RMSD değerlerinin önce azaldığı daha sonra ise değişmediği görülmektedir.

Eşit madde dağılımlı	*İçerik Ağırlıklandırmalı*
a.1 RMSD- Hata varyansı ilişkisi	a.2 RMSD- Hata varyansı ilişkisi
	
b.1 Güvenirlik- Hata varyansı ilişkisi	b.2 Güvenirlik- Hata varyansı ilişkisi
	
c.1 Ölçmenin standart hatası- Hata varyansı	c.2 Ölçmenin standart hatası- Hata varyansı
	

Şekil 4. Seçkisiz (Random) Madde Seçim Yöntemine İlişkin Grafikler

İçerik ağrılıklandırması uygulandığında ise Fisher'in puanlama yetenek kestirim yöntemlerine ait RMSD değerlerinin daha düşük olduğu görülmektedir. Bayesyen yöntemi için ise hata varyansı arttıkça RMSD değerlerinin azalıp arttığı görülmektedir. Genel olarak, hem Bayesyen MAP hem de Fisher'in puanlama yöntemine ait her bir boyuta ilişkin RMSD değerlerini yüksek olduğu görülmektedir.

Şekil 4'te yer alan seçkisiz madde seçim yönteminin kullanıldığı çok boyutlu BOB testlerine ait her bir boyuta ilişkin güvenirlik katsayıları ile hata varyansı arasındaki ilişkiyi veren b.1 ve b.2 grafiklerine bakıldığında, hata varyansı arttıkça her bir yetenek kestirim yöntemine ait güvenirlik katsayılarının değişmediği ve düşük olduğu görülmektedir. İçerik ağırlıklandırmasının yapıldığı durumda da benzer sonuçlar elde edilmiştir. Bunun temel sebebi, boyutlara ilişkin hata varyansı artmasına karşın testteki ortalama madde sayısının değişmemesidir.

Son olarak, Şekil 4'te çok boyutlu BOB testlerine ait her bir boyuta ilişkin standart hata ile hata varyansı arasındaki ilişkiyi veren c.1 ve c.2 grafiklerine bakıldığında, hem Bayesyen MAP hem de Fisher'in puanlama yönteminin benzer sonuçlar verdiği görülmektedir. Ayrıca, her bir boyuta ilişkin hata varyansı artmasına rağmen standart hata değerlerinin değişmediği görülmektedir. Diğer taraftan, içerik ağırlıklandırması uygulandığında boyutlara ilişkin standart hata değerlerinde her hangi bir iyileşme olmadığı ve eşit madde dağılımı ile benzer sonuçlar verdiği görülmektedir. Bu bulgular doğrultusunda, bireyin yeteneğinin çok boyutlu BOB testleri ile ölçüldüğü durumda, maddelerin seçkisiz olarak seçilmesi yerine, Fisher'in bilgi matrisine dayalı A-Optimality ve D-Optimality madde seçim yönteminin kullanılması daha az madde ile daha yüksek güvenilirlikte ölçümler yapılmasına olanak sağlar.

SONUÇLAR ve TARTIŞMA

Bu çalışmada içerik ağırlıklandırmasının maddeler-arası boyutluluk modeline dayalı çok boyutlu BOB testleri üzerindeki etkisini incelemek amacıyla, farklı madde seçim yöntemleri, yetenek kestirim yöntemleri ve durdurma kuralının kullanıldığı çok boyutlu MTK modellerinden maddeler-arası boyutluluk modellerine dayalı çok-boyutlu bilgisayar ortamında bireyselleştirilmiş (BOB) test (MCAT) yöntemlerinin performansları karşılaştırılmıştır.

Genel olarak, maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi analiz sonuçları karşılaştırıldığında, Fisher'in puanlama yetenek kestirim yöntemi için A-Optimality madde seçim yöntemlerine ait güvenirlik katsayılarının daha yüksek, RMSD ve standart hata değerlerinin ise daha düşük olduğu görülmektedir. Diğer taraftan, Bayesyen MAP yetenek kestirim yönteminin hem A-Optimality hem de D-Optimality madde seçim yöntemi için benzer sonuçlar verdiği görülmektedir. Dolayısıyla, Fisher'in puanlama yönteminin madde seçim yöntemlerinden etkilendiği sonucu çıkartılabilir.

İçerik ağırlıklandırması uygulandığında ise hem A-Optimality hem de D-Optimality madde seçim yöntemi için Bayesyen MAP yetenek kestirim yöntemine ait testteki ortalama madde sayısı az da olsa artarken, boyutlara ilişkin güvenirlik katsayılarının ise değişmediği görülmektedir. Buna karşın boyutlara ilişkin RMSD ve standart hata değerinin düştüğü görülmektedir. Bunun temel sebebi, kâğıt-kalem formatındaki testte her bir boyuta ait maddelerin oranlarının içerik ağırlıklandırması yöntemiyle korunarak, her bir boyut için seçilecek maddelerin ve madde sayılarının sınırlandırılmasıdır. Diğer taraftan, A-Optimality madde seçim yöntemi kullanıldığı çok boyutlu BOB testi yöntemlerinde içerik ağırlıklandırması uygulanmadığında Bayesyen MAP ve Fisher'in puanlama yöntemi farklı sonuçlar verirken, içerik ağırlıklandırması uygulandığında ise benzer ve tutarlı sonuçlar verdiği söylenebilir.

Çok boyutlu BOB testlerinde madde seçim yöntemlerinden hangisinin kullanılması gerektiğini testin amacı belirler. Eğer testin ölçtüğü bütün boyutlar ölçülmek istendiğinde, A-Optimality ve D-Optimality madde seçim yöntemleri en iyi sonucu verir (Mulder ve van der Linden, 2009; Lin, 2012). Dolayısıyla, yetenek kestirim yöntemlerinden Bayesyen MAP yöntemi kullanıldığında, madde seçim yöntemi olarak hem A-Optimality hem de D-Optimality madde seçim yöntemi benzer

sonuçlar vereceğinden tercih edilebilir. Eğer yetenek kestirim yöntemlerinden Fisher'in puanlama yöntemi kullanılacaksa, madde seçim yöntemlerinden A-Optimality yönteminin içerik ağırlıklandırması uygulanarak kullanılması önerilmektedir.

Diao (2009) yapmış olduğu çalışmada, yetenek kestirim yöntemi olarak MLE yönteminin ve madde seçim yöntemlerinden A-Optimality ve D-Optimality yöntemlerinin kullandığında testteki madde sayısını 50 olduğunda A-Optimality ve D-Optimality madde seçim yöntemlerine ait RMSE ve ortalama yanlılık değerlerinin birbirine çok yakın olduğunu belirtmektedir. Ayrıca testteki madde sayısı 50 olduğunda her iki yöntemin benzer sonuçlar verdiğini belirtmektedir. Nitekim bu çalışmada da testteki madde sayısı arttıkça her iki madde seçim yönteminin benzer sonuçlar verdiği bulgusuna ulaşılmıştır. Ayrıca bireyin yeteneğinin çok boyutlu BOB testleri ile ölçüldüğü durumda, maddelerin seçkisiz olarak seçilmesi yerine, Fisher'in bilgi matrisine dayalı A-Optimality ve D-Optimality madde seçim yönteminin kullanılması daha az madde ile daha yüksek güvenilirlikte ölçümler yapılmasına olanak sağlar.

Çok-boyutlu BOB testlerine ilişkin yapılan çalışmalara bakıldığında, A-optimality ve D-optimality madde seçim yöntemlerinin karşılaştırıldığı birçok çalışma yapılmıştır (Segall, (1996); Luecht, (1996); van der Linden,1999; Mulder ve van der Linden, 2009; Diao, 2009; Diao ve Reckase; 2009; Yoo, 2011; Lin, 2012). Genel olarak D-optimality yönteminin daha avantajlı olduğu ve daha yaygın olarak kullanıldığı belirtmektedir (Berger ve Veerkamp, 1996; Passo, 2007). Bunun temel sebebi olarak D-optimality madde seçim yönteminin daha güvenilir ve daha kararlı sonuçlar verdiği belirtilmektedir. Ancak bu yapılan çalışmalara bakıldığında sadece madde-içi boyutluluk modelinin kullanıldığı görülmektedir. Bu çalışmada ise maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testlerinde A-optimality ve D-optimality madde seçim yöntemlerinin Bayesyen MAP yetenek kestirim yöntemi kullanıldığında benzer sonuçlar verdiği bulgusuna ulaşılmıştır.

Genel olarak, maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi analiz sonuçları karşılaştırıldığında, aynı madde seçim yöntemi ve durdurma kuralı koşulları altında Bayesyen MAP yöntemine ait güvenirlik ve kestirilen yetenek parametrelerine ilişkin korelasyon değerlerinin daha yüksek, RMSD ve standart hata değerlerinin ise daha düşük olduğu görülmektedir. Dolayısıyla her bir madde seçim ve durdurma kuralı için maddeler-arası boyutluluk modelinde Bayesyen MAP yetenek kestirim yönteminin daha az madde ile daha güvenilir sonuçlar verdiği yorumu yapılabilir. Bu bulgular doğrultusunda, maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi yöntemlerinde her bir madde seçim yöntemi için en uygun yetenek kestirim yönteminin Bayesyen MAP yetenek kestirim yöntemi olduğu söylenebilir

Bayesyen yetenek kestirim yöntemlerini MLE yetenek kestirim yöntemlerine göre üstün kılan en önemli özelliği, Segall (1996)'in de belirttiği gibi, Bayesyen yöntemlerin boyutlar arasındaki korelasyon ve yetenek parametrelerine ait önsel dağılım bilgisini kullanarak kestirim yapmasıdır (Diao ve Reckase, 2009). Bundan dolayı Bayesyen yöntemleri gerçek θ değerine daha çabuk yakınsar ve daha az madde ile daha güvenirlik ve kararlı kestirimler yapar. Buna karşın, kestirilen yetenek parametreleri hakkında yeterli bilgi olmadığı veya belirlenen önsel parametreler zayıf olduğu durumda, Bayesyen yöntemler ile kestirilen yetenek parametreleri yanlılık gösterir.

Hata varyansı durdurma kuralına ilişkin maddeler-arası boyutluluk modeline dayalı A-optimality madde seçim yöntemi ve Bayesyen MAP yetenek kestirim yönteminin kullanıldığı çok-boyutlu BOB testlerinde en uygun hata varyansı durdurma kuralının 0,25 olduğu görülmektedir. Bu durumda maddeler-arası boyutluluk modeli için testi alan her bir bireyin test sürecinde cevaplamış olduğu ortalama madde sayısı 22,8'e eşit olurken, içerik ağırlıklandırması yapıldığında testteki ortalama madde sayısı 29,6'ya yükseldiği görülmektedir.

Hata varyansı durdurma kuralı daha güvenilir ve etkili olmasına karşın bireylerin test sürecinde farklı sayıda maddelere cevap vermesi testin adil olmadığı algısını oluşturabilir (Gershon, 2005). Bu yüzden eğitimdeki gerçek BOB testi uygulamalarında genellikle sabit madde sayısı durdurma kuralı kullanılır (Yoo, 2011). Sabit madde sayısı durdurma kuralının hata varyansı durdurma kuralına tercih edilmesinin bir diğer nedeni ise test süresi ve testten sıkılma gibi koşulların her bir birey için standartlaştırılmasına olanak sağlamasıdır (Segall, 2004). Buna karşın, Rizavi ve Swaminathan

(2001) sabit madde sayısı durdurma kuralının kullanıldığı durumlarda testteki madde sayısı bireyin yeteneğini güvenilir bir şeklide ölçmek için yeterli olmadığında problemlere neden olacağını ve testin güvenirlik ve geçerliğini düşüreceğini savunmuştur.

Maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi yöntemlerinin kullanıldığı durumda madde seçim yöntemlerinden ve madde düzeyindeki boyutluluk modellerinden daha az etkilendiği için Bayesyen MAP yetenek kestirim yönteminin kullanılması önerilmektedir. Ayrıca, çok-boyutlu BOB testi yöntemleri için testteki madde sayısının az olduğu ve boyutlar arasındaki korelasyonun yüksek olduğu durumlarda Bayesyen MAP yetenek kestirim yönteminin kullanılması önerilmektedir.

Sonuç olarak, kâğıt-kalem testleri ile karşılaştırıldığında hem maddeler-arası modeline dayalı çok-boyutlu BOB testlerinin daha az madde ile daha yüksek güvenirlikte ölçümler yaptığından benzer formattaki testlerin çok-boyutlu bilgisayar ortamında bireyselleştirilmiş testler ile uygulanması önerilmektedir.

Bu çalışmada, gerçek veriye dayalı simülasyon yöntemi (post-hoc simulation method) kullanılarak, testin üç boyutlu olduğu durumda farklı yetenek kestirimi, madde seçim ve durdurma kurallarına ilişkin çok-boyutlu BOB testi sonuçları karşılaştırılmıştır. Farklı simülasyon çalışmaları yapılarak, testin ölçtüğü boyut sayısının, boyutlara ilişkin madde havuzu büyüklüğünün ve farklı çok boyutlu modellerin çok-boyutlu BOB testi yöntemleri üzerindeki etkisi incelenebilir.

Bu çalışmada, çok-boyutlu BOB testi sürecinde her bir boyut için sorulacak madde sayısını ve testin formatını kontrol altında tutmak için kâğıt-kalem testi formatındaki her bir boyuta ait madde oranına bağlı olarak içerik ağırlıklandırması (content balancing) yapılmıştır. Gelecekte yapılacak çalışmalarda, daha önce geliştirilmiş olan içerik ağırlıklandırması yöntemleri (örn. Düzgünleştirilmiş-alfa deseni, Chang, Qian ve Ying, 2001; Sympson ve Hetter's yöntemi, Sympson ve Hetter 1985) kullanılarak BOB testi süreci daha gerçekçi hale getirilebilir. Ayrıca, çok-boyutlu BOB testlerinde kullanılan madde kullanım sıklığı ve içerik ağırlıklandırması yöntemlerinin birbirini nasıl etkilediği üzerinde çalışmalar yapılabilir.

Ülkemizde tek boyutlu BOB testi uygulamalarına yönelik çalışmalar olmasına karşın, çok-boyutlu BOB testlerinin gerçek hayatta uygulamalarına ilişkin bir çalışma henüz yapılmamıştır. Bu çalışma sonuçları ışığında gerçek çok-boyutlu BOB testi uygulamaları yapılması önerilmektedir. Ayrıca bu çalışmada bireylerin sadece İngilizce dil becerilerinin çok-boyutlu BOB testleri ile ölçülmesi amaçlanmıştır. Matematik ve fen bilgisi gibi alanlarda da benzer çalışmalar yapılabilir.

KAYNAKÇA

Berger, M.P. F., & Veerkamp, W. J. J. (1996). A review of selection methods for optimal tests design. In M. Wilson & G. Engelhard (Eds.), Objective measurement: Theory into practice (Vol. 3, pp. 437-455). Norwood, NJ: Ablex

Bloxom, B., & Vale, C.D. (1987).Multidimensional adaptive testing: An approximate procedure for updating. In *Meeting of the psychometric society*. Montreal, Canada, June.

Chang, H.-H., Qian, J. and Ying, Z. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement,* 25, 333-341.

Choi, S. W. & King D. R. (2011). *MAT: Multidimensional adaptive testing.* [Çevirim içi: https://cran.r-project.org/web/packages/MAT/MAT.pdf], Erişim tarihi: 15 Temmuz 2015.

Diao, Q. (2009). *Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing*. Unpublished Doctoral Dissertation. Michigan State University.

Diao, Q. & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In: Weiss DJ (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. pp. 1-13.

Fan, M., & Hsu, Y. (1996). Multidimensional computer adaptive testing. *In Annual meeting of the American educational research association*. New York City, NY, April.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement* 6:109-27.

Green, B. G., Bock, R.D., Humphries, L. G., Linn, R.L., & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360

IACAT Official Web Site. [Çevirim-içi: http://iacat.org/content/research-strategies-cat]. Erişim tarihi: 24 Aralık 2015.

Lin, H. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidimensional generalized partial credit model*. Unpublisjed Doctoral Dissertation. University of Illinois at Urbana-Champaign.

Lord, F. M. (1971a). Tailored testing, an approximation of stochastic approximation. Journal of the American Statistical Association, 66, 707–711.

Lord, F. M. (1971b). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31,* 805–813.

Lord, F. M. (1971c). A theoretical study of two-stage testing. *Psychometrika, 36,* 227–242.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20* (4), 389-404.

McBride, J.R. & Martin, J.T. (1983). Reliability and Validity of Adaptive Ability Tests in a military setting. in Weiss D.J. (Ed.) *"New Horizons in Testing"* New York: Academic Press.

Mulder, J., & van der Linden, W. J. (2009). Muldimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74* (2), 273-296.

Passos, V. L., Berger, M. P. F., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. Applied Psychological Measurement, 31, 213-232.

Rizavi, S. & Swaminathan, H. (2001). The effect of test and examinee characteristics on the occurrence of aberrant response patterns in a computerized adaptive test. *Paper presented at the annual meeting of the American Educational Research Association*, Seattle WA. (2001)

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

Segall, D.O. (2000). Principles of multidimensional adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–73). Boston: Kluwer Academic.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66, 79-97.

Silvey, S.D. (1980). *Optimal design*. London: Chapman & Hall.

Sympson, J.B. and Hetter, R.D. (1985, october). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977).* San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W.J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20,* 373–388.

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

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

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

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., et al. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Wang, W., Chen, P., & Cheng, Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116–136.

Wang, W.-C., Wilson, M., and Adams, R. (1997). Rasch models for multidimensionality between items and within items. In G. Englehard, Wilson, Mark (Ed.), *Objective Measurement* (Vol. 4,): Greenwich, CN: Ablex Publishing.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.*

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

Yoo, H. (2011). *Evaluating several multidimensional adaptive testing procedures for diagnostic assessment*. Unpublished Doctoral Dissertation. University of Massachusetts Amherst)

EXTENDED ABSTRACT

Developments in computer technologies not only affect our social life, environment and our life styles but also affect our education systems, measurement and evaluation tools and learning methods. These developments also provide new methods to measure students' different abilities. Along with these developments, stakeholders in education benefit from computer technologies so as to provide education in higher standards. As a result of these developments, new measurement and evaluation methods that utilize computer technologies has been developed to measure students' abilities or skills. An example of this is computer-based tests (CBT) that uses computer environment instead of paper-pencil tests. Another alternative measurement method is computerized adaptive testing (CAT) methods in which students abilities are tailored to items' features by means of a computer program (McBride and Martin, 1983; Weiss and Kingsbury, 1984).

There are a lot of advantages of measuring students' abilities with CAT methods compared to paper&pencil tests. One of the most important advantage of CAT is that it enable us to measure students' abilities with shorter tests and higher reliabilities (Wainer, 1993). İn addition, it provides flexible testing time and provides result of test to examinees as soon as test is terminated (Lin, 2012).

Purpose of study

The purpose of this study is to compare the performance of Between-item dimensionality model-based Multidimensional CAT designs and to examine the effect of content balancing on different MCAT designs.

Methods

For this purpose, real data set from English Proficiency Test (EPT) administered by Hacettepe University was used to create multidimensional item pool, in which each test consist of three dimensions listening, reading and grammar respectively. In this study, 10 EPT data sets, which consist of 628 items in total, administered between 2009 and 2013 were used to construct item pool. Number of items in each test ranged from 59 to 64 and administered to students ranged between 1200 and 2000. Item parameters were estimated with compensatory multidimensional 2 parameter logistic model (CM-2PLM) based on between-item dimensionality model. After item calibration, 73 items with low discrimination parameters and items which had difficulty parameters out of the [-4, 4] interval were excluded from the item pool. Finally, multidimensional item pool consist of 555 items of which 240 items are related to grammar, 115 items are related to listening and 200 items are related to reading dimension. For each MCAT algorithm, examinee ability parameters (θ) were drawn from the multivariate standard normal distribution with the population variance-covariance matrix and the number of examinee was restricted to 500.

In order to determine the best MCAT algorithm for the EPT, two different theta estimation methods (MLE based Fisher scoring and Bayesian MAP), three different item selection methods (fisher information based A-optimality and D-optimality, Random item selection) and two different termination methods (fixed number of item, precision) were used. In addition, results of MCAT algorithms with content distribution and without content distribution were compared. Results of these conditions were compared with respect to, reliability index, RMSE, averaged number of items administered and RMSD values between full bank theta and estimated MCAT theta.

Results and Discussion

Results indicated that using different theta estimation and item selection methods affected RMSE, averaged number of items administered and RMSD values for each MCAT algorithm. When Bayesian MAP ability estimation method was used both A-optimality and D-optimality yielded similar results with respect to reliability coefficients, SEM and RMSD values. On the other hand, A-

optimality item selection method outperformed both D-optimality and non-adaptive random item selection methods when MLE based Fisher's scoring ability estimation method was used. In multidimensional case, there are several studies investigating A- and D-optimality for dichotomous MIRT models. Segall (1996), Luecht (1996), and Mulder and van der Linden (2009) used the D-optimality. van der Linden (1999) studied A-optimality. As Mulder and van der Linden (2009) stated, A-optimality and D-optimality yield the most accurate estimates in which all measured abilities were intentional.

Using A-optimality rather than D-optimality to select items both decreased average number of items administered and RMSD values between true theta and estimated theta, and increased test reliability. Overall, MCAT design with A-optimality and Bayesian theta estimation method outperformed other MCAT designs.

The comparison of MCAT with content balancing and without content balancing showed that content balancing yielded more accurate and consistent results. In addition, using content balancing yielded higher reliability coefficients with shorter test. For each MCAT condition SEM and RMSD statistics associated with each dimension tended to decrease when content distribution was taken into account. Therefore, there was a trade-off between reliability coefficients and error statistics associated with each dimension. All in all, post-hoc simulation based on MCAT with content balancing for EPT provided ability estimations with higher reliability with fewer items compared to paper and pencil format. Results of this study would also provide an important guideline for live MCAT application of EPT.