ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2016; 7(2); 285-295

Nitelik Sayısının Madde Seçme Algoritmalarının Performansı Üzerindeki Etkisi

The Impact of Attribute Size on the Performance of Item Selection Algorithms

Mehmet KAPLAN*

Öz
Bilişsel tanı modelleri (BTM) ve bilgisayar ortamında bireye uyarlanmış test (BOBUT) uygulamaları hem eğitim alanında hem de psikoloji gibi diğer alanlarda hızla artmaktadır. Günümüze kadar BOBUT uygulamaları üzerine yapılan araştırmaların çoğunluğu, genellikle özetleyici bir değerlendirme sağlayan madde tepki kuramına dayalı modeller ile gerçekleştirilmiştir. Oysa BTM gibi bireylerin uzmanlaştığı ya da yetersiz kaldığı konularda daha ayrıntılı sonuç sağlayan biçimlendirici değerlendirme teknikleri son zamanlarda giderek önem kazanmıştır. BTM'nin BOBUT (BiTBOBUT) uygulamalarındaki kullanımı bireylerin yetenek düzeyleri hakkında daha ayrıntılı ve etkin bir değerlendirme sağlamada etkili bir yöntemdir. Bu çalışmanın amacı, BiTBOBUT uygulamalarındaki nitelik sayısının madde seçme algoritmalarının performansları üzerindeki etkisini araştırmaktır. Simülasyon çalışmasında farklı nitelik sayısı ve BTM'nin, madde seçme algoritmalarının performansları üzerindeki etkisi, ortalama test uzunlukları gibi betimsel istatistik değerlerine bakılarak araştırılmıştır. Elde edilen sonuçlara göre farklı nitelik sayısının ortalama test uzunlukları üzerinde önemli değişmelere sebep olduğu fakat farklı BTM kullanımının algoritmalar üzerinde test uzunluklarına herhangi bir etkisinin olmadığı gözlemlenmiştir.

Anahtar Kelimeler: bilişsel tanı modelleri, bilgisayar ortamında bireyselleştirilmiş testler, nitelik sayısı.

Abstract
The use of cognitive diagnosis models (CDMs) and computerized adaptive testing (CAT) has been increasing in both education and other fields such as psychology. To date, most of the research in CAT has been done using item response theory models which provide summative scores. However, formative assessment techniques (e.g., CDMs) that provide more detailed information about individuals' strengths and weaknesses have become popular in the recent years. The use of cognitive diagnosis computerized adaptive testing (CD-CAT) can produce more diagnostic information with an efficient testing design. This paper aims to investigate the impact of attribute size on the performance of item selection indices in terms of average test lengths. The result of this study showed that increasing the attribute size resulted in longer average test lengths; however, using different CDMs did not change the average test lengths.

Keywords: cognitive diagnosis models, computerized adaptive testing, attribute size.

GİRİŞ

İzleme veya ünite testleri gibi öğrenme niteliğinin izlenmesinde kullanılan biçimlendirici değerlendirme tekniklerinin önemi son zamanlarda eğitim literatüründe giderek artmıştır. Biçimlendirici değerlendirme teknikleri, özetleyici değerlendirme tekniklerine göre özellikle öğretme ve öğrenme stratejilerini geliştirmek için bireylerin yetenek düzeyleri hakkında daha açıklayıcı bilgi sağlayabilir (DiBello & Stout, 2007). Örneğin dönem sonunda açıklanan karne notları; bireylerin hangi konuları iyi derecede öğrendiği, hangi konularda ise yetersiz kaldıkları

* Araş. Gör. Dr., Artvin Çoruh Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Artvin-Türkiye, e-posta: mehmetkaplan@artvin.edu.tr

Geliş Tarihi: 17.07.2016
DOI: 10.21031/epod.268486 *Kabul Tarihi: 13.10.2016*

hakkında detaylı bir bilgi sağlamayabilir. Oysa etkinlik sonunda biçimlendirici değerlendirme yöntemleriyle ilgili kazanımlara dayalı ve detaylı bir şekilde öğrencilere dönüt vermek, hem öğrenci hem de öğretmen için daha iyi öğrenme ve öğretme ile sonuçlanabilir. Bu bağlamda bilişsel tanı modelleri (BTM; *cognitive diagnosis models*) bireylerin yetenek düzeyleri hakkında daha ayrıntılı bir değerlendirme sağlayabilmek için geliştirilmiş psikometrik modellerdir (de la Torre, 2009). BTM'nin en önemli amacı, belirli bir alanda bireylerin uzman ve yetersiz oldukları *noktaları*, *özellikleri* ya da *nitelikleri* tespit etmektir. Bu çalışmada *nitelik* kelimesi terim olarak bireyin uzman ve yetersiz olduğu *nokta* ya da *özellik* için kullanılmıştır. Türkiye'de de BTM ile ilgili çalışmalar gün geçtikçe literatürde yerini almaya başlamaktadır (Başokçu, 2014).

Gerek eğitim alanında gerekse psikoloji gibi diğer alanlarda kullanımı giderek artan diğer bir uygulama ise bilgisayar ortamında bireye uyarlanmış test (BOBUT; *computerized adaptive testing*) uygulamalarıdır. BOBUT uygulamaları, kağıt ve kalem testlere alternatif olarak geliştirilmiş olup daha kısa test uzunluğu ile birlikte daha iyi yetenek kestirimine olanak sağlamaktadır (Meijer & Nering, 1999; van der Linden & Glas, 2000). Her ne kadar madde tepki kuramı (MTK; *item response theory*) modelleri kullanılarak gerçekleştirilen BOBUT uygulamaları, bireylerin yetenek seviyeleri hakkında özetleyici değerlendirme sağlayarak mevcut test sisteminin ihtiyaçlarını karşılamış gibi görünse de BOBUT uygulamalarının BTM ile birlikte kullanımıyla bireylerin bilişsel seviyeleri hakkında daha etkin bir tanısal değerlendirme elde edilebilir. Son zamanlarda bilişsel tanı modelleri kullanılarak bilgisayar ortamında bireye uyarlanmış test (BiTBOBUT) uygulamalarının kullanımı biçimlendirici değerlendirme çalışmalarında hızla artmaktadır.

BTM

Daha önce de değinildiği gibi BTM'nin amacı bireyleri sıralama amaçlı bir puanlama yapmak yerine, bir konuyu öğrenmek için sahip olunması gereken niteliklerde bireylerin uzmanlaşması ya da yetersiz kalmasına göre sınıflandırma yapmaktır (DiBello & Stout, 2007). Bunun için her bir bireye nitelik vektörü olarak adlandırılan çok boyutlu $\boldsymbol{\alpha}$ tayin edilir. Bu vektör genellikle 0 ve 1 olmak üzere iki değerli değişkenlerden oluşur. Bireyin nitelik vektörü her bir nitelik $k = 1, 2, \ldots, K$ için $\boldsymbol{\alpha} = \{\alpha_k\} = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ şeklinde ifade edilir. Eğer birey k sıradaki nitelik için uzmanlaşmışsa $\alpha_k = 1$, uzmanlaşmamış ve yetersiz kalmışsa, $\alpha_k = 0$ şeklinde gösterilir. Örneğin beş niteliğin bulunduğu bir konu alanında, eğer birey birinci ve üçüncü niteliklerde uzmanlaşıp diğerlerinde yetersiz kalmışsa bireyin nitelik vektörü $\boldsymbol{\alpha} = (1,0,1,0,0)$ şeklinde gösterilir. Ayrıca BTM'de her bir madde için hangi niteliğin kullanılıp kullanılmayacağı Q-matris yardımıyla belirlenebilir (Tatsuoka, 1983). Bireylere tayin edilen nitelik vektörü gibi maddenin doğru cevaplanması için hangi niteliği gerektirdiğine göre her bir maddeye q-vektörü tayin edilir. Bu vektörler her bir madde için birleştirildiğinde Q-matris elde edilmiş olur. Toplam madde sayısı (J) ve nitelik sayısının (K) boyutlarını oluşturduğu Q-matris genellikle iki değerli değişkenlerden oluşur. Eğer j sırasındaki madde k niteliğini gerektiriyorsa, Q-matrisin bu madde ve nitelik için olan elementi $q_{jk} = 1$ değerini alır. Aksi halde bu element sıfır değerini alır. Örneğin yine beş niteliğin bulunduğu bir konu alanında, j sıradaki madde sadece ikinci ve dördüncü nitelikleri gerektiriyorsa, bu maddeye ait q-vektörü $\boldsymbol{q_j} = (0,1,0,1,0)$ şeklinde gösterilir.

Günümüze kadar BTM'nin uygulanabilirliğini arttırmak için birçok model geliştirilmiştir. Bu modeller, üzerindeki varsayımlar düşünülerek *kısıtlayıcı (restricted)* ve *genel (general)* modeller olarak iki gruba ayrılabilir. Kısıtlayıcı modellere örnek olarak; *the deterministic inputs, noisy "and" gate* (DINA; de la Torre, 2009; Haertel, 1989; Junker & Sijtsma, 2001) modeli ve *the deterministic input, noisy "or" gate* (DINO; Templin & Henson, 2006) modeli verilebilir. Kısıtlayıcı modeller nitelik vektörleri ile görev performansı arasında güçlü varsayımlar gerektirir. Fakat ürettikleri sonuçlar genel modellere nazaran daha kolay ve anlaşılırdır. Örneğin DINA modelinde varsayım gereği bir maddenin doğru cevaplanabilmesi için bireyin o maddenin gerektirdiği tüm niteliklere sahip olması gerekmektedir. DINO modelinde ise bireyin maddenin gerektirdiği en az bir niteliğe sahip olması yeterlidir. Bu bağlamda DINA modeli telafi edici olmayan bir model iken DINO modeli telafi edici özellikte bir modeldir. Genel modellere örnek olarak ise *the log-linear CDM* (Henson, Templin, & Willse, 2009), *general diagnostic* (von Davier, 2008) modeli ve *the*

generalized DINA (G-DINA; de la Torre, 2011) modeli gösterilebilir. Kısıtlayıcı modellerin aksine genel modellerde güçlü varsayımlar yoktur. Genel modellerin en önemli özelliği, verinin modele uyumunun kısıtlayıcı modellere göre her zaman daha iyi sağlanmasıdır. Ancak genel modeller daha çok parametreyle çalıştıkları için bu modellerden çıkan sonuçları yorumlamak kısıtlayıcı modellere göre daha zordur. Verinin modele uyumu sağlanması koşuluyla kısıtlayıcı modelleri tercih etmek, analiz sonuçlarının daha anlaşılır yorumlanması açısından faydalıdır.

BiTBOBUT

BOBUT uygulamalarının en önemli amacı, test ilerledikçe bireylerin yetenek seviyeleri hakkında daha fazla bilgi elde etmektir. BOBUT uygulamalarını bireye özgü yapan unsur ise maddelerin bir algoritma tarafından bireylerin yetenek seviyelerine göre uyarlanmış olarak uygulama esnasında seçilmesidir. Maddelerin bu şekilde seçilmesi, düşük yetenek düzeyindeki bireylere daha az sayıda zor madde sorulmasına ve aynı şekilde, yüksek yetenek düzeyindeki bireylere daha az sayıda kolay madde sorulmasına neden olur. Bu durum, hem test uzunluğunun kısalmasını hem de bireylerin yetenek kestirimlerinin daha doğru yapılmasını sağlar (Meijer & Nering, 1999).

BOBUT uygulamaları; madde havuzu, başlangıç noktası, madde seçme algoritması, puanlama ve testi sonlandırma kuralı olmak üzere beş ana bileşenden oluşur (Thissen & Mislevy, 2000). BOBUT uygulamaları genellikle MTK modelleri üzerine kurulmuştur ve bilindiği gibi bu modellerde kullanılan yetenek değişkeni θ süreklidir. Ancak BTM'de kullanılan nitelik vektörü $\boldsymbol{\alpha}$ sürekli olmayıp kategorik değişken olduğundan, BOBUT uygulamalarında bulunan bazı kavramlar BiTBOBUT uygulamalarında kullanılamaz. Örneğin BOBUT uygulamalarında, en yüksek Fisher bilgi fonksiyonu madde seçme algoritması olarak yaygın bir şekilde kullanılmakta iken BiTBOBUT uygulamalarında bu fonksiyon kullanılamaz çünkü Fisher bilgi fonksiyonu sadece sürekli değişkenlerde çalışabilmektedir. Bu soruna çözüm olarak BiTBOBUT uygulamalarında ilk olarak, hem sürekli hem de kategorik değişkenlerle çalışabilen *Kullback-Leibler* (K-L) *bilgi istatistiği* ve *Shannon entropi* yöntemi, kullanılmıştır (Xu, Chang, & Douglas, 2003). K-L bilgi istatistiği, *t* sırasındaki maddenin uygulanması sonucunda elde edilen nitelik vektörü kestiriminin, diğer nitelik vektörleri ile uzaklığıdır. Shannon entropi ise rastgele değişkenle ilişkilendirilmiş belirsizlik ölçüsü olarak tanımlanmıştır. Bu çalışmanın sonucunda her iki algoritmanın performansı karşılaştırıldığında, Shannon entropi yönteminin, K-L bilgi istatistiğine göre nitelik vektörlerinin sınıflandırmasında daha iyi sonuçlar sergilediği gösterilmiştir.

Daha sonra Cheng (2009), K-L bilgi istatistiğini geliştirerek BiTBOBUT uygulamalarında kullanılmak üzere yeni bir madde seçme algoritması önermiştir. Önerilen bu algoritmada bir önceki algoritmadan farklı olarak K-L bilgi istatistiği, nitelik vektörlerinin sonsal dağılımları ile ağırlıklandırılmıştır. Diğer bir ifadeyle bireyin nitelik vektörü kestiriminde, *t'*nci madde uygulandıktan sonra bazı nitelik vektörlerinin diğer vektörlere göre olma olasılığının daha fazla olmasından dolayı, vektörler arası uzaklığın sonsal dağılıma göre ağırlıklandırılması daha doğru sonuçlar doğuracaktır. Yapılan simülasyon çalışmasının sonucunda, *sonsal ağırlıklandırılmış Kullback-Leibler indeksi* (PWKL; *Posterior-Weighted Kullback-Leibler*), K-L bilgi istatistiği ve Shannon entropi yöntemine göre daha yüksek nitelik vektör sınıflandırılma değerleri vermiştir. PWKL algoritmasının matematiksel gösterimi Eşitlik 1'de yer almaktadır.

$$PWKL_j(\hat{\boldsymbol{\alpha}}_i^{(t)}) = \sum_{c=1}^{2^K}\left[\sum_{x=0}^{1}\log\left(\frac{P(X_j = x \mid \hat{\boldsymbol{\alpha}}_i^{(t)})}{P(X_j = x \mid \boldsymbol{\alpha}_c)}\right)P(X_j = x \mid \hat{\boldsymbol{\alpha}}_i^{(t)})\boldsymbol{\pi}_i^{(t)}(\boldsymbol{\alpha}_c)\right] \quad (1)$$

Eşitlik 1'de, $\hat{\boldsymbol{\alpha}}_i^{(t)}$, *t* sayıda madde uygulandıktan sonraki; *i* bireyinin *j* sırasındaki madde için nitelik vektörü kestirimi; $P(X_j = x \mid \boldsymbol{\alpha}_c)$, 2^K adet nitelik vektörünün herbirinin $(\boldsymbol{\alpha}_c)$ verilen yanıt örüntüsüne göre başarı olasılık değerleri ve $\pi_i^{(t)}(\boldsymbol{\alpha}_c)$; *t* sayıda madde uygulandıktan sonraki her bir 2^K nitelik vektörünün sonsal dağılım değerleridir.

BiTBOBUT uygulamalarında kullanılmak üzere iki yeni madde seçme algoritması daha yakın zamanda geliştirilmiştir (Kaplan, de la Torre, & Barrada, 2015). Bu algoritmalardan ilki K-L bilgi istatistiğine dayalı olan *değiştirilmiş sonsal ağırlıklandırılmış K-L* indeksi (MPWKL; *modified PWKL*), ikincisi ise *G-DINA modeli ayrım* indeksidir (GDI; *G-DINA model discrimination index*). Yapılan simülasyon çalışması sonucuna göre her iki madde seçme algoritması, benzer şekilde sonuç sergilemiş olup her ikisi de PWKL algoritmasından daha yüksek nitelik vektör sınıflandırması sağlamıştır. Ayrıca GDI algoritması, indirgenmiş nitelik vektörleri ile çalıştığından diğer algoritmalara göre daha hızlı uygulama süresine, MPWKL ise en yavaş uygulama süresine sahip olmuştur. Kaplan ve diğerlerinin (2015) belirttiği gibi yeni önerilen madde seçme algoritmalarının daha iyi sonuç sağlaması, BiTBOBUT uygulama esnasında nitelik vektör kestirimine ihtiyaç duymamasından kaynaklanmaktadır. Özellikle testin başlangıcında kestirilen nitelik vektörlerinin genellikle doğru olmaması ya da diğer bir ifadeyle yeterlik düzeyinin doğru kestirilememesi, BiTBOBUT uygulamalarındaki madde sayısının artmasına neden olmaktadır. Oysa yeni önerilen algoritmalarda test süresince bireylerin nitelik vektör kestirimi yapılmadığından böyle bir sorun yoktur. GDI, her bir maddenin uygulanmasından sonra hesaplanan sonsal ağırlıklandırılmış 2^K nitelik vektörlerinin basit varyans hesabıdır. GDI madde seçme algoritmasının matematiksel gösterimi Eşitlik 2'de verilmiştir.

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} \left[P(X_{ij}=1 \mid \boldsymbol{\alpha}_{cj}^*) - \bar{P}_j \right]^2 \boldsymbol{\pi}(\boldsymbol{\alpha}_{cj}^*) \tag{2}$$

Eşitlik 2'de $\bar{P}_j$, bireyin yanıt örüntüsüne göre her bir 2^K nitelik vektörü için verilen başarı olasılık değerlerinin ortalamasıdır. $\boldsymbol{\alpha}_c^*$ ise indirgenmiş nitelik vektörüdür. İndirgenmiş nitelik vektörü şu şekilde açıklanabilir: Örneğin beş niteliğin bulunduğu bir konu alanında, maddenin doğru yanıtlanması ikinci, dördüncü ve beşinci nitelikleri gerektirsin. Yukarıda da gösterildiği gibi bu madde için q-vektörü $\boldsymbol{q_j}$ = (0,1,0,1,1) şeklinde gösterilir. İndirgenmiş nitelik vektörü ise $\boldsymbol{q}_j^*$ = (1,1,1) şeklinde gösterilir. GDI ile ilgili daha ayrıntılı bilgi için de la Torre (2011) çalışmasına, madde seçme algoritmaları için ise Kaplan ve diğerlerinin (2015) araştırmasına bakılabilir. Bu çalışmada, madde seçme algoritması olarak PWKL ve GDI kullanılmış olup nitelik sayısının bu iki algoritma üzerindeki etkisi araştırılmıştır.

Araştırmanın Amacı

Bu çalışmanın amacı, BiTBOBUT uygulamalarındaki nitelik sayısının (*attribute size*) madde seçme algoritma performansı üzerindeki etkisini araştırmaktır. Bu bağlamda aşağıdaki sorulara yanıt aranmıştır:

1. Nitelik sayısının artması, iki farklı madde seçme algoritması düşünüldüğünde, test uzunluklarında ne gibi değişikliğe yol açmıştır?
2. Yanıt örüntülerinin oluşturulmasında kullanılan BTM'nin, farklı nitelik sayısı göz önüne alındığında, test uzunluklarında ne gibi değişikliğe yol açmıştır?
3. Test sonlandırma kriterinin, yine farklı nitelik sayısı göz önüne alındığında, test uzunluklarında ne gibi değişikliğe yol açmıştır?

YÖNTEM

Araştırmanın Türü

Bu çalışma, BTM'de genel bir model olan G-DINA modeli çatısı altında bir simülasyon çalışması olup BiTBOBUT uygulamalarında kullanılan madde seçme algoritmalarının farklı nitelik sayısı

kullanıldığındaki performansı, ürettiği farklı test uzunluk değerlerine göre incelenmiştir. Bu sebepten dolayı bu çalışma temel bir araştırmadır.

Madde Havuzu Üretimi

Bu çalışmanın amacı olan nitelik sayısının madde seçme algoritmaları üzerindeki etkisi, K = 4, 5, 6 ve 7 şeklinde dört farklı değer kullanılarak incelenmiştir. Madde havuzunun oluşturulması için nitelik sayısı da göz önüne alınarak bütün olası 2^K-1 tane q-vektörü düşünülmüş ve her bir q-vektör için 20 madde üretilerek madde havuzu oluşturulmuştur. Sonuç olarak K = 4, 5, 6 ve 7 için madde havuz büyüklükleri sırasıyla 300, 620, 1260 ve 2540 olarak belirlenmiştir. Nitelik sayısı artınca madde havuzunun büyüklüğünün de artması, nitelik kestiriminde madde seçme algoritmalarının ihtiyacı kadar olan maddeyi havuzda bulmalarında sorun yaşamaması için gerekli görülmüştür.

Yanıt Örüntülerinin Üretimi

Bireylerin yanıt örüntülerinin üretiminde bazı kısıtlamalarla G-DINA modelinden de elde edilebilen DINA ve DINO modelleri kullanılmıştır. Ayrıca bu modellerde kullanılan tahmin (*guessing*) ve kaydırma (*slip*) parametreleri sırasıyla $g = s = 0{,}1$ olarak sabitlenmiştir. Diğer bir deyişle parametre değerleri sıfıra yakın olduğu için bu simülasyon çalışmasında ayırt ediciliği yüksek maddeler kullanılmıştır. Örneklem büyüklüğü olarak 3.000 bireyin yanıt örüntüleri bu iki modelin varsayımlarına göre nitelik sayısına göre oluşan farklı havuz büyüklükleri için ayrı ayrı üretilmiştir. Yanıt örüntülerinin üretimi Ox (Doornik, 2011) yazılımı kullanılarak kodlanmıştır.

Test Başlangıç Kuralı ve Madde Seçme Algoritmaları

Daha önceden BiTBOBUT uygulamalarında tanımlanmış iki farklı madde seçme algoritması olan PWKL ve GDI kullanılmıştır. Her iki madde seçme algoritması nitelik vektörlerinin sonsal dağılım değerlerini kullandığındantest başlamadan önce madde seçme algoritmalarının değerlerinin hesaplanmasında 2^K nitelik vektörü için tekbiçimli (*uniform*) dağılım kullanılmıştır. Bu şekilde her bir durum için uygulanacak ilk madde rastgele bir şekilde seçilmiştir.

Test Sonlandırma Kuralı ve Nitelik Kestirme Yöntemi

Bu simülasyon çalışmasında BiTBOBUT uygulamaları, nitelik sayısına göre oluşan (2^K) nitelik vektörlerinin sonsal dağılım değerlerinin en büyük-en küçük (EBEK; *minimax*) değerleri baz alınarak sonlandırılmıştır. Bu şekildeki sonlandırma kuralı ile her birey için farklı test uzunlukları elde edilmiştir. Bu sonlandırma kuralıyla ilgili diğer bir nokta ise madde havuzunda yeterli madde bulunduğunda, bu sonlandırma kuralı ile hedeflenen nitelik vektörlerinin doğru sınıflandırılma (NVDS; *correct attribute vector classification*) değerleri EBEK değerlerinden yüksek olacak şekilde elde edilir. Sonlandırma kuralı olarak ayrıca en uzun testin de 80 maddeyi geçmemesi koşul olarak simülasyon çalışmasına tanımlanmış ve üç farklı EBEK değeri olan P = 0,65, 0,75 ve 0,85 BiTBOBUT uygulamasını sonlandırmak için ölçüt olarak alınmıştır.

PWKL algoritması yukarıda da bahsedildiği gibi her madde uygulanmasından sonra nitelik vektörünün kestirimine ihtiyaç duyduğundan, bireylerin nitelik vektörlerinin kestirimi her madde uygulamasından sonra en yüksek sonsal (MAP; *maximum a posteriori*) yöntemi kullanılarak elde edilmiştir. Bireylerin BiTBOBUT uygulaması sonucunda oluşan en son nitelik vektörlerinin kestirimi için de yine MAP yöntemi her iki madde seçme algoritması için kullanılmıştır. GDI madde seçme algoritması bireylerin nitelik vektörlerinin kestirimini sadece BiTBOBUT uygulamasının en sonunda gerçekleştirdiğini belirtmek yerinde olacaktır.

Verilerin Analizi

Verilerin analizinde her durumda bireyler için elde edilen farklı test uzunluklarının en kısa, en uzun, ortalama ve değişim katsayısı değerleri hesaplanmıştır. Değişim katsayısı test uzunluklarının standart sapmalarının ortalama değerlerine oranlanmasıyla bulunur. BiTBOBUT uygulaması Ox (Doornik, 2011) yazılımı kullanılarak kodlanmıştır ve işlemci hızı 2,5 olan bir bilgisayarda simülasyon çalışması gerçekleştirilmiştir.

BULGULAR

Veri analizi sonucunda elde edilen bulgular simülasyon çalışmasında kullanılan her bir etmen için ayrı ayrı incelenmiş ve sonuçlar; DINA ve DINO modelleri için sırasıyla Tablo 1 ve 2'de gösterilmiştir. Test sonlandırma kuralı bölümünde de değinildiği gibi bulguların ilk incelenmesinde elde edilen NVDS oranlarının her bir durum için EBEK değerlerinden yüksek olup olmadığı kontrol edilmiştir. Tablo 1 ve 2'den de görüleceği gibi NVDS oranları her bir durum için EBEK değerlerinden yüksek çıkmıştır. Örneğin K = 4, EBEK değeri 0,65 ve DINA modeli kullanıldığında, PWKL ve GDI için NVDS oranları sırasıyla 0,66 ve 0,74 çıkmıştır. Bu da hedeflenen NVDS oranının (0,65) elde edildiğini göstermiştir. Diğer bir ifadeyle genel olarak BiTBOBUT uygulamasından önce hedeflenen bireylerin nitelik vektörlerinin doğru sınıflandırılmasına, uygulama sonrasında en az EBEK değeri kadar ulaşılmıştır. Elde edilen bulgulardaki diğer ilginç bir nokta ise GDI madde seçme algoritması, K = 4 olduğu zamanda gerçekleşmiştir. Bu durumda kullanılan modelden bağımsız olarak test uzunluğu tüm bireyler için dört (test uzunluklarının standart sapması da sıfır) olarak gerçekleşmiştir. Yani PWKL madde seçme algoritmasından farklı olarak nitelik sayısı dört olduğunda GDI madde seçme algoritması her bir tek nitelikli maddeleri kullanarak ve başka madde tipini kullanmaya ihtiyaç duymadan hedeflenen NVDS oranlarına ulaşmıştır.

Tablo 1. DINA Modeli İçin Elde Edilen Sonuçlar

		GDI					PWKL				
K	EBEK	NVDS	En Kısa	En Uzan	Ortalama	DK	NVDS	En Kısa	En Uzun	Ortalama	DK
4	0,65	0,66	4	4	4,00	0,00	0,74	2	19	5,64	0,37
	0,75	0,82	5	18	6,32	0,29	0,81	2	21	6,84	0,40
	0,85	0,88	5	19	7,36	0,30	0,88	3	29	8,12	0,39
5	0,65	0,74	6	21	7,32	0,27	0,73	2	28	8,55	0,42
	0,75	0,79	6	24	8,20	0,30	0,80	3	29	9,83	0,42
	0,85	0,86	6	28	9,55	0,31	0,89	3	34	11,31	0,41
6	0,65	0,72	7	21	8,86	0,26	0,74	3	44	12,16	0,47
	0,75	0,77	7	36	10,02	0,29	0,81	3	44	13,49	0,46
	0,85	0,91	7	46	12,40	0,25	0,91	3	46	15,10	0,44
7	0,65	0,72	8	31	10,78	0,28	0,73	3	58	17,23	0,50
	0,75	0,78	8	32	12,04	0,28	0,80	3	60	18,73	0,49
	0,85	0,90	8	36	14,43	0,25	0,89	4	64	20,53	0,47

Not: K = Nitelik sayısı, EBEK = En büyük-en küçük, DK = Değişim katsayısı.

Nitelik Sayısı

Tahmin edildiği gibi nitelik sayısı madde seçme algoritmalarının performansını ciddi bir şekilde etkilemiştir. Örnek olarak EBEK değeri 0,65 seçildiğinde nitelik sayısının ortalama test uzunluklarına etkisi DINA ve DINO modelleri için sırasıyla Şekil 1 ve 2'de gösterilmiştir. Elde edilen sonuçlarda dikkati çeken ilk nokta ortalama test uzunluklarının nitelik sayısına göre artışın, PWKL madde seçme algoritması için GDI algoritmasına göre daha hızlı olmasıdır. Örneğin K = 4 olduğunda, her iki madde seçme algoritması için ortalama test uzunlukları PWKL ve GDI için sırasıyla 4 ve 5,64 iken K = 7 olduğunda ortalama test uzunlukları arasındaki fark açılarak sırasıyla 10,78 ve 17,23 olmuştur. Benzer durum en uzun test değerleri için de gözlenmiştir. Örneğin K = 4 ve

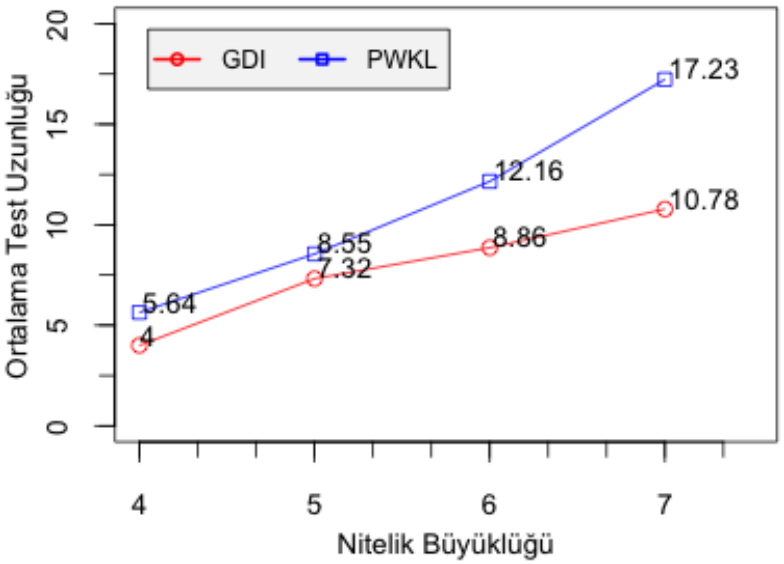
DINA modeli kullanıldığında, en uzun test Tablo 1'den de görüleceği gibi sırasıyla 4 ve 19 iken K = 7 olduğunda test uzunlukları, en uzun 31 ve 58 olmuştur.

PWKL ve GDI madde seçme algoritmalarından elde edilen test uzunlukları karşılaştırıldığında ise en kısa test PWKL madde seçme algoritmasından elde edilmiştir. Örneğin PWKL algoritması için en kısa test uzunluk değerleri iki ile dört arasında alırken GDI için bu değerler dört ile sekiz arasında değişmiştir. Değişim katsayı değerleri PWKL algoritması için nitelik sayısıyla artmakta iken GDI için böyle bir örüntü ortaya çıkmamıştır.

Tablo 2. DINO Modeli İçin Elde Edilen Sonuçlar

		GDI					PWKL				
K	EBEK	NVDS	En Kısa	En Uzun	Ortalama	DK	NVDS	En Kısa	En Uzun	Ortalama	DK
	0,65	0,65	4	4	4,00	0,00	0,73	2	18	5,85	0,40
4	0,75	0,82	5	17	6,38	0,30	0,81	2	21	7,01	0,43
	0,85	0,88	5	22	7,44	0,31	0,88	3	23	8,29	0,41
	0,65	0,75	6	21	7,30	0,28	0,75	2	30	8,89	0,46
5	0,75	0,79	6	25	8,16	0,30	0,82	3	38	10,17	0,46
	0,85	0,86	6	26	9,47	0,30	0,89	3	40	11,54	0,44
	0,65	0,70	7	22	8,93	0,27	0,72	3	46	12,95	0,51
6	0,75	0,76	7	30	10,09	0,30	0,80	3	55	14,27	0,50
	0,85	0,89	7	33	12,46	0,26	0,89	3	56	15,83	0,47
	0,65	0,71	8	26	10,79	0,28	0,75	3	67	18,34	0,54
7	0,75	0,77	8	28	12,08	0,28	0,82	3	71	19,79	0,52
	0,85	0,90	8	36	14,50	0,25	0,90	4	75	21,52	0,50

Not: K = Nitelik sayısı, EBEK = En büyük-en küçük, DK = Değişim katsayısı.



Şekil 1. DINA Modeli ve 0,65 EBEK Değeri İçin Ortalama Test Uzunlukları

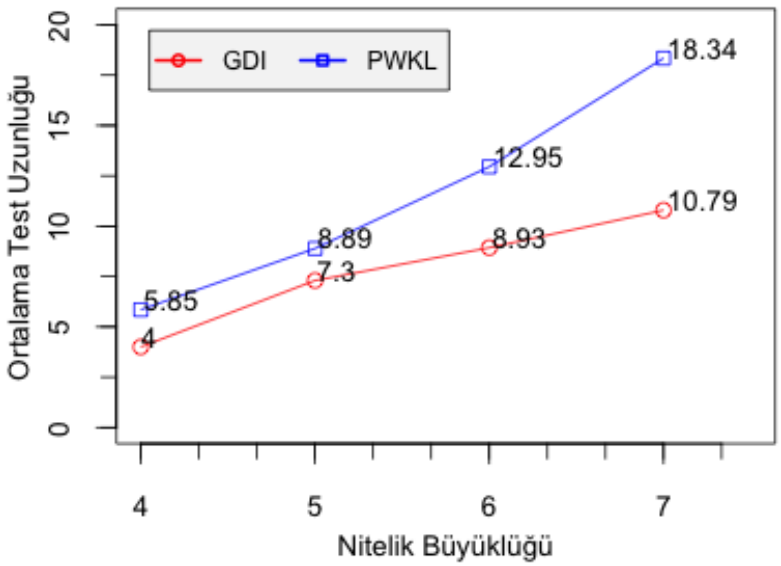
EBEK Değeri

Daha iyi NVDS değerlerinin elde edilmesi hedeflendiğinde test sonlandırma kuralındaki EBEK değerinin arttırılması gerekmektedir. EBEK değerinin arttırılması da her iki madde seçme algoritması için tahmin edildiği gibi en uzun ve ortalama test değerlerinin artmasına neden olmuştur. Örneğin K = 4 ve DINA modeli GDI madde seçme algoritması ile birlikte kullanıldığında, EBEK değerinin 0,65'ten 0,85'e çıkarılması en uzun test değerlerinin 4'ten 19'a ve ortalama test uzunluklarını da 4'ten 7,36'ya yükseltmiştir. Fakat her bir durum için hesaplanan en kısa test

değerleri bazı durumlarda artış gösterirken bazı durumlarda sabit kalmıştır. Özellikle K = 4 olduğunda ve GDI madde seçme algoritması kullanıldığında, en kısa test değerleri modelden bağımsız olarak EBEK değerlerinin artmasıyla artarken nitelik sayısı dörtten fazla olduğu durumlarda (örneğin K = 5, 6 ve 7) ise en kısa test değerleri nitelik sayısı artarken sabit kalmıştır. Aynı durum PWKL madde seçme algoritması kullanıldığında gözlemlenmemiştir. PWKL kullanıldığında en kısa test değerleri bazen artarken bazen sabit kalmıştır. Test uzunluklarının değişim katsayısı incelendiğinde EBEK değerlerine göre belirli bir örüntü ortaya çıkmamıştır.

Bilişsel Tanı Modeli

Bu simülasyon çalışmasında bireylerin yanıt örüntülerinin oluşturulmasında kullanılan iki farklı modelin madde seçme algoritmaları üzerindeki performansları da incelenmiştir. Tablo 1 ve 2'de elde edilen sonuçlardan da görüleceği gibi kullanılan modeller elde edilen test uzunluklarının en kısa, ortalama ve değişim katsayısı değerlerinde gözle görülür bir etkisi olmamıştır. Ancak istisnai olarak PWKL algoritması kullanıldığında ve K = 7 olarak alındığında, EBEK değerlerinden bağımsız olarak iki model arasındaki ortalama test uzunluğu farkı sadece bir madde olarak çıkmıştır. Örneğin sözü edilen durumdaki EBEK değeri 0,65 için; DINA modeli kullanıldığında ortalama test uzunluğu 17,23 olarak; DINO modeli kullanıldığında da test uzunluğu 18,34 olarak elde edilmiştir. Ayrıca kullanılan model en uzun test değerlerinin üzerinde etkisi olmuştur fakat genel bir örüntü ortaya çıkmamıştır. Bu sonuç genel olarak bu iki model için ileride yapılacak olan gerçek BiTBOBUT uygulamalarında model uyumunun ortalama test uzunluklarında bir etkisinin olmayacağını göstermiştir.



Şekil 2. DINO Modeli ve 0,65 EBEK Değeri İçin Ortalama Test Uzunlukları

SONUÇLAR ve TARTIŞMA

MTK ve BOBUT uygulamaları eğitimde ölçme ve değerlendirme alanında üzerinde çalışılmaya devam edilen en önemli konulardandır ve iki konu hakkında da birçok araştırma yapılmıştır. Her ne kadar MTK modelleri kullanılarak uygulanan BOBUT uygulamaları var olan sınav sisteminin ihtiyaçlarını gideriyor olsa da BOBUT uygulamalarının BTM ile kullanılması öğrenme ve öğretme becerilerinin geliştirilmesinde daha etkili olacağı, araştırmacıların ilgisini çekmektedir. Bu bağlamda BiTBOBUT uygulamalarının önemi biçimlendirici değerlendirme tekniklerinin arasında giderek artmaktadır.

Bu çalışmada farklı nitelik sayısının madde seçme algoritmalarının performansı üzerindeki etkileri, test uzunluklarına bakılarak araştırılmıştır. Ayrıca test sonlandırmada kullanılan farklı EBEK

değerlerinin ve bireylerin yanıt örüntülerinin oluşturulmasında kullanılan farklı BTM'nin algoritmalar üzerindeki etkileri de incelenmiştir. Yapılan simülasyon çalışması sonucunda nitelik sayısının arttırılması, ortalama test uzunluklarının artmasıyla sonuçlanmıştır. Daha önemlisi nitelik sayısının artmasıyla ortalama test uzunluk artışı, PWKL algoritması kullanıldığında GDI algoritmasına göre daha hızlı olmuştur. EBEK değerlerinin arttırılması, her iki madde seçme algoritması için ortalama test uzunluklarını arttırırken yanıt örüntülerinin oluşturulmasında kullanılan modellerin ortalama test uzunluklarında herhangi bir etkisi gözlemlenmemiştir. Bu çalışmanın sonucu olarak nitelik sayısının etkisi GDI algoritması için daha az olduğundan ileride yapılacak olan pratik uygulamalarda özellikle fazla nitelik kullanıldığında PWKL algoritmasının kullanılmamasıdır.

Her ne kadar bu çalışma nitelik sayısının algoritmalar üzerindeki etkisini araştırmada bir ön çalışma olarak kabul edilebilse de simülasyonda kullanılan etmenler sonuçları genelleyebilmek için yeterli değildir. Örneğin yanıt örüntülerinin oluşturulmasında kullanılan madde parametreleri her bir madde için sabitlenmiştir. Halbuki gerçek BTM uygulamalarında bu durum söz konusu olmayacaktır. Ayrıca yanıt örüntülerinin oluşturulmasında sadece iki kısıtlayıcı model kullanılmıştır. Oysa literatürde bulunan diğer modeller kullanıldığında madde seçme algoritmalarının performansı merak konusudur. Bu çalışmadaki diğer kısıtlı durum ise madde havuzunun oluşturulmasında her bir q-vektörün kullanılmasıdır. Yine gerçek BTM uygulamalarında her bir q-vektör için madde yazımı mümkün olmayabilir.

KAYNAKÇA

Başokçu, T. O. (2014). Öğrenci yeteneğinin kestiriminde bilişsel tanı modelleri ve uygulamaları. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi.* http://www.efdergi.ibu.edu.tr/index.php/efdergi/article/view/1341 adresinden erişildi.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619-632.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

DiBello, L. V., & Stout, W. (2007). Guest editors introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*, 285-291.

Doornik, J. A. (2011). Object-oriented matrix programming using Ox (Versiyon 6.21) [Computer software]. London, England: Timberlake Consultants Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210.

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*, 563-582.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167-188.

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

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer et al. (Eds.). Computerized adaptive testing: A primer (ss. 101-133). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

von Davier, M. (2008). A general diagnostic model applied to language testing data. *The British Journal of Mathematical and Statistical Psychology, 61*, 287-307.

Xu, X., Chang, H.-H., & Douglas, J. (2003, Nisan). *Computerized adaptive testing strategies for cognitive*

diagnosis. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Quebec, Canada.

EXTENDED ABSTRACT

Introduction

Cognitive diagnosis models (CDMs) have been developed to detect mastery and nonmastery of attributes in a particular area. In contrast to unidimensional item response models (IRTs), CDMs provide more detailed evaluation about the strengths and weaknesses of students (de la Torre, 2009). A general CDM called generalized deterministic inputs, noisy "and" gate (G-DINA; de la Torre, 2011) model, a generalization of the deterministic inputs, noisy "and" gate (DINA; de la Torre, 2009; Haertel, 1989; Junker & Sijtsma, 2001) model, relaxes some of the strict assumptions of the DINA model. The G-DINA model considers the effects of all the possible interactions among the required attributes, which can contribute to the probability of success both in positive and negative ways. Several reduced models can be obtained from the G-DINA model by adding some constrictions to the model.

Computerized adaptive testing (CAT) has been developed as an alternative to paper-and-pencil tests, and it provides more accurate ability estimations of examinees with shorter and tailored tests (Meijer & Nering, 1999; van der Linden & Glas, 2000). Because attribute vectors in CDMs are discrete, some of the concepts in traditional CAT such as Fisher information are not applicable in cognitive diagnosis CAT (CD-CAT). Fortunately, several indices have been proposed for CD-CAT. First, the efficiency of the Kullback-Leibler (K-L) information was investigated as an item selection method in CD-CAT, and the results showed that it outperformed random selection in terms of attribute classification accuracy (Xu, Chang, & Douglas, 2003). Later, another index, namely, the posterior-weighted K-L (PWKL), was proposed, and it is a modified version of the K-L information in which it is weighted by the posterior distribution of the latent classes, and the simulation study showed that PWKL outperformed K-L information (Cheng, 2009). Recently, two new item selection indices, namely, the modified PWKL (MPWKL) and the G-DINA discrimination index (GDI) were proposed for CD-CAT. The results showed that both of the indices performed similarly and had higher classification accuracy compared to the PWKL (Kaplan, de la Torre, & Barrada, 2015).

Method

This simulation study was conducted under the G-DINA model framework. The design of the simulation study consisted of several factors. First, item pools were created considering different levels of attribute size (i.e., K = 4, 5, 6, and 7). For each of 2^K-1 q-vectors, 20 items were created. Therefore, item pool sizes were 300, 620, 1260, and 2540 for K = 4, 5, 6, and 7, respectively. Second, examinees' response patterns were generated using two reduced models, namely, the DINA and DINO models, and the guessing and slip parameters were fixed at 0.1 to eliminate the impact of item quality. Also, the number of examinees was set to 3,000. Third, two item selection indices, namely, the PWKL and GDI, were used in the CD-CAT administration. Uniform distribution was used in the calculation of the indices at the beginning of the test. This also means that the first item was selected randomly for each condition. Last, the CD-CAT administration was terminated after the largest posterior probability of an attribute vector was at least as large as prespecified minimax values, 0.65, 0.75, and 0.85 (Hsu, Wang, & Chen, 2013). This termination rule resulted in different test lengths for different examinees. In addition, maximum a posteriori (MAP) was used to estimate examinees' attribute vectors. Different test lengths for different examinees were examined to analyze the results using the minimum, maximum, mean, and coefficient of variation of the test lengths.

Results and Discussion

Several results can be gleaned from this simulation study. First, as expected, correct attribute vector classification (CVC) rates were always higher than the prespecified minimax values for each condition. For example, using K = 4, the minimax of 0.65 and DINA model, the CVC rates were higher than the minimax, and they were 0.66 and 0.74 for the PWKL and GDI, respectively. Second, interestingly, the test lengths were the same for each examinee (and therefore the standard deviation of the test lengths was zero) when the GDI and K = 4 were used regardless of the model. Third, as expected, increasing the attribute size resulted in longer test lengths regardless of the other factors. In addition, the PWKL yielded even longer test lengths compared to the GDI when the attribute size was larger. For example, when K = 4, the mean test lengths were 4 and 5.64 for the PWKL and GDI, respectively; however, when K = 7, they were 10.78 and 17.23, respectively. The maximum test lengths showed similar patterns as the mean test lengths; however, the PWKL yielded smaller minimum test lengths compared to the GDI. Increasing the attribute size resulted in larger CV rates; however, there was no clear pattern for the GDI. Forth, increasing the prespecified minimax values resulted in larger maximum and mean test lengths for each item selection index regardless of the model. For example, when K = 4 and the DINA model was used, increasing the minimax value from 0.65 to 0.85 increased the maximum test lengths from 4 to 19, and increased the mean test lengths from 4 to 7.36. However, increasing the the prespecified minimax values did not show a clear pattern for the minimum and CV of test lengths. Last, using different CDMs in generating response patterns did not have an impact on the performance of the indices except that when K = 7 and the PWKL was used, the difference of the mean test lengths between two models was only one item. However, using different CDMs resulted in different maximum test lengths for different conditions, but there was no clear pattern to generalize the results.

Compared to traditional unidimensional IRT models, CDMs provide more information that can be used to inform instruction and learning. These models can reveal examinees' strengths and weaknesses by diagnosing whether they have mastered a specific set of attributes. CAT is a tool that can be used to create tests tailored for different examinees. This allows for a more efficient determination of what students know and do not know. In this study, the impact of different levels of attribute size on the item selection indices was examined using different factors in the simulation study. Although this study showed that the GDI is promising when larger attribute sizes to be used, more research needs to be done to determine its viability. For example, it would be interesting to investigate the impact of different CDMs, item quality, and item pool design on the performance of item selection indices.