Araştırma Makalesi

Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 54: 188-214 [2022]
doi:109779.pauefd.769548

BT-BOBUT Uygulamalarında Madde Seçim Algoritmalarının Performanslarının Ölçme Doğruluğu Açısından İncelenmesi

Semih AŞİRET* Seçil ÖMÜR SÜNBÜL**

• ***Geliş Tarihi:*** *15.07.2020* • ***Kabul Tarihi:*** *23.08.2021* • ***Çevrimiçi Yayın Tarihi:*** *19.11.2021*

Öz

Bu çalışmanın amacı, Bilişsel Tanıya Dayalı Bilgisayar Ortamında Bireye Uyarlanmış Testlerde (BT-BOBUT), DINA ve DINO model için farklı nitelik sayısında, madde kalitesinde ve test uzunluklarında madde seçim algoritmalarının performanslarını ölçme doğruluğuna göre incelemektir. Çalışma kapsamında, nitelik sayısı 5 ve 8 olarak değişimlenmiş ve her madde en az bir nitelik ve en fazla 4 nitelik ölçecek şekilde sınırlandırılmıştır. Veri üretiminde, g ve s parametreleri yüksek madde kalite düzeyi için U(0,05-0,25) ve düşük madde kalite düzeyi için U(0.10-0.30) tekbiçimli dağılımdan çekilmiştir. Her bireyin her niteliğe sahip olma şansı %50 olacak şekilde 3000 bireye ait bilişsel örüntüler üretilmiştir. Sonlandırma kuralı olarak 8, 16 ve 24 sabit test uzunlukları kullanılmıştır. Çalışmada kullanılan madde seçim algoritmaları GDI, JSD, MI, PWCDI ve PWKL'dir. Madde seçim algoritmalarının performansları, nitelik ve örüntü koruma oranlarına göre değerlendirilmiştir. Çalışmada veri üretimi ve analizleri R 3.6.3 yazılımı kullanılarak gerçekleştirilmiştir. Çalışma sonucunda, madde kalitesi ve test uzunluğu arttıkça tüm algoritmaların ölçme doğruluk değerlerinin arttığı, nitelik sayısı arttıkça ölçme doğruluğunun azaldığı tespit edilmiştir. JSD algoritmasının ölçme doğruluğu tüm koşullarda en yüksek iken, PWKL algoritmasının en düşük olduğu bulunmuştur. DINA ve DINO modellerde PWKL algoritması dışındaki algoritmaların performansı yaklaşık aynı iken, DINO modelde PWKL algoritmasının ölçme doğruluğunun DINA modelden daha düşük olduğu bulgusu elde edilmiştir.

Anahtar sözcükler: bilişsel tanı modeli, bilgisayar ortamında bireye uyarlanmış test, madde seçim yöntemi, DINA model, DINO model

Atıf: Aşiret, S. & Ömür-Sünbül, S. (2022). BT-BOBUT Uygulamalarında madde seçim algoritmalarınınım performanslarının ölçme doğruluğu açısından incelenmesi. *Pamukkale Eğitim Fakültesi Dergisi 54,* 188-214 doi: 109779.pauefd.769548

* Öğretmen, Milli Eğitim Bakanlığı, ORCID ID: 0000-0002-0577-2603 , semihasiret@gmail.com
** Doç. Dr., Mersin Üniversitesi Eğitim Fakültesi, ORCID ID: 0000-0001-9442-1516, secilomur@gmail.com

Giriş

Eğitimde bilişsel tanıya dayalı değerlendirmeler, model bazlı ölçmeye ve biçimlendirici değerlendirmeye dayalıdır (Embretson, 2001). Son yıllarda bilişsel tanıya dayalı birçok model geliştirilmiştir. Bilişsel tanıya dayalı yapılan değerlendirmelerin öncelikli amacı sonuç odaklı değerlendirme yapmak değildir. Buradaki asıl amaç, bireyin güçlü ve zayıf yanları detaylı tespit edilerek, bireylere etkili geri bildirimde bulunmak, bireylerin öğrenme profillerini ortaya koymak ve bireylerin öğrenme durumlarını kolaylaştırmaktır.

Bilişsel Tanı Modeli (BTM), bir testteki problemi çözmek için gerekli işlemleri veya birçok küçük becerilerin varlığını veya yokluğunu tanılamayı sağlayan kesikli örtük değişkenli modellerdir (de la Torre, 2009). BTM'lerin amacı, bireyin sahip olduğu ve olmadığı nitelikleri (becerilerini) ortaya koymaktır. Madde Tepki Kuramı'nın (MTK) aksine, BTM bireyin niteliklere sahip olma durumunu 1-0'dan oluşan örtük bir vektör ile ifade eder. Birçok BTM mevcuttur. BTM'ler tamamlayıcı, tamamlayıcı olmayan ve genel modeller olarak üç farklı şekilde sınıflandırılmaktadır. Bireyin verilen maddeye doğru cevap vermesi için tamamlayıcı modellerde, maddenin ölçtüğü niteliklerden en az birine, tamamlayıcı olmayan modellerde ise maddenin ölçtüğü tüm niteliklere sahip olması gerekir.

BTM'ye dayalı değerlendirmelerde, öncelikle nitelikler belirlenir. Belirlenen niteliklere göre madde yazımı gerçekleştirilir. Maddeler geliştirildikten sonra Q matrisinin oluşturulur. Q matrisinde satırda maddeler, sütunlarda ise nitelikler yer alır. Maddenin ölçtüğü niteliklere denk gelen hücrelere 1, ölçmediği niteliklere denk gelen hücrelere 0 yazılır. Böylece her maddenin ölçtüğü nitelikler belirlenir.

Tamamlayıcı olmayan ve tamamlayıcı BTM'lerde en bilinen ve en sık kullanılan DINA (deterministic-input, nosiy-and-gate) (Haertel, 1989; Junker ve Sijtsma, 2001) ve DINO (deterministic-input, nosiy-or-gate) (Templin ve Henson, 2006) modellerdir. Her iki model de, kısıtlayıcı modellerdir.

DINA model, tamamlayıcı olmayan ve birleştirici yoğunlaştırma kuralına sahip bir modeldir. Yani, bireyin maddeyi doğru cevaplayabilmesi için Q matrisinde tanımlanan maddenin ölçtüğü tüm niteliklere sahip olması gerekir. Birey, Q matrisinde madde için tanımlanan niteliklerden herhangi birine sahip olmadığı durumlarda maddeye yanlış cevap vereceği sayıltısına sahiptir. DINA modelde, her madde için tahmin etme (g) ve kaydırma (s) olmak üzere iki farklı parametre kestirilmektedir. g parametresi, bireyin maddenin ölçtüğü tüm niteliklere sahip değilken, maddeye doğru cevap verme olasılığını, s parametresi ise,

bireyin, Q matrisinde tanımlanan maddenin ölçtüğü tüm niteliklere sahipken, maddeye yanlış cevap verme olasılığını göstermektedir (Rupp, Templin ve Henson, 2010).

DINA modelde, *i* bireyinin *j* maddesine doğru cevap verme olasılığı Eşitlik 1'de gösterilmektedir.

$$\pi_{ij} = P\left(X_{ij} = 1 \middle| \alpha_i\right) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} \quad (1)$$

Eşitlik 1'de α_i, bireyin bilişsel örüntüsünü, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, örtük tepki örüntüsünü, g, tahmin parametresini, s, kaydırma parametresini göstermektedir. Örneğin $\eta_{ij} = 1$ olduğu durumda, bireyin maddeyi doğru cevaplama olasılığı Eşitlik 2 ile gösterilmektedir.

$$(\pi_{ij}) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} = (1 - s_j)^1 g_j^{1-1} = (1 - s_j) \quad (2)$$

DINO model, DINA modelin tamamlayıcı halidir. DINO model, en az bir tane nitelik ölçümünün olup olmadığını gösteren *ayrıştırıcı yoğunlaştırma kuralı*na sahiptir (Rupp ve diğerleri 2010). Birey, Q matrisinde tanımlanan maddenin ölçtüğü niteliklerden herhangi birine sahip olması durumunda maddeye doğru cevap verecektir. Bireyin, Q matrisinde tanımlanan maddenin ölçtüğü niteliklerden herhangi birine sahip olmaması durumunda maddeye yanlış cevap verecektir sayıltısına sahiptir. Yani maddeyi ölçen herhangi bir niteliğe sahip oluş, diğerlerinin eksikliğinin gidererek tamamlar. DINO modelde, DINA model gibi her madde için kaydırma ve tahmin etme olmak üzere iki parametre kestirilir.

DINO modelde, *i* bireyinin *j* maddesine doğru cevap verme olasılığı Eşitlik 3'te gösterilmektedir.

$$\pi_{ij} = P\left(X_{ij} = 1 \middle| \alpha_i\right) = (1 - s_j)^{\omega_{ij}} g_j^{1-\omega_{ij}} \quad (3)$$

Eşitlik 3'te α_i, bireyin bilişsel örüntüsünü, $\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{qjk}$, örtük tepki örüntüsünü, g, tahmin parametresini, s, kaydırma parametresini göstermektedir.

Bilişsel Tanıya Dayalı Bilgisayar Ortamında Bireye Uyarlanmış Testler

Bilişsel Tanıya Dayalı Bilgisayar Ortamında Bireye Uyarlanmış Testler (BT-BOBUT), Bilişsel Tanı (BT) ve Bilgisayar Ortamında Bireye Uyarlanmış Testlerin (BOBUT) bir araya gelmesiyle oluşmuştur. BOBUT'ta bireyler örtük süreklilik üzerinden bir noktaya yerleştirilirken, bireylere tanısal geri bildirim verilmemektedir. Bunların aksine BT- BOBUT, bireyleri örtük durumlarına göre sınıflamayı ve bu örtük sınıflar üzerinde örtük sınıf modellerini uygulamayı amaçlamaktadır (Cheng, 2009).

BT-BOBUT uygulaması, BOBUT uygulamasıyla benzer aşamalara sahiptir. BOBUT uygulamalarında olduğu gibi, öncelikle ilk madde seçilir. Seçilen madde bireylere uygulanır ve bireylerin bilişsel örüntüleri kestirilir. Kestirilen bilişsel örüntüye göre madde seçim algoritmaları aracılığıyla bir sonraki madde, madde bankasından seçilerek bireye gönderilir. Bu süreç sonlandırma kuralı gerçekleştirilinceye kadar devam eder. Sonlandırma kuralı gerçekleştikten sonra, bireylerin nihai bilişsel örüntüleri kestirilir ve süreç sona erer.

BOBUT uygulamalarında, Fisher En Yüksek Bilgisi (Maksimum Fisher Information-MFI) en popüler madde seçim algoritmalarından biridir (Thissen ve Mislevy, 2000). Ancak MFI, şans başarısından etkilenmesi, kısa testlerde yetenek kestiriminin yeterli olmaması ve kesikli örtük sınıflamalarda kullanılamamasından dolayı BT-BOBUT uygulamalarında kullanılmamaktadır. İlgili literatür incelendiğinde, Kullback-Leibler Bilgisi (Xu, Chang ve Douglas, 2003), Shannon Entropy (Tatsuoka, 2002; Tatsuoka ve Ferguson, 2003), Sonsal Ağırlıklandırılmış Kullback-Leibler bilgisi PWKL ve Hibrid Kullback-Leibler bilgisi (HKL) (Cheng, 2009), Karşılıklı (mutual) bilgi indeksi (Wang, 2013), Değiştirilmiş Sonsal Ağırlıklandırılmış Kullback-Leibler bilgisi (MPWKL) ve GDINA ayırtedicilik indeksi (GDI) (Kaplan, de la Torre ve Barrada, 2015), Sonsal ağırlıklandırılmış bilişsel ayırtedicilik indeksi (PWCDI) ve Sonsal ağırlıklandırılmış nitelik düzeyinde bilişsel ayırtedicilik indeksi (PWACDI) (Zheng ve Chang, 2016) ve Jensen-Shannon uyumsuzluk indeksi (JSD) (Minchen ve de la Torre, 2016) algoritmalarının tek amaçlı BT-BOBUT uygulamalarında madde seçiminde kullanıldığı görülmektedir. Aşağıdaki kısımda bu çalışma kapsamında kullanılan algoritmalardan kısaca bahsedilmiştir.

Sonsal ağırlıklandırılmış KL bilgisi (PWKL)

Cheng (2009), KL algoritmasının düşük etkililiğinden dolayı, her bilişsel örüntünün KL algoritmasına katkısını nicelleştirmek amacıyla KL bilgisi ile bu bilgeye denk gelen sonsal ağırlığı çarparak, sonsal ağırlıklı KL bilgi algoritmasını geliştirmiştir. PWKL algoritması matematiksel olarak Eşitlik 4'te gösterilmektedir.

$$PWKL_j(\hat{\alpha}_l) = \sum_{c=1}^{2^K} \left\{ \sum_{x=0}^{1} \left[\log \left(\frac{P(X_{ij} = x|\hat{\alpha}_l)}{P(X_{ij} = x|\alpha_c)} \right) P(X_{ij} = x|\hat{\alpha}_l) \right] \pi(\alpha_c|x_{t-1}) \right\} \quad (4)$$

Eşitlik 4'te K, toplam nitelik sayısı, $P(X_{ij} = x|\alpha_c)$ verilen α_c bilişsel örüntüsü için i bireyinin j maddesine verdiği x tepkisinin olasılık değeri ve $\pi(\alpha_c|x_{t-1})$, (t-1) madde uygulandıktan sonra bilişsel örüntülerin sonsal ağırlık değerleridir.

Karşılıklı (Mutual) Bilgi İndeksi (MI)

Wang (2013) tarafından geliştirilen MI algoritması, birbirini takip eden sonsal dağılım arasındaki KL uzaklığının eşdeğeri olarak tanımlanır. MI algoritmasına ilişkin eşitlik, Eşitlik 5'te tanımlanmıştır.

$$MI_{ij} = \sum_{c=1}^{2^K} \pi_i(\alpha_c|x_{t-1}) \sum_{x=0}^{1} P\left(X_{ij} = x\middle|\alpha_c\right) \log\left(\frac{P\left(X_{ij} = x\middle|\alpha_c\right)}{P(X_{ij}=x)}\right) \quad (5)$$

Jensen-Shannon Uyumsuzluğu (JSD)

JSD algoritması (Minchen ve de la Torre, 2016), iki seçkisiz dağılımın katışık dağılımı ve bunların marjinal dağılımlarının çarpımı arsında bağıl entropinin ölçüsüdür (Yiğit, Sorrel ve de la Torre, 2019). Maksimum JSD değeri maksimum olan madde, sonraki madde olarak seçilir. JSD algoritmasına ilişkin eşitlik Eşitlik 6'da verilmiştir.

$$JSD_j = S\left(P_j \times \pi'\right) - \sum_{c}^{2^K} \pi_c S\left(P_{jc}\right) \quad (6)$$

Eşitlik 6'da H, seçenek sayısını, $S\left(P_{jc}\right)$, Shannon entropiyi, P_j, $H \times 2^K$ matrisini ve π sonsal olasılık ağırlığını göstermektedir.

Sonsal Ağırlıklandırılmış Bilişsel Ayırtedicilik İndeksi (PWCDI)

Zheng ve Chang (2016) Bilişsel Ayırtedicilik İndeksine (CDI) bilişsel örüntülerin sonsal olasılık dağılımlarını dahil ederek PWCDI algoritmasını geliştirmiştir. Olası bilişsel örüntülerin tepki dağılımları arsındaki KL bilgileri 2^Kx2^K boyutundaki D matrisinde saklanmaktadır. PWCDI algoritmasında bilişsel örüntülerin sonsal olasılık dağılımları, D matrisine dahil edilerek PWD matrisi elde edilir. Bu yönüyle PWKL algoritmasına benzerdir ancak PWKL algoritmasından farklı olarak, matriste satır ve sütun için ağırlıklandırma yapılmaktadır. PWD matrisi Eşitlik 7 ile tanımlanmaktadır.

$$PWD_{juv} = E_{\alpha_u}\left[\pi(\alpha_u) \times \pi(\alpha_v) \times log\left(\frac{P\left(X_j|\alpha_u\right)}{P\left(X_j|\alpha_v\right)}\right)\right] \quad (7)$$

PWCDI algoritması ise Eşitlik 8 ile gösterilmektedir.

$$PWCDI_j = \frac{1}{\sum_{u\neq v} h(\alpha_u,\alpha_v)^{-1}} \sum_{u\neq v} h(\alpha_u, \alpha_v)^{-1} PWD_{juv} \quad (8)$$

Eşitlik 8'de $h(\alpha_u, \alpha_v) = \sum_{k=1}^{K} |\alpha_{uk} - \alpha_{vk}|$, iki bilişsel örüntü arasındaki hamming mesafesini göstermektedir.

Araştırmanın Amacı ve Önemi

BT-BOBUT çalışmalarında anahtar durum madde seçim algoritmalarıdır (Cheng, 2009). Hsu ve Wang (2015) BT-BOBUT çalışmalarında, iyi madde seçiminin ölçme doğruluğunu artırdığını belirtmiştir. BT-BOBUT uygulamaları için son yıllarda yeni madde seçim algoritmaları geliştirilmiştir. Ancak geliştirilen bu algoritmaların performanslarını ölçme doğruluğu açısından farklı modellerde ve farklı koşullarda değerlendiren yeterli çalışma yer almamaktadır.

Bu çalışmanın amacı, madde seçim algoritmalarının (GDI, JSD, MI, PWCDI, PWKL) DINA ve DINO modellerinde, çeşitli test uzunluklarında (8, 16 ve 24), madde kalitesi düzeylerinde (düşük ve yüksek) ve nitelik sayılarında (5 ve 8) performanslarını ölçme doğruluğu (örüntü koruma oranı ve nitelik koruma oranı) açısından değerlendirmektir. Bu amaç doğrultusunda belirlenen koşullarda en yüksek ölçme doğruluğunun hangi madde seçim algoritmalarından elde edildiği ortaya konarak, pratikteki uygulamalara yardımcı olacağı düşünülmektedir.

Yöntem

Araştırmanın Türü

Bu çalışmada, BT-BOBUT uygulamasında madde seçim algoritmalarının performanslarının analiz edilen modele, nitelik sayısına, madde kalitesine ve test uzunluğuna göre ölçme doğruluğu açısından incelenmesi amaçlanmıştır. Bu açıdan çalışmada madde seçim algoritmalarının çeşitli faktörlere göre nitelik ve örüntü koruma oranlarının incelenmesi amaçlandığından bu çalışma, temel araştırmadır.

Araştırma Kapsamında Değişimlenen Faktörler

Analiz Modeli: Çalışmada, tamamlayıcı BTM olarak DINO ve tamamlayıcı olmayan BTM olarak DINA model kullanılmıştır. Her iki modelin tercih edilmesindeki temel gerekçe, tamamlayıcı ve tamamlayıcı olmayan modeller olması, pratikte sıklıkla tercih edilmesi ve hesaplama kolaylığıdır. Hesaplama kolaylığı, bireyselleştirilmiş testlerde istenen bir özelliktir.

Madde kalitesi: Çalışmada madde kalitesi düşük ve yüksek olmak üzere iki farklı şekilde değişimlenmiştir. Madde parametrelerinin üretiminde Zheng ve Chang (2016) tarafından belirlenen parametre değerleri kullanılmıştır. Her iki model için yüksek madde kalitesi için g

ve s parametreleri U(0,05-0,25) ve düşük madde kalitesi için U(0,10-0,30) tekbiçimli dağılımdan üretilmiştir.

Nitelik sayısı: İlgili literatür incelendiğinde, Cheng (2009) ve Wang (2013) 5 nitelik düzeyinin orta olduğunu ve von Davier (2005), pratikte nitelik sayısının en fazla 8 olması gerektiğini ifade etmiştir. Bu çalışmada nitelik sayıları, orta ve yüksek olmak üzere 5 ve 8 olarak değişimlenmiştir. Ayrıca gerçek uygulamalarla benzerlik sağlaması amacıyla her madde en fazla 4 nitelik ölçecek şekilde sınırlandırılmıştır.

Test uzunluğu: DiBello, Roussos ve Stout (2007) BT-BOBUT çalışmalarının sıklıkla sınıf içi değerlendirmelerde, biçimlendirici değerlendirme amacıyla kullanıldığını ve test uzunluklarının kısa olması gerektiğini belirtmiştir. Bu açıdan, nitelik sayıları da dikkate alınarak test uzunlukları 8, 16 ve 24 olarak değişimlenmiştir.

Madde Seçim Algoritmaları: Çalışma kapsamında madde seçim algoritması olarak Cheng (2009) tarafından geliştirilen PWKL, Wang (2013) tarafından geliştirilen MI, Kaplan ve diğerleri (2015) tarafından geliştirilen GDI, Zheng ve Chang (2016) tarafından geliştirilen PWCDI ve Minchen ve de la Torre (2016) tarafından geliştirilen JSD algoritmaları kullanılmıştır.

Veri Üretimi ve İşlem

Çalışmadan verilerin üretimi ve analizi R 3.6.3 (R Core Team, 2020) yazılımı kullanılarak gerçekleştirilmiştir. Verilerin üretiminde GDINA paketi (v2.8; Ma ve de la Torre, 2020) ve grafiklerin oluşturulmasında ggplot2 (v3.3.2; Wickham, 2016) paketi kullanılmıştır. Diğer işlemler için kodlar araştırmacılar tarafından R 3.6.3 (R Core Team, 2020) yazılımında yazılmıştır.

Madde bankası ve bireylerin üretimi: Stocking (1994) madde bankasının test uzunluğunun en az 12 katı olacak şekilde üretilmesi gerektiğini ifade etmiştir. Bu amaçla, 5 ve 8 nitelik düzeyleri için 500 madden oluşan iki farklı madde bankası üretilmiştir. Q matrisi, her bir niteliğin madde tarafından ölçülme şansı %30 olacak şekilde oluşturulmuştur. Q matrisi, madde madde ve nitelik nitelik olarak üretilmiştir. Ayrıca gerçek uygulamalara benzerliğin sağlanması amacıyla Q matrisinde yer alan her madde en az bir nitelik, en çok dört nitelik ölçecek şekilde sınırlandırılmıştır. Böylelikle Q matrisinde 5 nitelik düzeyinde 30, 8 nitelik düzeyinde ise 162 farklı bilişsel örüntü yer almaktadır. Her bireyin her niteliği başarma olasılığı % 50 olacak şekilde, 5 ve 8 nitelik düzeyleri için 3000 bireye ait bilişsel örüntüler üretilmiştir. Üretilen bireylerin bilişsel örüntülerine ve Q matrisine göre bireylerin maddelere

verdiği tepkiler 1-0 olarak üretilmiştir. Bu işlemin ardından her örüntü için DINA ve DINO modele göre maddelere doğru cevap verme olasılıkları hesaplanmıştır.

Tablo 1'de 5 ve 8 nitelik düzeyinde, her niteliği ölçen madde sayısı ve niteliğe sahip olan birey sayıları verilmiştir. Tablo 1 incelendiğinde, Q matrisinin madde madde ve nitelik nitelik oluşturulmasından dolayı her niteliği ölçen madde sayıları yaklaşık olarak eşittir. Tablo 1'e göre 5 nitelik düzeyi için, birinci niteliği ölçen madde sayısı 169, ikinci niteliği ölçen madde sayısı 179, üçüncü niteliği ölçen madde sayısı 185, dördüncü niteliği ölçen madde sayısı 189 ve beşinci niteliği ölçen madde sayısı 182'dir. Benzer dağılım 8 nitelik düzeyi içinde geçerlidir. Tablo 1'de her bir niteliğe sahip birey sayıları incelendiğinde, 5 ve 8 nitelik sayıları için her nitelik düzeyinde birey sayılarının yaklaşık eşit olduğu görülmektedir. Bu durum, bireylerin üretiminde her bireyin her niteliği başarma olasılığı % 50 olacak şekilde sınırlandırılmasından kaynaklanmaktadır.

Tablo 1. *5 ve 8 nitelik düzeyinde, her niteliği ölçen madde sayısı ve niteliğe sahip olan birey sayıları*

	Nitelikler							
K=5	1	2	3	4	5			
Madde Sayısı (J=500)	169	179	185	189	182			
Birey Sayısı (N=3000)	1493	1543	1492	1495	1499			
K=8	1	2	3	4	5	6	7	8
Madde Sayısı (J=500)	156	133	135	161	146	153	154	170
Birey Sayısı (N=3000)	1493	1503	1524	1547	1475	1494	1450	1509

Not: K, toplam nitelik sayısı, J, madde bankasında yer alan madde sayısı, N, toplam birey sayısı

Tablo 2'de 5 ve 8 nitelik düzeyinde, olası nitelik sayısını ölçen madde ve niteliğe sahip olan birey dağılımları verilmiştir. Madde bankasında, her madde en az bir nitelik ve en fazla 4 nitelik ölçecek şekilde sınırlandırma getirildiğinden, hiçbir niteliği ölçmeyen ve dörtten fazla niteliği ölçen madde bulunmamaktadır. Tablo 2 incelendiğinde, 5 nitelik düzeyinde en çok bir (204) ve iki nitelik (203) ölçen maddelerin olduğu görülürken, üç nitelik ölçen madde

sayısı 78 ve dört nitelik ölçen madde sayısı 15'tir. 8 nitelik düzeyinde, en çok iki nitelik ölçen (168) maddeler yer almaktadır. Bir nitelik ölçen madde sayısı, 105, üç nitelik ölçen madde sayısı 141 ve dört nitelik ölçen madde sayısı 86'dır. 5 nitelik düzeyinde, hiçbir niteliğe sahip olmayan birey sayısı 97 iken 8 nitelik düzeyinde ise 6'dır.

Tablo 2. *5 ve 8 nitelik düzeyinde, olası nitelik sayısını ölçen madde ve niteliğe sahip olan birey dağılımları*

	Nitelikler								
Nitelik Sayısı (K=5)	0	1	2	3	4	5			
Madde Sayısı (J=500)	0	204	203	78	15	0			
Birey Sayısı (N=3000)	97	449	942	955	461	96			
Nitelik Sayıları (K=8)	0	1	2	3	4	5	6	7	8
Madde Sayısı (J=500)	0	105	168	141	86	0	0	0	0
Birey Sayısı (N=3000)	6	86	334	665	834	658	312	92	13

Not: K, toplam nitelik sayısı, J, madde bankasında yer alan madde sayısı, N, toplam birey sayısı

***İlk madde seçimi*:** BT-BOBUT uygulaması ilk madde seçimiyle başlar. Bu çalışmada ilk madde seçimi seçkisiz olarak yapılmıştır ve madde seçim algoritmalarını eş koşullarda değerlendirmek amacıyla seçkisiz seçilen madde, tüm algoritmalarda ilk madde olarak kullanılmıştır.

Bilişsel örüntünün kestirilmesi: BT-BOBUT uygulamaları sıklıkla sınıf içi değerlendirmelerde kullanılmaktadır. Ders sürecinde uygulanan testlerin uzunlukları genellikle kısadır. Bu durumda, maddelerin tamamına doğru (1) veya tamamına yanlış (0) tepki veren bireylerin olma olasılığı yüksek olabilmektedir. Bireylerin tepki örüntülerinin tamamı 0 veya 1 olduğunda, En çok olabilirlik (MLE) yöntemi doğru kestirim yapamamaktadır. Bu nedenle çalışmada bireylerin bilişsel örüntüleri maximum a posteriori (MAP) kestirim yöntemiyle kestirilmiştir.

Değerlendirme ölçütü: Bu çalışmada, madde seçim algoritmalarının performansları nitelik ve örüntü düzeyinde değerlendirilmiştir. Nitelik düzeyinde, madde seçim algoritmalarının nitelik

koruma oranları (NKO) ve bilişsel örüntü düzeyinde madde seçim algoritmalarının örüntü koruma oranları (ÖKO) hesaplanmıştır. NKO ve ÖKO Eşitlik 9 ve Eşitlik 10 ile hesaplanmıştır.

$$NKO_k = \frac{\sum_{i=1}^{N} A_{ik}}{N} = \frac{\sum_{i=1}^{N} \left(I_{\widehat{\alpha_{ik}},\alpha_{ik}}\right)}{N}, \qquad (k=1, 2, \ldots., K) \tag{9}$$

$$ÖKO_k = \frac{\sum_{i=1}^{N} R_i}{N} = \frac{\sum_{i=1}^{N} \left(I_{\widehat{\alpha_i},\alpha_i}\right)}{N}, \qquad (k=1, 2, \ldots., K) \tag{10}$$

BT-BOBUT süreci: Çalışmada, Q matrisleri, bireylere ait bilişsel örüntüler ve bireylerin maddelere verdiği tepkiler üretildikten sonra, her örüntünün maddeye doğru cevap verme olasılıkları hesaplanmıştır. Bu aşamadan sonra, BT-BOBUT süreci başlatılmıştır. BT-BOBUT sürecinde, seçkisiz olarak seçilen başlangıç maddesi tüm bireylere uygulanmış ve her bireyin olası bilişsel örüntüleri, MAP kestirim yöntemi ile kestirilmiştir. Ardından, her birey için madde seçim algoritması tarafından seçilen madde, sonraki madde olarak uygulanmıştır. Bu süreç sonlandırma kuralı gerçekleştirilinceye kadar tekrarlanmıştır. Süreç tamamlandığında, her madde seçim algoritmasının, her koşulda NKO ve ÖKO değerleri hesaplanmıştır. Elde edilen bu değerlere ilişkin tablolar ve grafikler oluşturulmuştur.

Bulgular

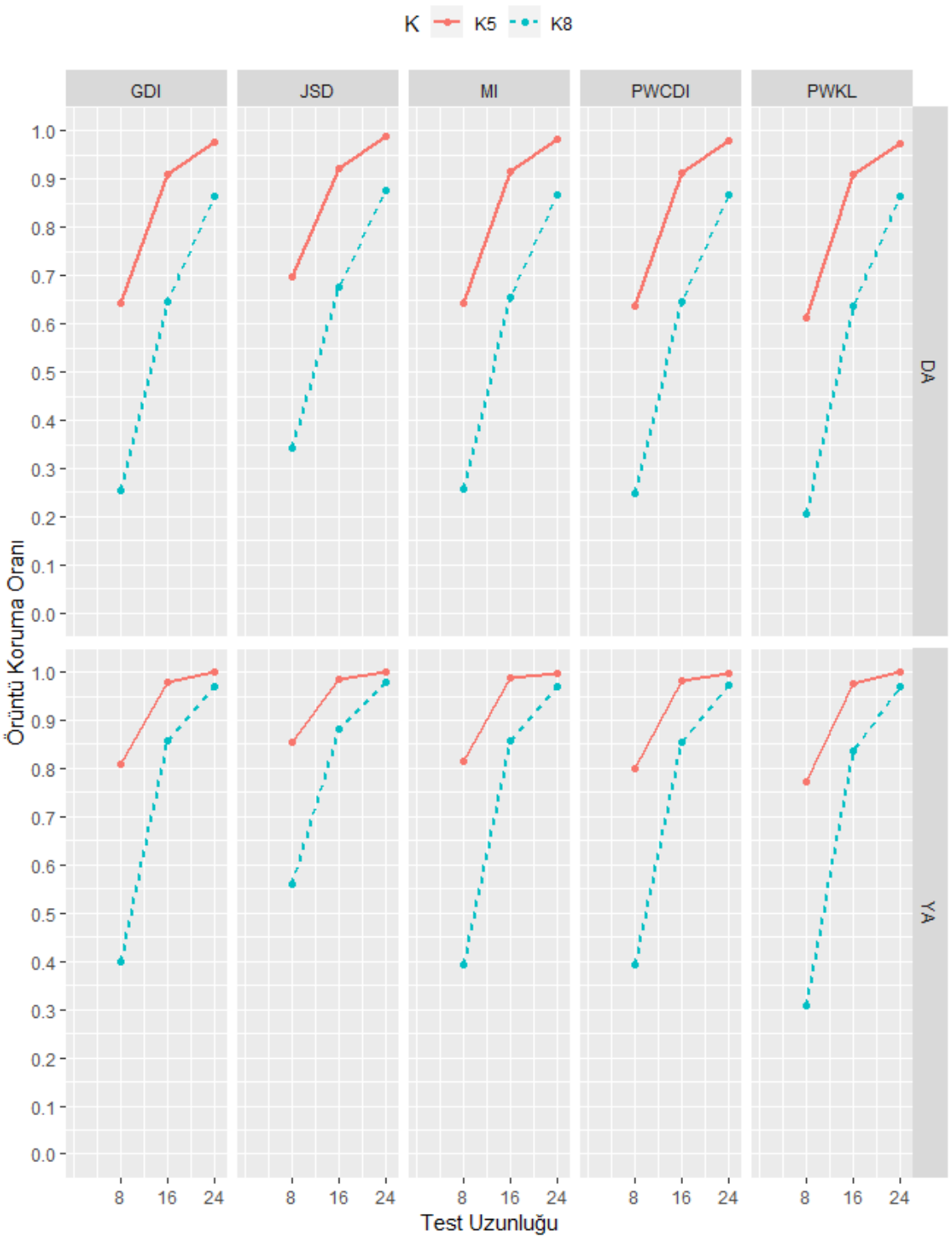
DINA modelde, madde kalitesine ve nitelik sayısına göre, sabit test uzunluğu sonlandırma kuralında, madde seçim algoritmalarının örüntü koruma oranlarına ilişkin bulgular

Tablo 3'te araştırmada yer alan faktör düzeylerinde, DINA modele göre madde seçim algoritmalarının örüntü koruma oranları verilmiştir. Ayrıca bu oranlar grafiksel olarak Şekil 1'de verilmiştir.

Tablo 3. *DINA Modele Göre Madde Seçim Algoritmalarının Örüntü Koruma Oranları*

Model	K	Madde Kalitesi	Madde Sayısı	Algoritmalar				
				GDI	JSD	MI	PWCDI	PWKL
DINA	5	Düşük	8	0,644	0,699	0,644	0,638	0,612
			16	0,909	0,921	0,917	0,914	0,910
			24	0,978	0,989	0,982	0,979	0,975
		Yüksek	8	0,808	0,855	0,814	0,799	0,773
			16	0,980	0,985	0,987	0,982	0,976
			24	0,999	0,999	0,998	0,997	0,999
	8	Düşük	8	0,254	0,343	0,256	0,248	0,207
			16	0,647	0,676	0,655	0,647	0,638
			24	0,866	0,876	0,869	0,867	0,864
		Yüksek	8	0,399	0,560	0,394	0,392	0,308
			16	0,857	0,881	0,859	0,853	0,836
			24	0,971	0,98	0,971	0,973	0,971

Tablo 3 ve Şekil 1 birlikte incelendiğinde, madde seçim algoritmalarının ÖKO değerlerinin, maddenin ölçtüğü nitelik sayısı arttıkça önemli ölçüde azaldığı görülmektedir. Tüm koşullarda, 5 nitelik düzeyindeki ÖKO değerleri daha yüksektir. Ancak test sonlandırma kuralının 24 madde olduğu ve madde kalitesinin yüksek olduğu durumlarda, 5 ve 8 nitelik düzeyinde algoritmaların ÖKO değerleri 1'e yaklaşmaktadır.



Şekil 1. *DINA Modele Göre Madde Seçim Algoritmalarının Örüntü Koruma Oranları*

Sonlandırma kuralı olarak 8 test uzunluğu kullanıldığında, 5 nitelik düzeyi için düşük madde kalitesinde algoritmaların örüntü koruma oranları 0,612 - 0,699 aralığında, yüksek madde kalitesinde ise 0,855 - 0,975 aralığında değişmektedir. Test uzunluğu arttıkça, farklı madde kalite düzeylerinde algoritmaların ÖKO değerleri artmaktadır. 24 test uzunluğunda ve 5 nitelik düzeyinde, tüm düşük ve yüksek madde kalite düzeyinde, madde seçim

algoritmalarının ÖKO değerleri 1'e yakındır. 8 nitelik düzeyinde ise, sadece madde kalitesi yüksek olduğunda ve 24 test uzunluğu sonlandırma kuralında, ÖKO değerlerinin 1'e yakın olduğu söylenebilir. 5 nitelik düzeyinde, madde kalitesi yüksek ve test uzunluğu 16 olduğunda, madde seçim algoritmalarının örüntü korumalarının da 1'e yakın olduğu söylenebilir. Tablo3'e göre, tüm koşullarda en yüksek ÖKO değerleri JSD algoritmasından, en düşük ÖKO değerleri ise PWKL algoritmasından elde edildiği görülmektedir. Diğer algoritmaların ÖKO değerleri ise birbirine çok yakındır.

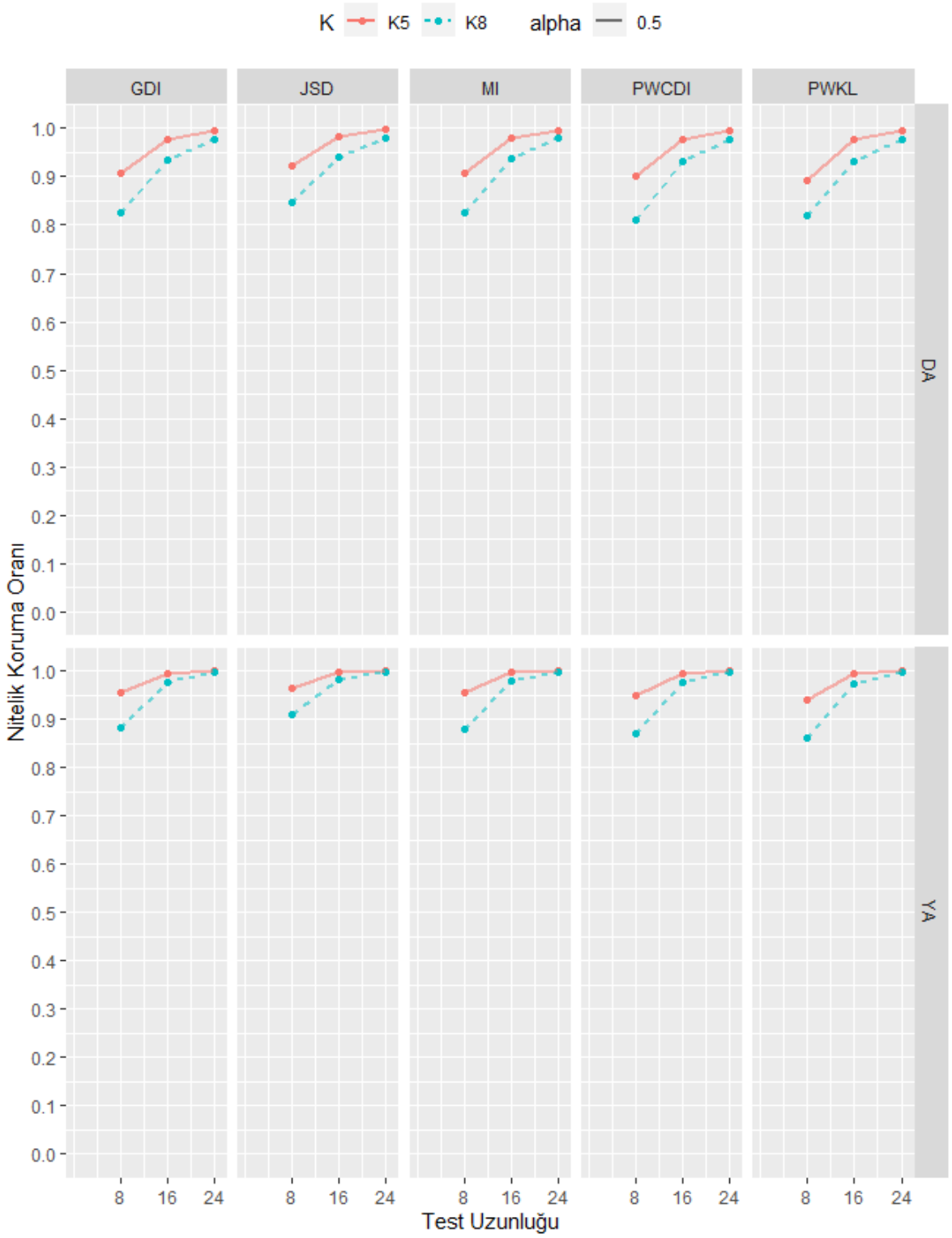
DINA modelde, madde kalitesine ve nitelik sayısına göre, sabit test uzunluğu sonlandırma kuralında, madde seçim algoritmalarının, nitelik koruma oranlarına ilişkin bulgular

Tablo 4'te araştırmada yer alan faktör düzeylerinde DINA modele göre madde seçim algoritmalarının nitelik koruma oranları verilmiştir. Ayrıca bu oranlar grafiksel olarak Şekil 2'de verilmiştir.

Tablo 4 ve Şekil 2 birlikte incelendiğinde, test uzunluğu arttıkça madde seçim algoritmalarının NKO değerlerinin arttığı, test uzunluğu 24 ve madde kalitesi yüksek olduğunda madde seçim algoritmaların NKÖ değerlerinin yaklaşık 1,00 olduğu, test uzunluğu 24 ve madde kalitesi düşük olduğunda ise bu değerlerin 5 nitelik düzeyinde 1, 8 nitelik düzeyinde 1'e yakın olduğu görülmektedir. JSD algoritmasına ait NKO değerlerinin diğer algoritmalara göre az da olsa daha yüksek olduğu söylenebilir.

Tablo 4. *DINA Modele Göre Madde Seçim Algoritmalarının Nitelik Koruma Oranları*

Model	K	Madde Kalitesi	Madde Sayısı	Algoritmalar				
				GDI	JSD	MI	PWCDI	PWKL
DINA	5	Düşük	8	0,906	0,922	0,907	0,9	0,891
			16	0,978	0,982	0,98	0,978	0,978
			24	0,995	0,998	0,996	0,995	0,994
		Yüksek	8	0,954	0,964	0,955	0,948	0,939
			16	0,995	0,997	0,997	0,995	0,994
			24	1,00	1,00	1,00	0,999	1,00
	8	Düşük	8	0,826	0,846	0,825	0,812	0,819
			16	0,936	0,942	0,937	0,931	0,932
			24	0,978	0,98	0,978	0,977	0,976
		Yüksek	8	0,881	0,91	0,88	0,871	0,861
			16	0,977	0,981	0,978	0,975	0,972
			24	0,996	0,997	0,996	0,996	0,996



Şekil 2. *DINA Modele Göre Madde Seçim Algoritmalarının Nitelik Koruma Oranları*

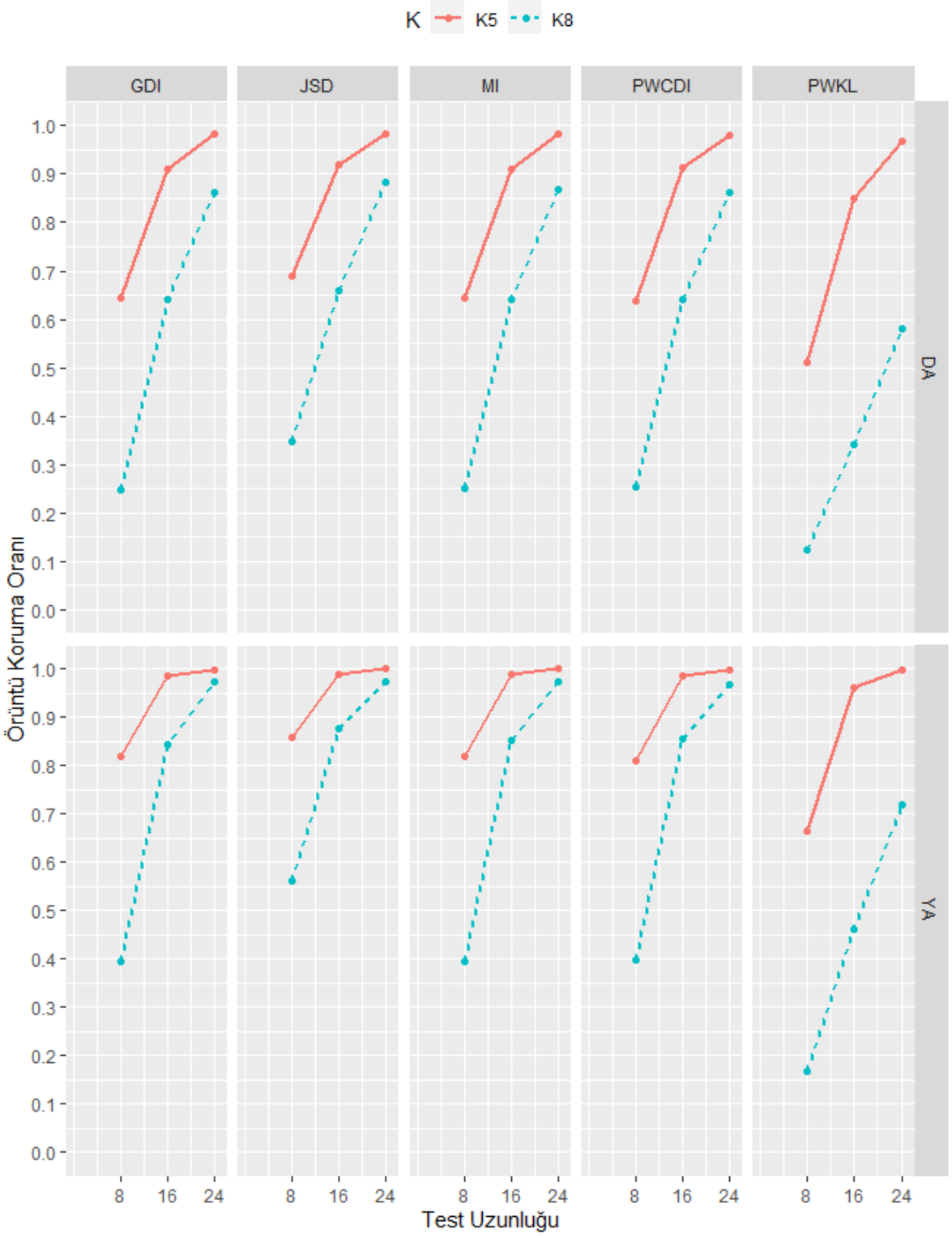
DINO modelde, madde kalitesine ve nitelik sayısına göre, sabit test uzunluğu sonlandırma kuralında, madde seçim algoritmalarının örüntü koruma oranlarına ilişkin bulgular

Tablo 5'te, araştırmada yer alan faktör düzeylerinde DINO modele göre madde seçim algoritmalarının örüntü koruma oranları verilmiştir. Ayrıca bu oranlar grafiksel olarak Şekil 3'te verilmiştir.

Tablo 5. *DINO Modele Göre Madde Seçim Algoritmalarının Örüntü Koruma Oranları*

Model	K	Madde Kalitesi	Madde Sayısı	Algoritmalar				
				GDI	JSD	MI	PWCDI	PWKL
DINO	5	Düşük	8	0,644	0,689	0,645	0,637	0,511
			16	0,911	0,921	0,909	0,912	0,851
			24	0,982	0,984	0,983	0,981	0,968
		Yüksek	8	0,819	0,858	0,819	0,81	0,663
			16	0,986	0,988	0,987	0,984	0,962
			24	0,998	0,999	0,999	0,997	0,997
	8	Düşük	8	0,247	0,348	0,25	0,254	0,124
			16	0,642	0,66	0,64	0,641	0,341
			24	0,861	0,884	0,868	0,863	0,579
		Yüksek	8	0,396	0,56	0,394	0,398	0,167
			16	0,844	0,876	0,853	0,854	0,46
			24	0,972	0,973	0,973	0,966	0,719

Şekil 3 incelendiğinde, DINA modelde olduğu gibi, test uzunluğu ve madde kalitesi arttıkça madde seçim algoritmalarının ÖKO değerlerinin arttığı, nitelik sayısı arttığında ise ÖKO değerlerinin azaldığı söylenebilir.



Şekil 3. *DINO Modele Göre Madde Seçim Algoritmalarının Örüntü Koruma Oranları*

Tablo 5'e göre 5 nitelik düzeyinde, düşük madde kalitesinde ve 8 test uzunluğunda algoritmaların ÖKO değerleri 0,511 - 0,689 aralığında, yüksek madde kalitesinde ise 0,663 - 0,858 aralığında değişmektedir. 8 nitelik düzeyinde, 8 test uzunluğu ve düşük madde kalitesinde, ÖKO değerleri 0,124 - 0,348 aralığında, yüksek madde kalitesinde ise, 0,167 -

0,56 arlığında değişmektedir. Sonlandırma kuralı 24 test uzunluğu olduğunda, 5 nitelik düzeyinde algoritmaların ÖKO değerleri 1'e yakındır. 8 nitelik düzeyinde ve yüksek madde kalitesinde, PWKL algoritması haricinde diğer algoritmaların ÖKO değerleri 1'e yakınken, düşük madde kalitesinde bu değerler 0.589 - 0,884 arasında değişmektedir. Algoritmaların ÖKO değerleri aralığının bu kadar geniş olmasının nedeni, PWKL algoritmasının diğer algoritmalara göre ÖKO değerlerinin önemli ölçüde düşük olmasıdır. Özellikle 8 nitelik düzeyinde, PWKL algoritmasının ÖKO değerleri, 5 nitelik düzeyine göre önemli ölçüde azalmıştır. Tüm koşullarda JSD algoritmasının ÖKO değerleri daha yüksek iken, PWKL algoritmasının ÖKO değerleri en düşüktür. PWKL algoritması dışındaki algoritmaların ÖKO değerleri ise, JSD algoritmasının ÖKO değerlerine yakındır.

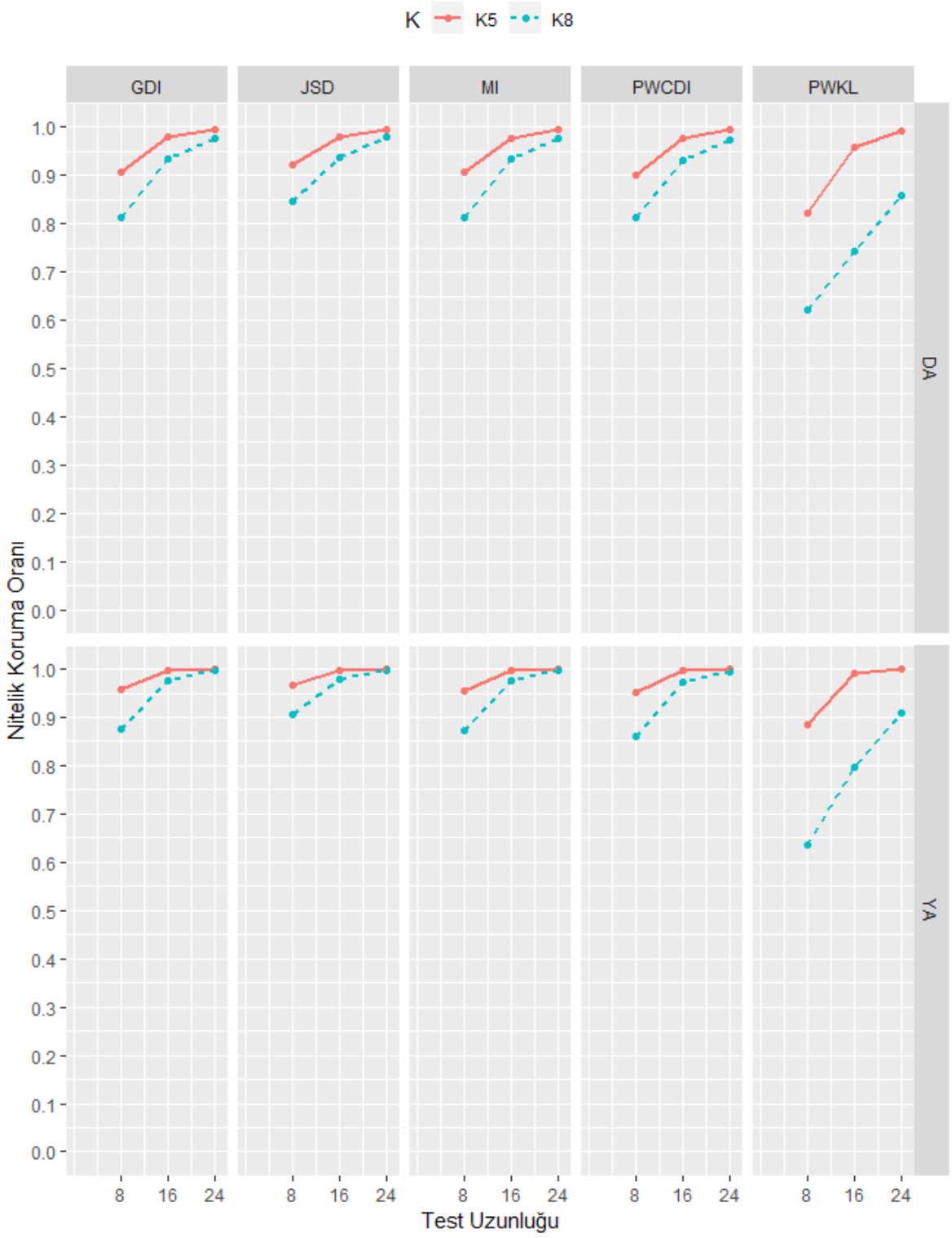
DINO modelde, madde kalitesine ve nitelik sayısına göre, sabit test uzunluğu sonlandırma kuralında, madde seçim algoritmalarının nitelik koruma oranlarına ilişkin bulgular

Tablo 6'da, araştırmada yer alan faktör düzeylerinde DINO modele göre madde seçim algoritmalarının nitelik koruma oranları verilmiştir. Ayrıca bu oranlar grafiksel olarak Şekil 4'te verilmiştir.

Tablo 6 ve Şekil 4 incelendiğinde, DINO modelde elde edilen NKO değerlerinin PWKL algoritması dışında, DINA modelde elde edilen NKO değerlerine benzer olduğu söylenebilir. Düşük madde kalitesinde ve 5 nitelik düzeyinde algoritmaların NKO değerleri tüm test uzunluklarında daha yüksek iken, madde kalitesi arttığında ve 24 test uzunluğunda PWKL algoritması dışındaki algoritmaların NKO değerlerinin 1'e yakın olduğu söylenebilir. Şekil 4'e göre, nitelik sayısının artmasıyla, PWKL algoritmasının NKO değerlerinin, diğer algoritmaların NKO değerlerine göre daha fazla düştüğü söylenebilir.

Tablo 6. *DINO Modele Göre Madde Seçim Algoritmalarının Nitelik Koruma Oranları*

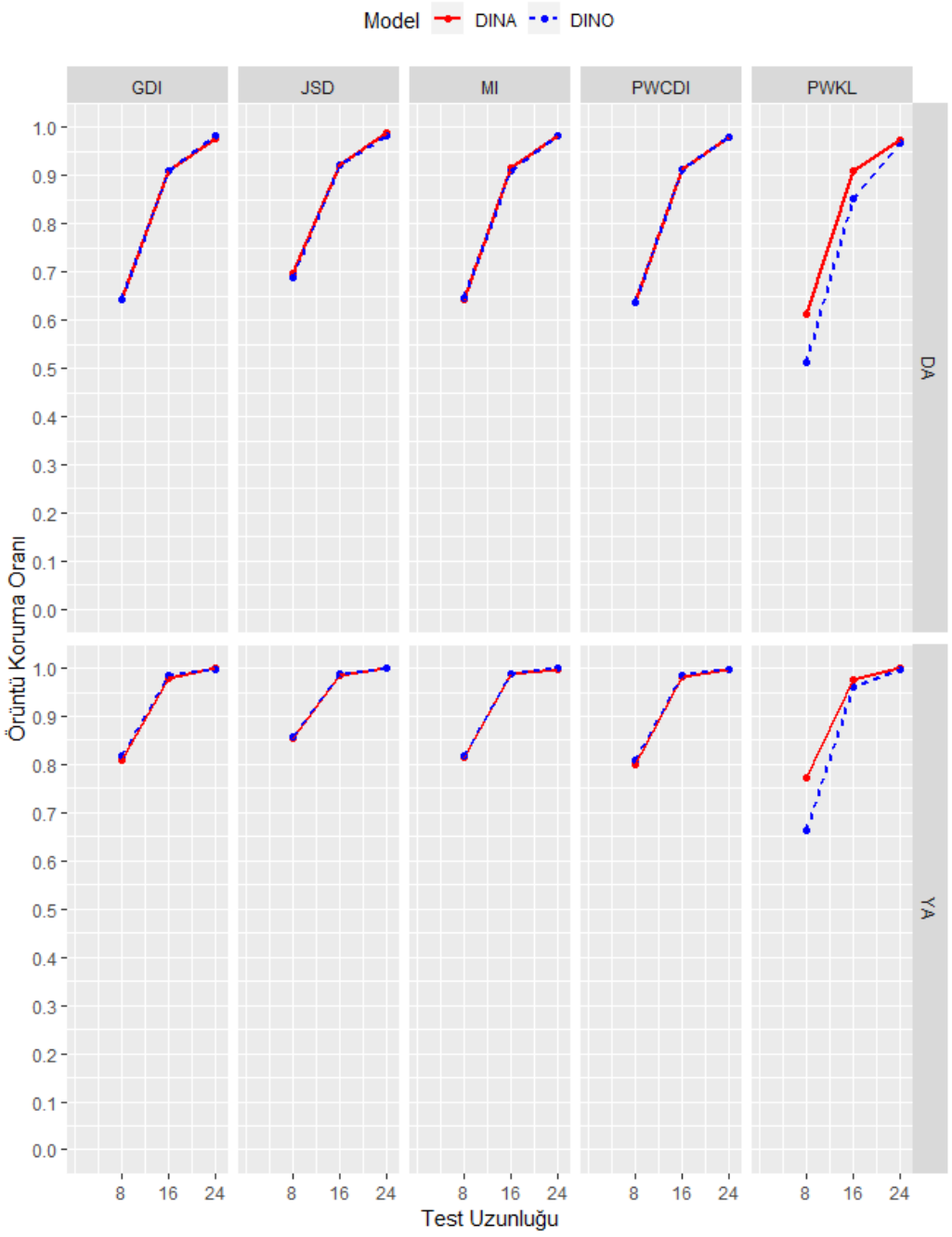
Model	K	Madde Kalitesi	Madde Sayısı	Algoritmalar				
				GDI	JSD	MI	PWCDI	PWKL
DINO	5	Düşük	8	0,906	0,921	0,906	0,9	0,821
			16	0,979	0,981	0,978	0,977	0,958
			24	0,996	0,996	0,996	0,995	0,992
		Yüksek	8	0,957	0,966	0,956	0,952	0,886
			16	0,997	0,997	0,997	0,996	0,99
			24	1	1	1	0,999	0,999
	8	Düşük	8	0,814	0,846	0,813	0,812	0,622
			16	0,933	0,938	0,933	0,931	0,743
			24	0,977	0,981	0,978	0,975	0,858
		Yüksek	8	0,875	0,906	0,873	0,861	0,635
			16	0,975	0,98	0,976	0,974	0,797
			24	0,996	0,996	0,996	0,995	0,91



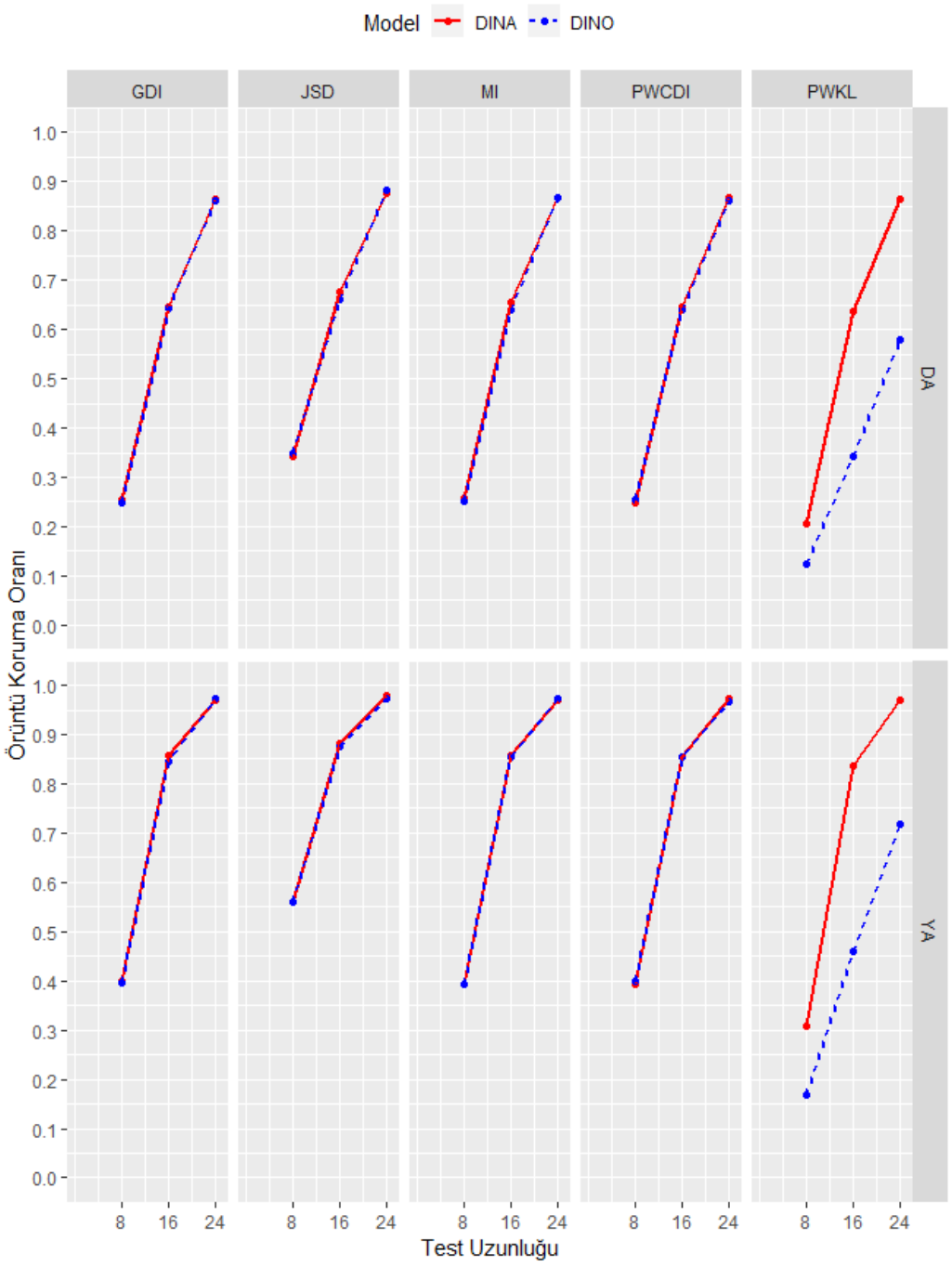
Şekil 4. *DINO Modele Göre Madde Seçim Algoritmalarının Nitelik Koruma Oranları*

DINA ve DINO modele göre madde seçim algoritmalarının örüntü koruma oranlarının incelenmesi

DINA ve DINO modeller için farklı madde kalite düzeylerinde ve test uzunluklarında, 5 nitelik düzeyine göre madde seçim algoritmalarının ÖKO değerleri Şekil 5'te ve 8 nitelik düzeyine göre madde seçim algoritmalarının ÖKO değerleri Şekil 6'da gösterilmiştir.



Şekil 5. *Nitelik sayısı 5 olduğunda DINA ve DINO modele göre madde seçim algoritmalarının ÖKO değerleri*



Şekil 6. *Nitelik sayısı 8 olduğunda DINA ve DINO modele göre madde seçim algoritmalarının ÖKO değerleri*

Şekil 5 ve 6 incelendiğinde, tüm koşullarda PWKL algoritması dışındaki diğer algoritmalardan elde edilen ÖKO değerlerinin hem DINA hem de DINO modellerinde yaklaşık olarak aynı olduğu görülmektedir. PWKL algoritmasından elde edilen ÖKO değenlerinin DINO modelde DINA modele göre daha düşük olduğu görülmektedir.

Tartışma

DINA modele göre madde seçim algoritmalarının ÖKO ve NKÖ değerleri incelendiğinde, algoritmaların ÖKO ve NKÖ değerleri, test uzunluğu ve madde kalitesi arttıkça artmakta, nitelik sayısı arttıkça azalmaktadır. Elde edilen bu bulgular, Cheng (2009), Wang (2013), Kaplan ve diğerleri (2015), Zheng ve Chang (2016), Yiğit ve diğerleri (2019) tarafından elde edilen bulgularla paralellik göstermektedir. Test uzunluğu 16 olduğunda, 5 nitelik sayısında ve yüksek madde kalite düzeyinde, madde seçim algoritmalarının ÖKO değerlerinin 1'e yakın olduğundan ölçme doğruluğu açısından 5 nitelik düzeyi için 16 test uzunluğunda tüm algoritmalar tercih edilebilir. Düşük madde kalitesinde ise 16 test uzunluğunda madde seçim algoritmalarından yüksek ölçme doğruluğu elde edilse de bu değerlerin yüksek madde kalitesine göre daha düşük olduğu görülmüştür. Bu açıdan düşük kalitede maddelerden oluşan banka kullanıldığında daha doğru sınıflama yapılabilmesi için daha uzun testlerin kullanılması gerekmektedir. Nitelik sayısı 8 olduğunda madde seçim algoritmalarının ÖKO değerleri azalmaktadır. Özellikle düşük ayırt edicilikteki maddelerden oluşan bankalarda, 24 test uzunluğunda bile en yüksek ÖKO değeri 0,876 olarak bulunmuştur. Bu açıdan bakıldığında, nitelik sayısı arttığında, madde bankasında yer alan maddelerin ayırt edicilik güçlerinin de yüksek olması gerekmektedir. 8 nitelik sayısında, yüksek madde kalitesinde ve 24 test uzunluğu sonlandırma kuralında, madde seçim algoritmalarından elde edilen ÖKO değerleri yüksektir. Bu açıdan 8 nitelik koşulunda, uzun testlerin ve yüksek ayırt edicilikte maddelerden oluşan bankaların kullanılması ölçme doğruluğunu artıracaktır. Xu ve diğerleri (2003) DINA modele göre 5 ve 8 nitelik düzeyinde madde seçim algoritmalarının örüntü koruma oranlarını incelediği çalışmada nitelik sayısı arttıkça algoritmaların örüntü koruma oranlarının azaldığını belirtmiştir. Wang (2013) çalışmasında, 5 ve 8 nitelik düzeyinde madde seçim algoritmalarını, nitelik ve örüntü koruma oranları açısından değerlendirilmiş ve nitelik sayısının artırılmasının, nitelik ve örüntü koruma oranlarını azalttığını ifade etmiştir. Xu ve diğerleri (2003) ve Wang (2013) tarafından yapılan çalışmalardan elde edilen bulgularla, bu çalışmadan elde edilen bulgular benzerlik göstermektedir. DINO modele göre yapılan simülasyon çalışmasından da benzer sonuçlar elde edilmiştir.

DINA ve DINO model için gerçekleştirilen simülasyon çalışmasında tüm koşullarda, JSD algoritmasının ölçme doğruluğunun, diğer algoritmalardan elde edilen ölçme doğruluğundan az da olsa daha yüksek olduğu görülmüştür. Wang (2013) kısa testler için, MI algoritmasını, SHE, PWKL ve KL algoritmalarıyla karşılaştırdığı çalışmasında, MI algoritmasının NKO ve ÖKO değerlerinin PWKL algoritmasından elde edilen değerlerden

daha yüksek olduğunu belirtmiştir. Yiğit ve diğerleri (2019), MC-DINA model kullanarak yaptıkları çalışmada, JSD algoritmasının ölçme doğruluğunun, GDI ve PWKL algoritmalarından elde edilen ölçme doğruluğundan daha yüksek olduğunu belirtmiştir. Zheng ve Chang (2016), 5 nitelik düzeyinde, DINA modele göre yaptıkları çalışmada, PWCDI algoritmasından elde edilen ölçme doğruluk değerlerinin PWKL algoritmasından daha yüksek olduğunu, kısa testlerde MI algoritmasının ölçme doğruluk değerlerinin bu algoritmalardan daha yüksek uzun testlerde ise bu algoritmalarının ölçme doğruluklarının benzer olduğunu belirtmiştir. Kaplan ve diğerleri (2015) yaptıkları çalışmada, GDI algoritmasından, PWKL algoritmasına göre daha yüksek ölçme doğruluk değerleri elde edildiğini belirtmiştir. Bu çalışmadan elde edilen bulguların, Wang (2013), Kaplan ve diğerleri (2015), Zheng ve Chang (2016) ve Yiğit ve diğerleri (2019) tarafından elde edilen bulgularla örtüştüğü söylenebilir.

Madde seçim algoritmalarının performansları kullanılan bilişsel tanı modeline göre değerlendirildiğinde, PWKL algoritması dışındaki algoritmaların ölçme doğruluk değerleri her iki model için birbirlerine çok yakındır. Ancak DINO modele göre gerçekleştirilen çalışmada, 8 nitelik sayısında, PWKL algoritmasından elde edilen ölçme doğruluk değeri, DINA modelden elde edilene değerlere göre önemli ölçüde düşüktür. Bu durumda, DINO modelde, nitelik sayısının yüksek olduğu durumlarda, PWKL algoritmasının ölçme doğruluğu açsından performansının, diğer algoritmalara göre önemli ölçüde daha düşük olduğu söylenebilir.

Sonuç

Bu çalışmanın amacı, DINA ve DINO modelde, farklı madde kalitesi, farklı nitelik sayılarında ve farklı sabit test uzunluğu sonlandırma kuralı düzeylerinde, madde seçim algoritmalarının performanslarını ölçme doğruluğu ölçütüne göre incelemektir.

Çalışmada, madde bankasında yer alan maddelerin ayırt edicilik güçleri arttıkça, madde seçim algoritmalarının ölçme doğruluğunun önemli ölçüde arttığı sonucuna ulaşılmıştır. Ayrıca, test uzunluğu arttıkça algoritmaların ölçme doğrulukları artarken, nitelik sayısının artmasıyla algoritmaların ölçme doğrulukları azalmaktadır. 8 nitelik sayısında, 24 test uzunluğunda ve yüksek ayırt edici maddelerden oluşan bankaların olduğu çalışmalarda, tüm madde seçim algoritmaları kullanılabilir. Nitelik sayısı 5 olduğunda ise, yüksek madde kalitesinde, 16 ve 24 test uzunluklarında, algoritmaların ölçme doğrulukları 1'e yaklaştığından dolayı algoritmaların kullanılması önerilmektedir. Tüm koşullarda en yüksek ölçme doğruluğu değerleri JSD algoritmasından elde edilmiştir. PWKL algoritmasından elde edilen ölçme doğruluk değerleri ise az da olsa diğer algoritmalara göre daha düşüktür. Çalışmada,

PWKL algoritması dışındaki algoritmaların performansları ise tüm koşullarda yaklaşık aynıdır. MI algoritmasından, çok az da olsa GDI ve PWCDI algoritmasına göre daha yüksek ölçme doğruluk değerleri elde edilmiştir. Çalışmadan elde edilen bir başka sonuç ise, PWKL algoritması dışında diğer madde seçim algoritmalarının performansları DINA ve DINO modeller için benzerdir. Özellikle DINO modelde ve nitelik sayısı yüksek olduğunda PWKL algoritmasının kullanılması önerilmemektedir.

Bu çalışmada madde kalitesi, nitelik sayısı ve sonlandırma kuralı olarak sabit test uzunluğu ele alınmıştır. Benzer bir çalışma, sonlandırma kuralı olarak değişken test uzunluğu kullanılarak da gerçekleştirilebilir. Çalışma kapsamında tamamlayıcı olmayan ve tamamlayıcı modellerden DINA ve DINO modeller kullanılmıştır. Her iki model kısıtlayıcı modellerdir. Benzer çalışmalar genelleştirilmiş modeller (örneğin GDINA, GDM) kullanılarak da gerçekleştirilebilir. Bu çalışmada madde seçim algoritmaları ölçme doğrulukları (ÖKO ve NKO) açısından değerlendirilmiştir. Benzer bir çalışma hesaplama süresi, madde kullanım sıklığı vb. değerlendirme ölçütleri kullanılarak, algoritmaların performanslarının daha bütüncül değerlendirilmesi için gerçekleştirilebilir.

Etik Kurul İzin Bilgisi: *Bu araştırma simulatif bir çalışma olduğundan, "Etik Kurul İzni gerektiren" araştırmalar arasında yer almamaktadır.*

Yazar Çıkar Çatışması Bilgisi: *Yazarlar araştırma, yazarlık ve/veya bu makalenin yayınlanmasıyla ilgili olarak herhangi bir çıkar çatışması beyan etmemiştir.*

Yazar Katkısı: *Tüm aşamalarda yazarlar ortak katkı sunmuştur.*

Kaynakça

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

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement, 33,* 163–183.

DiBello, L., Roussos, L. A., & Stout, W. F. (2007). Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models. C. R. Rao, & S. Sinharay (Eds). *Handbook of Statistics*. 26 (pp. 979-1030). Amsterdam, The Netherlands: Elsevier

Embretson, S.E. (2001). *The second century of ability testing: Some predictions and speculations*. http://www.ets.org/Media/Research/pdf/PICANG7.pdf.

Haertel, E.H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement*, 8, 333–346.

Hsu, C. L., & Wang, W. C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *Journal of Educational Measurement*, 52, 125–143.

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, 25(3), 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.

Ma. W., & de la Torre, J. (2020). GDINA: An R Package for Cognitive Diagnosis Modeling. *Journal of Statistical Software*, 93(14), 1-26. doi: 10.18637/jss.v093.i14.

Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence*. Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rupp, A. A., Henson, R. A., & Templin, J. L. (2010) *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press

Stocking, M.L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Rep. No. 94-5). Princeton: Educational Testing Service.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society*, 51, 337–350.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of Royal Statistics*, 65, 143–157.

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

Thissen, D., & Mislevy, R. J. (2000). Computerized Adaptive Testing: A primer. H. Wainer, (Ed). *Testing algorithms*, Mahwah, NH: Lawrence Erlbaum Associates, Inc, s. 101-133.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report 05-16). Princeton, NJ: Educational Testing Service.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length. *Educational and Psychological Measurement*, 73(6), 1017–1035.

Wickham, H. (2016). ggplot2: *Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Xu, X., Chang, H., & Douglas, J. (2003). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized Adaptive Testing for Cognitively Based Multiple-Choice Data. *Applied Psychological Measurement*, 43(5), 388–401.

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, 40, 608-624.

Research Article

Pamukkale University Journal of Education 54: 188-214 [2022]
doi: 109779.pauefd.769548

Investigating the Performance of Item Selection Algorithms in terms of Measurement Accuracy in CD-CAT

Semih AŞİRET* Seçil ÖMÜR SÜNBÜL**

• ***Received:*** *15.07.2020* • ***Accepted:*** *23.08.2021* • ***Online First:*** *19.11.2021*

This study aims to examine the performance of item selection algorithms according to the accuracy of measuring through different number of attributes, item quality, and test lengths for DINA and DINO models in cognitive diagnostic computerized adaptive testing (CD-CAT). Within the scope of the study, the number of attributes was manipulated as 5 and 8, and each item was limited to measure at least one attribute and at most four attributes. In data generation, the g and s parameters were drawn from the uniform distribution of U(0.05-0.25) for high item quality level and U (0.10-0.30) for low item quality level. Cognitive patterns of 3000 examinees were generated so that each examinee had a 50% chance of having each attribute. Fixed test lengths of 8, 16, and 24 were used as termination rules. GDI, JSD, MI, PWCDI and PWKL were used as item selection algorithms in the study. The performances of item selection algorithms were evaluated according to their attribute and pattern recovery rates. Data generation and analysis in the study were carried out using R 3.6.3 software. As a result of the study, it was determined that the measurement accuracy values of all algorithms increased as the item quality and test length increased, and the measurement accuracy decreased as the number of attributes increased. It was found that the measurement accuracy of the JSD algorithm was the highest in all conditions, while the PWKL algorithm was the lowest. While the performance of the algorithms except for the PWKL algorithm in DINA and DINO models was approximately the same, it was found that the measurement accuracy of the PWKL algorithm in the DINO model was lower than that of the DINA model.

Keywords: Cognitive diagnostic model, Computerized adaptive test, Item selection algorithms, DINA model, DINO model

Cited:

Aşiret, S. & Ömür-Sünbül, S. (2022). Investigating the performance of item selection algorithms in terms of measurement accuracy in CD-CAT. *Pamukkale University Journal of Education. 54, 188-214,* doi:109779.pauefd.769548

* Teacher, Republic of Turkey Ministry of National Education, ORCID ID: 0000-0002-0577-2603, semihasiret@gmail.com

** Assoc. Prof., Mersin University, Faculty of Education , ORCID ID: 0000-0001-9442-1516, secilomur@gmail.com

Introduction

In education, cognitive diagnostic assessments are model-based and based on formative assessment (Embretson, 2001). In recent years, many cognitive diagnostic models have been developed. The primary purpose of cognitive diagnostic assessment is not to make a summative assessment. Its main purpose is to identify the examinee's strengths and weaknesses in detail, provide effective and descriptive feedback to the examinees, reveal the learning profiles of the examinees, and facilitate their learning of them.

The cognitive diagnostic model (CDM) is a discrete latent variable model that allows diagnosing the steps required to solve the problem in a test or the presence or absence of many skills (de la Torre, 2009). The purpose of CDM is to reveal the skills that the examinee has and does not have. Unlike the item response theory (IRT), CDM expresses attributes that examinees have mastered or not with a cognitive pattern consisting of 1-0. There are many cognitive diagnostic models, and these models are classified in three different ways: complementary, non-complementary, and general. In complementary models, the examinee must have mastered at least one of the item's attributes to respond to the item correctly. In non-complementary models, on the other hand, the examinee must have mastered all attributes measured by the item to respond to the item correctly.

Firstly, in cognitive diagnostic assessments, attributes are determined, and items are written according to these attributes. After the items are written, a Q matrix is yielded. In the Q matrix, the rows contain the items and the columns the attributes. Cells corresponding to the attributes that the item measures are coded 1 and 0 otherwise. Thus, the attributes that each item measures are determined.

DINA (deterministic-input, noisy-and-gate) (Haertel, 1989; Junker & Sijtsma, 2001) and DINO (deterministic-input, noisy-or-gate) (Templin & Henson, 2006) are common and well known restrictive cognitive diagnostic models. The DINA is a non-complementary model and has a conjunctive condensation rule. Namely, an examinee must mastered all the attributes measured by the item in the Q matrix to answer that item correctly. If the examinee does not have mastered any of the attributes required in the Q matrix, it is assumed that the examinee will answer the item incorrectly. In the DINA model, two different parameters, guessing (g) and slipping (s), are estimated for each item. The parameter g indicates the probability of responding correctly to the item when the examinee does not have mastered all the attributes measured by the item, and the parameter s indicates the probability of the examinee giving an

incorrect answer to the item when they have all the attributes measured by the item defined in the Q matrix (Rupp, Templin, & Henson., 2010).

In the DINA model, the probability that examinee i responds the item j correctly is given in Equation 1.

$$\pi_{ij} = P\left(X_{ij} = 1 \middle| \alpha_i\right) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} \quad (1)$$

In Equation 1, α_i refers to a cognitive pattern of examinee i, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ refers latent response pattern and g is guessing parameter, s is slipping parameter. To illustrate, when $\eta_{ij} = 1$, the probability of the examinee answering the item correctly is given in Equation 2.

$$(\pi_{ij}) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} = (1 - s_j)^1 g_j^{1-1} = (1 - s_j) \quad (2)$$

The DINO model is complementary of the DINA model. The DINO model has a disjunctive condensation rule that indicates whether at least one required attribute has been mastered (Rupp et al., 2010). If the examinee has mastered any required attribute measured by the item defined in the Q matrix, the examinee will answer the item correctly. If the examinee has not mastered all of the required attributes, it is assumed that the examinee will answer the item incorrectly. In other words, mastering any required attributes completes the deficiency of the other required attributes. In the DINO model, just like the DINA model, two parameters are estimated for each item: guessing and slipping.

In the DINO model, the probability that examinee i responds the item j correctly is given in Equation 3.

$$\pi_{ij} = P\left(X_{ij} = 1 \middle| \alpha_i\right) = (1 - s_j)^{\omega_{ij}} g_j^{1-\omega_{ij}} \quad (3)$$

In Equation 3, α_i, represents the cognitive pattern of the examinee, $\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{qjk}$, latent response pattern, g, guessing parameter, and s, slipping parameter.

Cognitive Diagnostic Computerized Adaptive Testing

CD-CAT is a combination of cognitive diagnosis (CD) and computerized adaptive testing (CAT). While examinees are located at a point on the latent continuum in CAT, diagnostic feedback is not given to examinees. On the contrary, CD-CAT aims to classify examinees according to their latent states and to employ latent class models on these latent classes (Cheng, 2009).

CD-CAT has similar stages to CAT. As in CAT, Firstly, the first item is selected. The selected item is applied to examinees, and their cognitive patterns are estimated. The next item is selected from the item bank through item selection algorithms for the estimated cognitive pattern for each examinee and applied. This process continues until the termination rule is satisfied. After the termination rule is satisfied, the examinees' final cognitive patterns are estimated, and CD-CAT ends.

In traditional CAT, Maximum Fisher Information (MFI) is one of the most popular item selection algorithms (Thissen & Mislevy, 2000). However, MFI cannot be applied in CD-CAT because of being affected by chance success, insufficient for ability estimation in short tests, and cannot be used in discrete latent classifications. In the literature, Kullback-Leibler Information (Xu, Chang & Douglas, 2003), Shannon Entropy (Tatsuoka, 2002; Tatsuoka & Ferguson, 2003), Posterior Weighted Kullback-Leibler Information (PWKL) and Hybrid Kullback-Leibler Information (HKL) (Cheng, 2009), Mutual Information (MI) (Wang, 2013), Modified Posterior Weighted Kullback-Leibler Information (MPWKL) and GDINA Discrimination Index (GDI) (Kaplan, de la Torre & Barrada, 2015), Posterior Weighted Cognitive Discrimination Index (PWCDI) and Posterior Weighted Attribute-Level Cognitive Discrimination Index (PWACDI) (Zheng & Chang, 2016) and Jensen-Shannon Divergence (JSD) (Minchen & de la Torre, 2016) algorithms seem to be applied in single-purpose CD-CAT for item selection. The following section briefly explains the item selection algorithms used in this study.

Posterior Weighted Kullback-Leibler Information (PWKL)

Due to the low efficiency of the KL algorithm, Cheng (2009) developed the PWKL algorithm by multiplying the KL information with the posterior weight corresponding to this information to quantify the contribution of each cognitive pattern to the KL algorithm. The PWKL algorithm is mathematically given in Equation 4. Equation 4 is

$$PWKL_j(\widehat{\alpha}_\iota) = \sum_{c=1}^{2^K} \left\{ \sum_{x=0}^{1} \left[\log \left(\frac{P\left(X_{ij} = x \middle| \widehat{\alpha}_\iota\right)}{P\left(X_{ij} = x \middle| \alpha_c\right)} \right) P\left(X_{ij} = x \middle| \widehat{\alpha}_\iota\right) \right] \pi(\alpha_c | x_{t-1}) \right\} \quad (4)$$

where K is the number of the total attribute, $P\left(X_{ij} = x \middle| \alpha_c\right)$ is probability value of response x of examinee i to item j for cognitive pattern α_c and $\pi(\alpha_c|x_{t-1})$, is the posterior weight of cognitive patterns after (t-1) items are applied.

Mutual Information (MI) Index

The MI algorithm developed by Wang (2013) is defined as the equivalent of the KL distance between successive posterior distributions. The equation for the MI algorithm is given in Equation 5.

$$MI_{ij} = \sum_{c=1}^{2^K} \pi_i(\alpha_c|x_{t-1}) \sum_{x=0}^{1} P\left(X_{ij} = x|\alpha_c\right) \log\left(\frac{P\left(X_{ij} = x|\alpha_c\right)}{P(X_{ij}=x)}\right) \quad (5)$$

Jensen-Shannon Divergence (JSD)

The JSD (Minchen & de la Torre, 2016) measures the relative entropy between the joint distribution of two random distributions and the product of their marginal distributions (Yiğit, Sorrel, & de la Torre, 2019). The item with the maximum JSD value is selected as the next item. The equation for the JSD algorithm is given in Equation 6.

$$JSD_j = S\left(P_j \times \pi'\right) - \sum_{c}^{2^K} \pi_c S\left(P_{jc}\right) \quad (6)$$

In Equation 6, H refers to the number of options,$S\left(P_{jc}\right)$, Shannon entropy, P_j, $H \times 2^K$ matrix and π, posterior probability weight.

Posterior Weighted Cognitive Discrimination Index (PWCDI)

Zheng and Chang (2016) developed the PWCDI algorithm by including posterior probability distributions of cognitive patterns in the cognitive discrimination index (CDI). The KL information among the response distributions of possible cognitive patterns is stored in D matrix, which dimension is the 2^Kx2^K. In the PWCDI algorithm, the posterior probability distributions of the cognitive patterns are included in the D matrix to obtain the PWD matrix. In this respect, it is similar to the PWKL algorithm, but unlike the PWKL algorithm, rows and columns in the matrix are weighted. The PWD matrix is given by Equation 7.

$$PWD_{juv} = E_{\alpha_u}\left[\pi(\alpha_u) \times \pi(\alpha_v) \times log\left(\frac{P\left(X_j|\alpha_u\right)}{P\left(X_j|\alpha_v\right)}\right)\right] \quad (7)$$

PWCDI algorithm is presented by Zheng and Chang (2016) in Equation 8 as;

$$PWCDI_j = \frac{1}{\sum_{u\neq v} h(\alpha_u,\alpha_v)^{-1}} \sum_{u\neq v} h(\alpha_u, \alpha_v)^{-1} PWD_{juv} \quad (8)$$

where $h(\alpha_u, \alpha_v) = \sum_{k=1}^{K} |\alpha_{uk} - \alpha_{vk}|$ is hamming distance between two cognitive pattern.

Purpose and Significance of the Study

The key case in CD-CAT studies is item selection algorithms (Cheng, 2009). Hsu and Wang (2015) stated that good item selection increases measurement accuracy in CD-CAT. Many new item selection algorithms have been developed for CD-CAT in recent years. However, there are not enough studies evaluating the performance of these item selection algorithms in different models and under different conditions in terms of measurement accuracy.

The purpose of this study is to analyze the performance of item selection algorithms (GDI, JSD, MI, PWCDI, PWKL) for DINA and DINO models at various test lengths (8, 16, and 24), item quality levels (low and high) and several attributes (5 and 8) to evaluate their performance in terms of measurement accuracy (pattern recovery rate and attribute recovery rate). For this purpose, it is thought to help practical applications by revealing which item selection algorithms obtained the highest measurement accuracy in the conditions determined.

Method

Research Model

This study aims to examine the performance of item selection algorithms in CD-CAT in terms of measurement accuracy according to the model, number of attributes, item quality, and test length. In this respect, this study is theoretical research since it is aimed to examine the pattern and attribute recovery rates of item selection algorithms according to various factors.

Factors Manipulated in the Research

Analysis Model: The DINO model was used as a complementary CDM in the study, and the DINA model was used as a non-complementary CDM. The main reasons for choosing both models are complementary and non-complementary models; they are frequently preferred in practice and easy to calculate. Ease of computation is a desirable feature in adaptive testing.

Item quality: In the study, item quality was manipulated as low and high. The parameters used by Zheng and Chang (2016) were used to generate the item parameters. For both models, the g and s parameters were generated from a uniform distribution of U(0.05-0.25) for high item quality and U(0.10-0.30) for low item quality.

The number of attributes: In the literature, Cheng (2009) and Wang (2013) stated that 5 attributes are medium, and von Davier (2005) stated that the number of attributes should be 8 at most in practice. In this study, the number of attributes was manipulated as 5 and 8, as

medium and high. In addition, each item was limited to measuring a maximum of 4 attributes to provide similarity with real applications.

Test length: DiBello, Roussos, & Stout (2007) stated that CD-CAT studies are frequently used for formative evaluation in classroom assessments, and test lengths should be short. In this respect, the test lengths were manipulated as 8, 16, and 24 by considering the number of attributes.

Item Selection Algorithms: PWKL (Cheng, 2009), MI (Wang, 2013), GDI (Kaplan et al., 2015), PWCDI (Zheng & Chang, 2016), and JSD (Minchen & de la Torre, 2016) item selection algorithms were used.

Data Generation and Data Analysis

Generation and analysis of data in the study was carried out using R 3.6.3 (R Core Team, 2020) software. The GDINA package (v2.8; Ma & de la Torre, 2020) was used to generate the data, and the ggplot2 (v3.3.2; Wickham, 2016) package was used to generate the graphs. Other codes were written by the researchers in R 3.6.3 (R Core Team, 2020) software.

Item bank and generation cognitive pattern of examinees: Stocking (1994) reported that the item bank should be at least 12 times the test length. Therefore, two different item banks consisting of 500 items were generated for five and eight attributes. The Q matrix was constructed as each attribute has a 30% chance of being measured by the item. The Q matrix was generated item by item and attribute by attribute. In addition, each item in the Q matrix was limited to measure at least one attribute and at most four attributes to ensure similarity to real applications. Thus, there are 30 different cognitive patterns for five attributes and 162 different cognitive patterns for eight attributes in the Q matrix. Cognitive patterns of 3000 examinees were generated for the five-attributes and eight-attributes test separately, with each examinee having a 50% probability of achieving each attribute. 1-0 data were generated in respect to generated cognitive patterns of the examinees and the Q matrix. Then, the probability of answering the items correctly was calculated for each cognitive pattern for the DINA and DINO models.

In Table 1, the number of items measuring each attribute and the number of examinees with the attribute is given for five and eight attributes. As can be seen from Table 1, the number of items measuring each attribute is approximately equal since the Q matrix is composed of the item by item and the attribute by attribute. As is shown in Table 1, for K=5, the number of items measuring the first attribute is 169, the second attribute is 179, the third attribute is

185, the fourth attribute is 189, and the fifth attribute is 182. A similar distribution is valid for K=8. From Table 1 it can be seen that the number of examinees at each attribute is approximately equal for K=5 and K=8. This is because each examinee is limited to a 50% probability of achieving each attribute in the generation of examinees.

Table 1. *The number of items measures each attribute, and examinees have mastered each attribute for K=5 and K=8*

	Attributes							
K=5	1	2	3	4	5			
Number of Items (J=500)	169	179	185	189	182			
Number of Examinees (N=3000)	1493	1543	1492	1495	1499			
K=8	1	2	3	4	5	6	7	8
Number of Items (J=500)	156	133	135	161	146	153	154	170
Number of Examinees (N=3000)	1493	1503	1524	1547	1475	1494	1450	1509

Note: K, number of attributes, J, number of items in item bank, N, number of examinees

In Table 2, the distribution of the items measuring the number of possible attributes and the examinees for 5 and 8 attributes are given. In the item bank, since each item is limited to measure at least one attribute and at most four attributes, no item measures more than four attributes. When Table 2 is examined, it is seen that items are measuring one (204) and two attributes (203) at the most at the level of five attributes, while the number of items measuring three attributes is 78 and the number of items measuring four attributes is 15. Some items measure at most two attributes (168) for K=8. The number of items measuring one attribute is 105, the number of items measuring three attributes is 141, and the number of items measuring four attributes is 86. While the number of examinees with no attribution for K=5 is 97, it is 6 for K=8.

Table 2. *Distribution of examinees with items and attributes measuring the number of possible attributes at the 5 and 8 attribute levels*

	Attributes								
Attribute Number (K=5)	0	1	2	3	4	5			
Item Number (J=500)	0	204	203	78	15	0			
Examinee Number (N=3000)	97	449	942	955	461	96			
Attribute Number (K=8)	0	1	2	3	4	5	6	7	8
Item Number (J=500)	0	105	168	141	86	0	0	0	0
Examinee Number (N=3000)	6	86	334	665	834	658	312	92	13

Note: K, number of attributes, J, number of items in item bank, N, number of examinees

First item selection: CD-CAT application starts with the first item selected. In this study, the first item was selected randomly, and the randomly selected item was used as the first item in all algorithms to evaluate the item selection algorithms under equal conditions.

Estimation of cognitive pattern: CD-CAT applications are frequently used in classroom assessments. The length of the tests administered during the course is usually short. In this case, the probability of examinees who respond correctly (1) to all items or incorrectly (0) to all items may be high. When the response patterns of examinees are all 0 or 1, the maximum likelihood estimation (MLE) method cannot make an accurate prediction. For this reason, the cognitive patterns of examinees were estimated by the maximum a posteriori (MAP) estimation method in the study.

Evaluation criteria: In this study, item selection algorithms' performance was evaluated at the levels of attribute and pattern. At the attribute level, the item selection algorithms' attribute recovery ratios (ARR) and at the cognitive pattern level, the pattern recovery rates (PRR) were calculated. ARR and PRR were calculated with Equation 9 and Equation 10.

$$NKO_k = \frac{\sum_{i=1}^{N} A_{ik}}{N} = \frac{\sum_{i=1}^{N} \left(I_{\widehat{\alpha_{ik}},\alpha_{ik}}\right)}{N}, \qquad (k=1, 2, \ldots., K) \qquad (9)$$

$$\ddot{O}KO_k = \frac{\sum_{i=1}^{N} R_i}{N} = \frac{\sum_{i=1}^{N} \left(I_{\widehat{\alpha_i},\alpha_i}\right)}{N}, \qquad (k=1, 2, \ldots., K) \qquad (10)$$

CD-CAT process: In the study, after generating the Q matrix, cognitive patterns of examinees, and responses of examinees to the items, the probability of each pattern responding correctly to the item was calculated. After this stage, the CD-CAT was started. In the CD-CAT, the randomly selected starting item was applied to all examinees, and the possible cognitive patterns of each examinee were estimated using the MAP method. Then, the selected item by the item selection algorithm for each examinee was applied as the next item. This process is repeated until the termination rule is satisfied. When CD-CAT was completed, the ARR and PRR of each item selection algorithm were calculated under all conditions. Tables and plots were created for these obtained values

Findings

Findings on the pattern recovery rates of item selection algorithms in the DINA Model, according to item quality and many attributes in fixed test length termination rule.

In Table 3, the pattern recovery rates of the item selection algorithms according to the DINA model are given at the factor levels in the research. In addition, these ratios are given graphically in Figure 1.

Table 3. *Pattern Recovery Rates of Item Selection Algorithms by DINA Model*

Model	K	Item quality	Number of Items	Item Selection Algorithms				
				GDI	JSD	MI	PWCDI	PWKL
DINA	5	Low	8	0,644	0,699	0,644	0,638	0,612
			16	0,909	0,921	0,917	0,914	0,910
			24	0,978	0,989	0,982	0,979	0,975
		High	8	0,808	0,855	0,814	0,799	0,773
			16	0,980	0,985	0,987	0,982	0,976
			24	0,999	0,999	0,998	0,997	0,999
	8	Low	8	0,254	0,343	0,256	0,248	0,207
			16	0,647	0,676	0,655	0,647	0,638
			24	0,866	0,876	0,869	0,867	0,864
		High	8	0,399	0,560	0,394	0,392	0,308
			16	0,857	0,881	0,859	0,853	0,836
			24	0,971	0,98	0,971	0,973	0,971

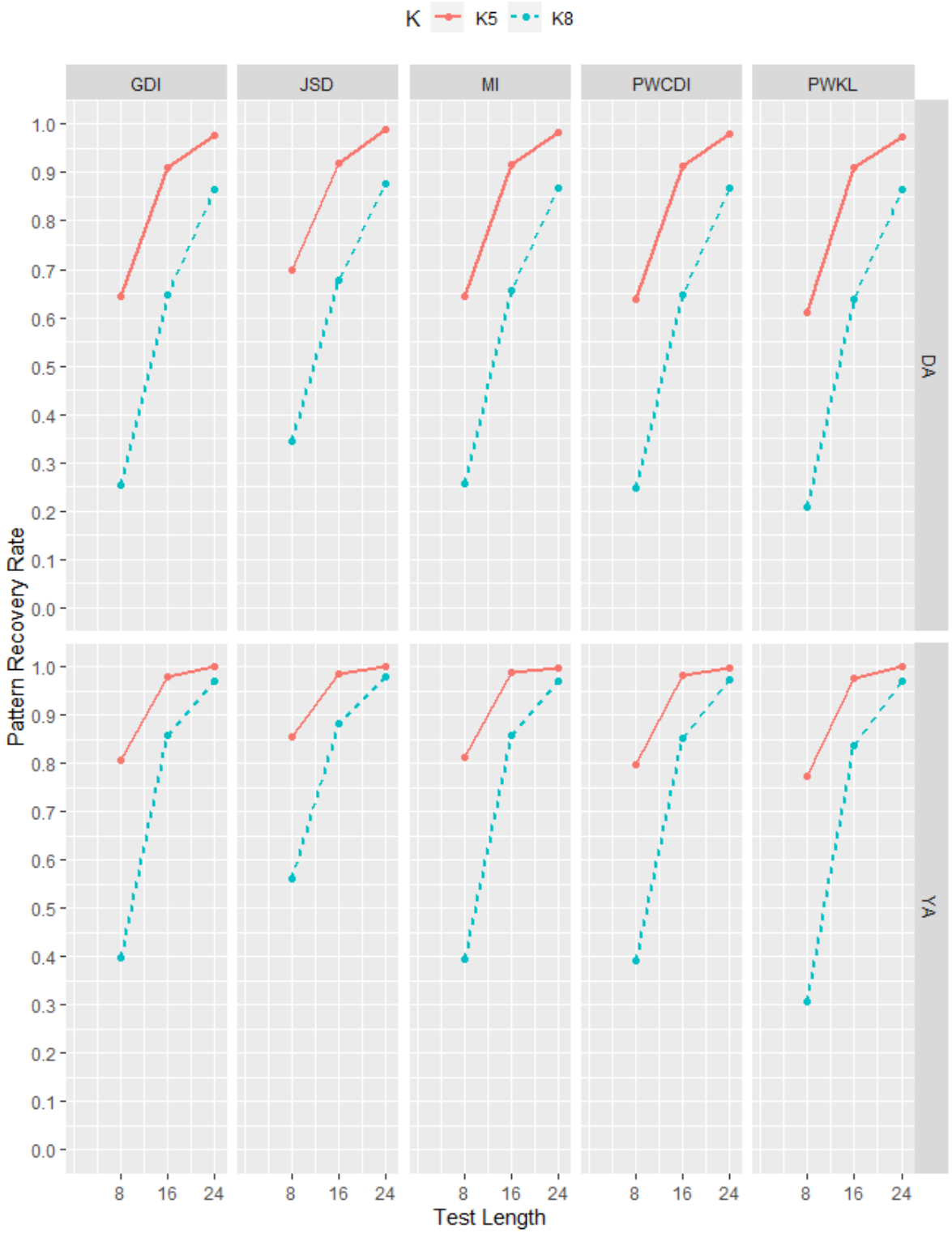


Figure 1. *Pattern Recovery Rates of Item Selection Algorithms by DINA Model*

When Table 3 and Figure 1 are examined together, it can be seen that the PRR values of the item selection algorithms decrease significantly as the number of attributes measured by the item increases. In all conditions, PRR values for K=5 are higher than for K=8. However,

in cases where the test termination rule is 24 items and the item quality is high, the PRR values of the algorithms at the 5 and 8 attribute levels draw near 1.

When 8 test lengths are used as the termination rule, the PRR of algorithms with low item quality varies between 0.612 - 0.699 for 5 attribute levels and 0.855 - 0.975 for high item quality.

As the test length increases, the PRR values of the algorithms increase at different item quality levels. The PRR values of the item selection algorithms are close to 1 in all low and high item quality levels, with 24 test lengths and 5 attribute levels. At the 8 attribute level, it can be said that the PRR values are close to 1 only when the item quality is high and the 24 test length termination rule. It can be seen in Figure 1 that the pattern recover PRR of the item selection algorithms is close to 1 when the item quality is high, and the test length is 16 at 5 attribute level. According to Table 3, it is seen that the highest PRR values were obtained from the JSD algorithm, and the lowest PRR values were obtained from the PWKL algorithm in all conditions. The PRR values of other algorithms are very close to each other.

Findings on the attribute recovery rates of item selection algorithms in the DINA Model, according to item quality and number of attributes, in fixed test length termination rule

In Table 4, the attribute recovery rates of the item selection algorithms according to the DINA model at the factor levels in the research. In addition, these ratios are given graphically in Figure 2.

When Table 4 and Figure 2 are examined together, it is seen that the ARR values of the item selection algorithms increase as the test length increases; the ARR values of the item selection algorithms are approximately 1.00 when the test length is 24 and the item quality is high while the test length is 24. The item quality is low. These values seem to be 1 at 5 attribute level and close to 1 at 8 attribute level. It can be said that the ARR values of the JSD algorithm are slightly higher than the other algorithms.

Table 4. *Attribute Recovery Rates of Item Selection Algorithms by DINA Model*

Model	K	Item Quality	Number of Items	Algorithms				
				GDI	JSD	MI	PWCDI	PWKL
DINA	5	Low	8	0,906	0,922	0,907	0,9	0,891
			16	0,978	0,982	0,98	0,978	0,978
			24	0,995	0,998	0,996	0,995	0,994
		High	8	0,954	0,964	0,955	0,948	0,939
			16	0,995	0,997	0,997	0,995	0,994
			24	1,00	1,00	1,00	0,999	1,00
	8	Low	8	0,826	0,846	0,825	0,812	0,819
			16	0,936	0,942	0,937	0,931	0,932
			24	0,978	0,98	0,978	0,977	0,976
		High	8	0,881	0,91	0,88	0,871	0,861
			16	0,977	0,981	0,978	0,975	0,972
			24	0,996	0,997	0,996	0,996	0,996

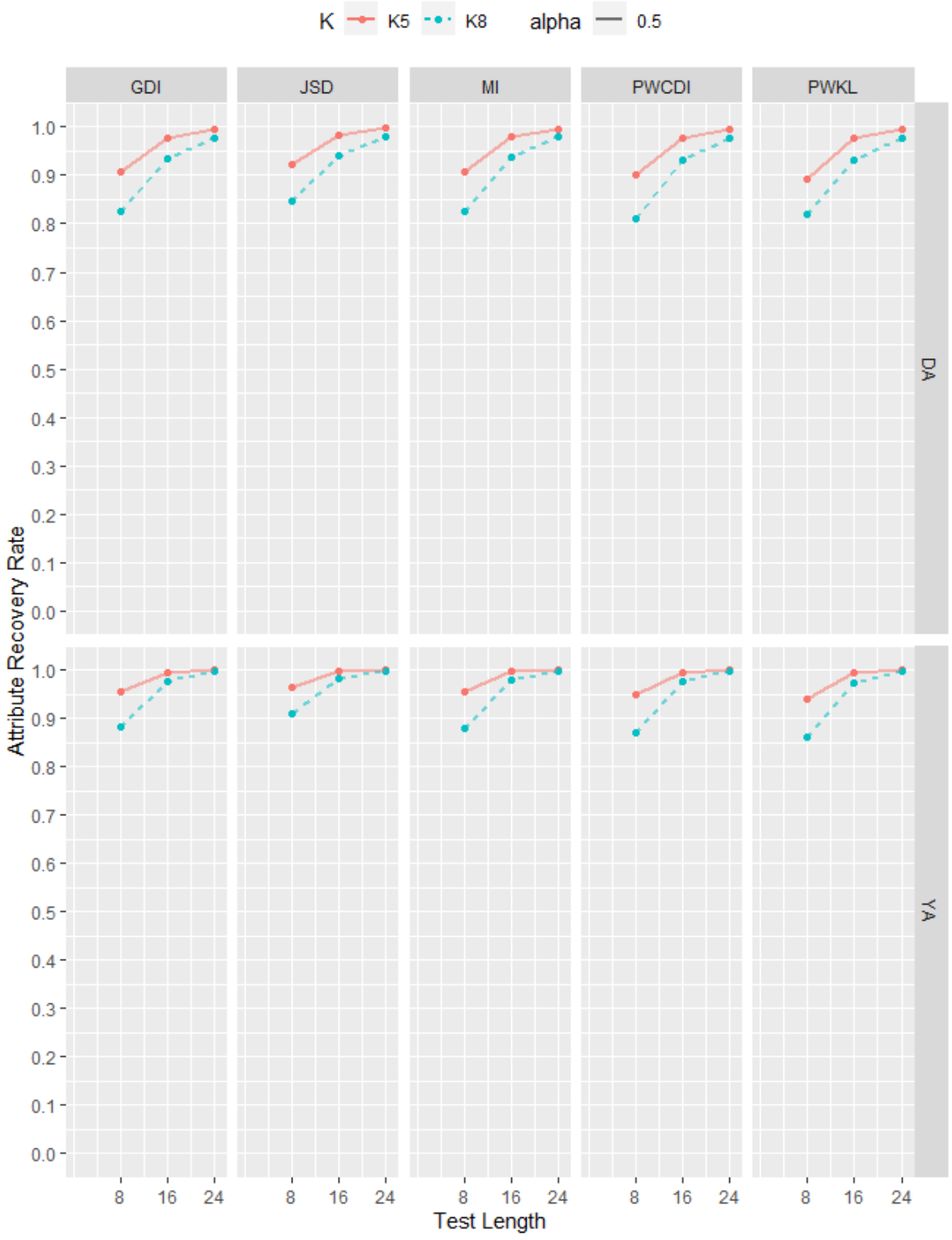


Figure 2. *Attribute Recovery Rates of Item Selection Algorithms by DINA Model*

Findings on the pattern recovery rates of item selection algorithms in the DINO Model, according to item quality and number of attributes, in fixed test length termination rule

In Table 5, the PRR of the item selection algorithms according to the DINO model at the factor levels in the research. In addition, these ratios are given graphically in Figure 3.

Table 5. *Pattern Recovery Rates of Item Selection Algorithms According to DINO Model*

Model	K	Item Quality	Number of Items	Algorithms				
				GDI	JSD	MI	PWCDI	PWKL
DINO	5	Low	8	0,644	0,689	0,645	0,637	0,511
			16	0,911	0,921	0,909	0,912	0,851
			24	0,982	0,984	0,983	0,981	0,968
		High	8	0,819	0,858	0,819	0,81	0,663
			16	0,986	0,988	0,987	0,984	0,962
			24	0,998	0,999	0,999	0,997	0,997
	8	Low	8	0,247	0,348	0,25	0,254	0,124
			16	0,642	0,66	0,64	0,641	0,341
			24	0,861	0,884	0,868	0,863	0,579
		High	8	0,396	0,56	0,394	0,398	0,167
			16	0,844	0,876	0,853	0,854	0,46
			24	0,972	0,973	0,973	0,966	0,719

When Figure 3 is examined, it can be said that as the test length and item quality increase, the PRR values of the item selection algorithms increase, as in the DINA model, and the PRR values of the item selection algorithms decrease when the number of attributes increases.

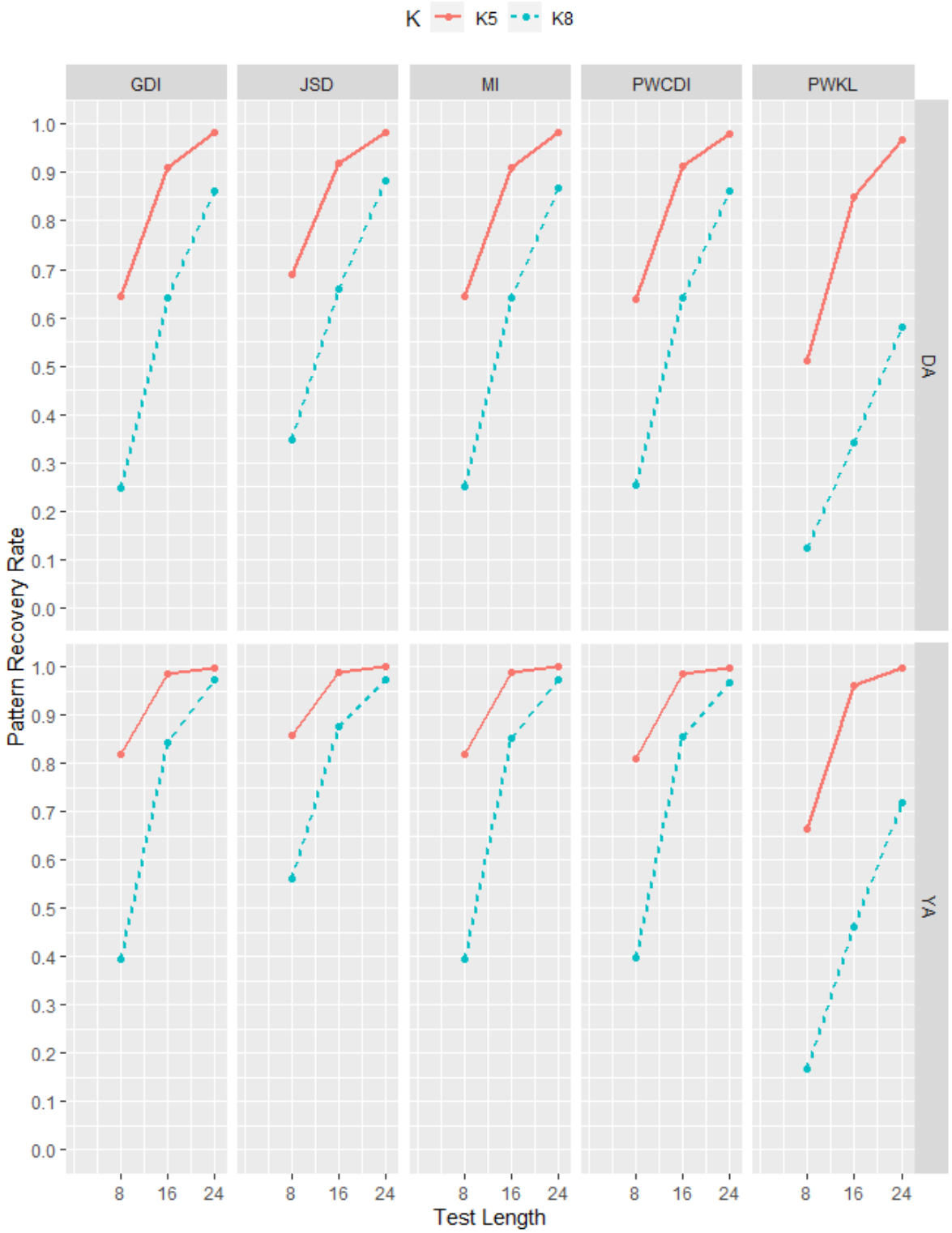


Figure 2. *Pattern Recovery Rates of Item Selection Algorithms According to DINO Model*

According to Table 5, the PRR values of the algorithms at 5 attribute levels, low item quality and 8 test lengths vary between 0.511 - 0.689 and between 0.663 - 0.858 for high item quality. At 8 attribute level, 8 test length, and low item quality, PRR values range from 0.124 to 0.348; at high item quality, it ranges from 0.167 to 0.56. When the termination rule is 24

test length, the PRR values of the algorithms at the 5 attribute level are close to 1. At the 8 attribute level and high item quality, the PRR values of other algorithms except the PWKL algorithm are close to 1, while these values vary between 0.589 and 0.884 at low item quality. The range of PRR of the item selection algorithms is so wide because the PWKL algorithm has significantly lower PRR values than other algorithms. Especially at the 8 attribute level, the PRR of the PWKL has decreased significantly compared to the 5 attribute level. In all conditions, the PRR of the JSD is higher, while the PRR of the PWKL is the lowest. The PRR of the algorithms other than the PWKL algorithm is close to the PRR of the JSD algorithm.

Findings on the attribute recovery ratios of item selection algorithms in fixed test length termination rule according to item quality and several attributes in the DINO model.

In Table 6, the attribute recovery rates of the item selection algorithms according to the DINO model at the factor levels in the research are given. In addition, these ratios are given graphically in Figure 4.

When Table 6 and Figure 4 are examined, it can be said that the ARR obtained in the DINO model are similar to the ARR obtained in the DINA model, except for the PWKL algorithm. While the ARR values of the algorithms with low item quality and 5 attribute levels are higher at all test lengths, it can be said that the ARR of the algorithms are close to 1 when the item quality increases, except for the PWKL algorithm with 24 test lengths. According to Figure 4, it can be said that with the increase in the number of attributes, the ARR of the PWKL algorithm decreases more than the ARR of other algorithms.

Table 6. *Attribute Recovery Rates of Item Selection Algorithms According to DINO Model*

Model	K	Item Quality	Number of Items	Algorithms				
				GDI	JSD	MI	PWCDI	PWKL
DINO	5	Low	8	0,906	0,921	0,906	0,9	0,821
			16	0,979	0,981	0,978	0,977	0,958
			24	0,996	0,996	0,996	0,995	0,992
		High	8	0,957	0,966	0,956	0,952	0,886

			16	0,997	0,997	0,997	0,996	0,99
			24	1	1	1	0,999	0,999
			8	0,814	0,846	0,813	0,812	0,622
		Low	16	0,933	0,938	0,933	0,931	0,743
			24	0,977	0,981	0,978	0,975	0,858
8								
			8	0,875	0,906	0,873	0,861	0,635
		High	16	0,975	0,98	0,976	0,974	0,797
			24	0,996	0,996	0,996	0,995	0,91

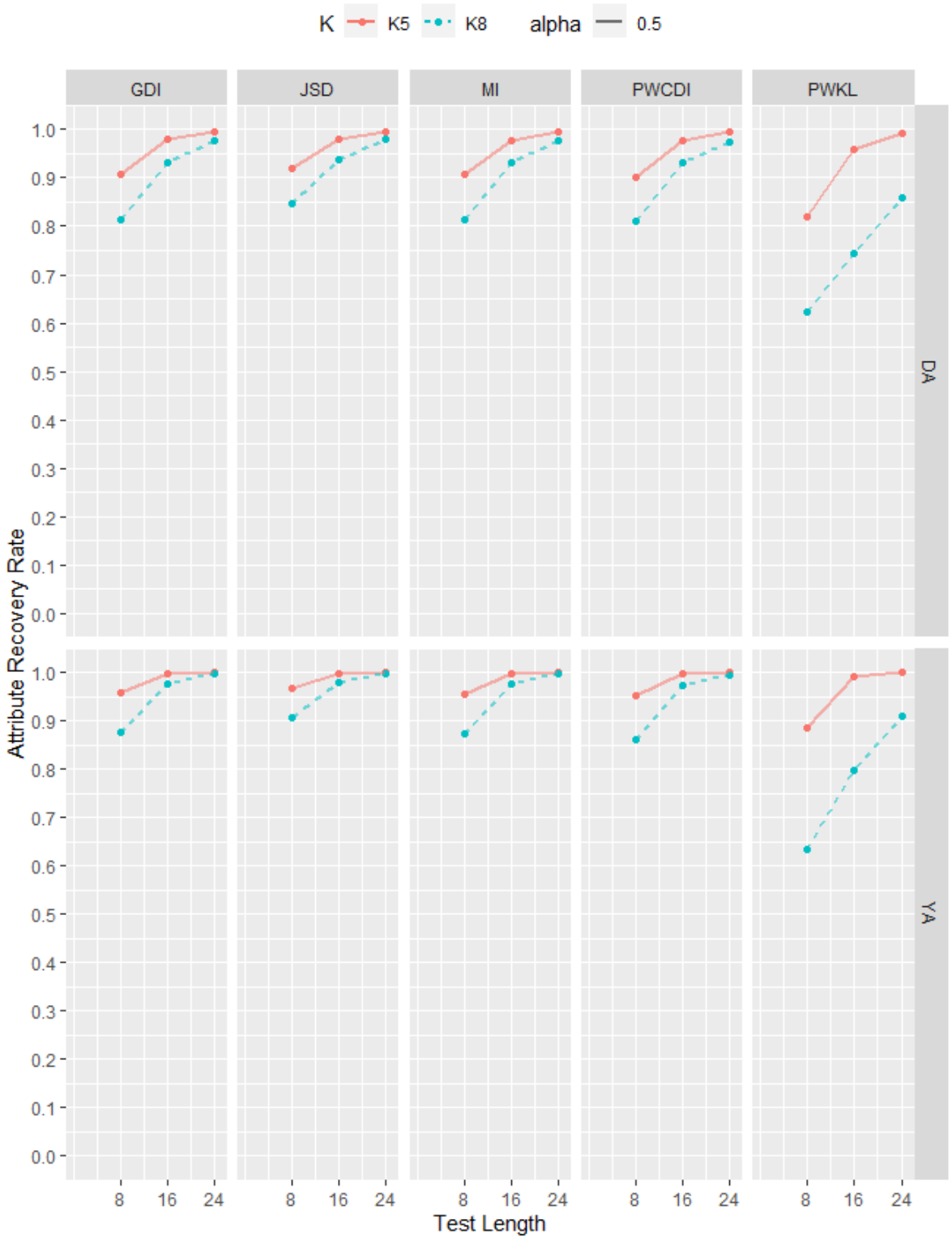


Figure 3. *Attribute Recovery Rates of Item Selection Algorithms According to DINO Model*

Analysis of pattern recovery rates of item selection algorithms according to DINA and DINO models

PRR of item selection algorithms according to 5 attribute levels at different item quality levels and test lengths for DINA and DINO models are shown in Figure 5, and PRR values of item selection algorithms according to 8 attribute levels are shown in Figure 6. When Figures 5 and 6 are examined, it is seen that the PRR values obtained from other algorithms except the PWKL algorithm are approximately the same in both DINA and DINO models under all conditions. It is seen that the PRR values obtained from the PWKL algorithm are lower in the DINO model than in the DINA model.

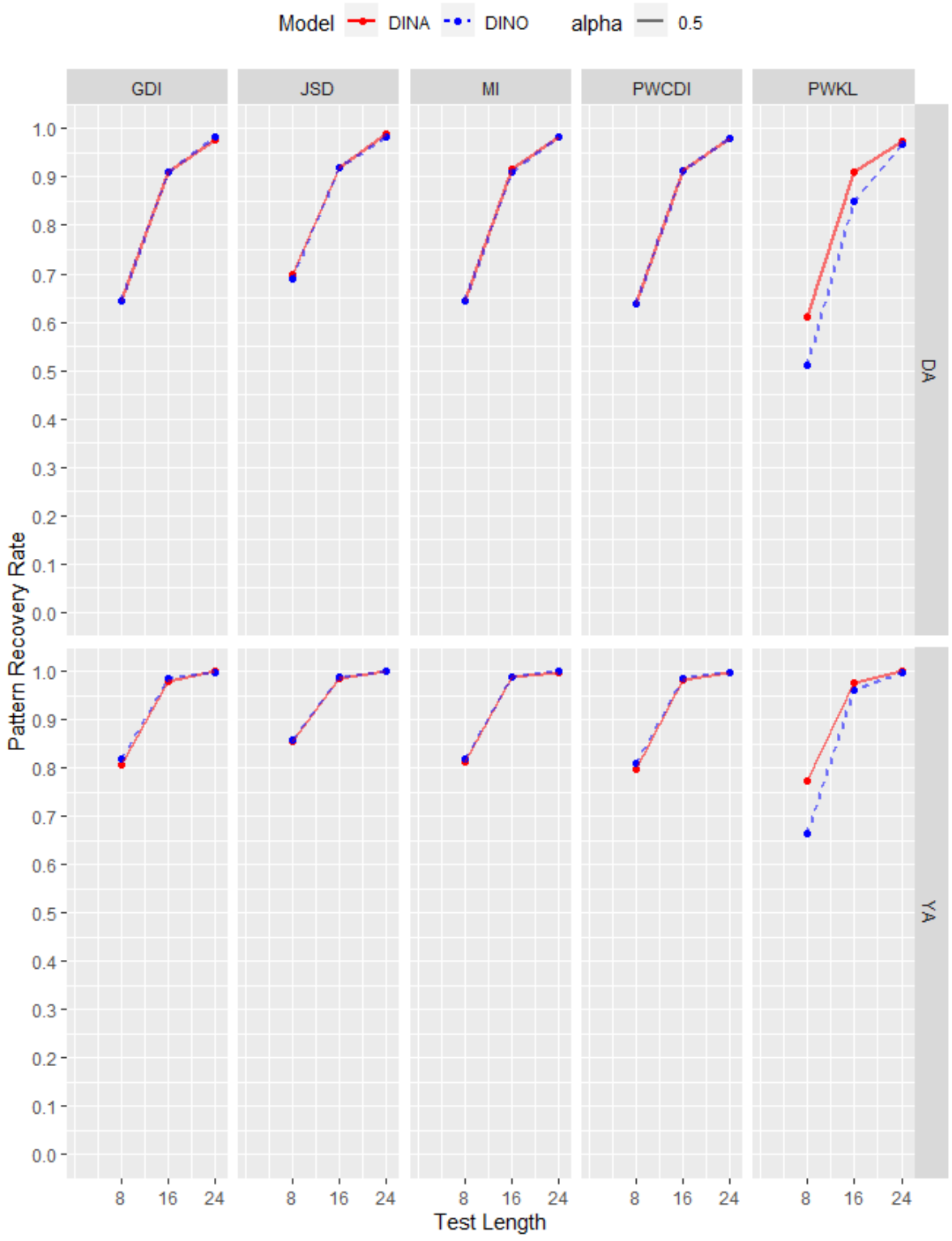


Figure 4. *The PRR values of the item selection algorithms according to the DINA and DINO model when the number of attributes is 5.*

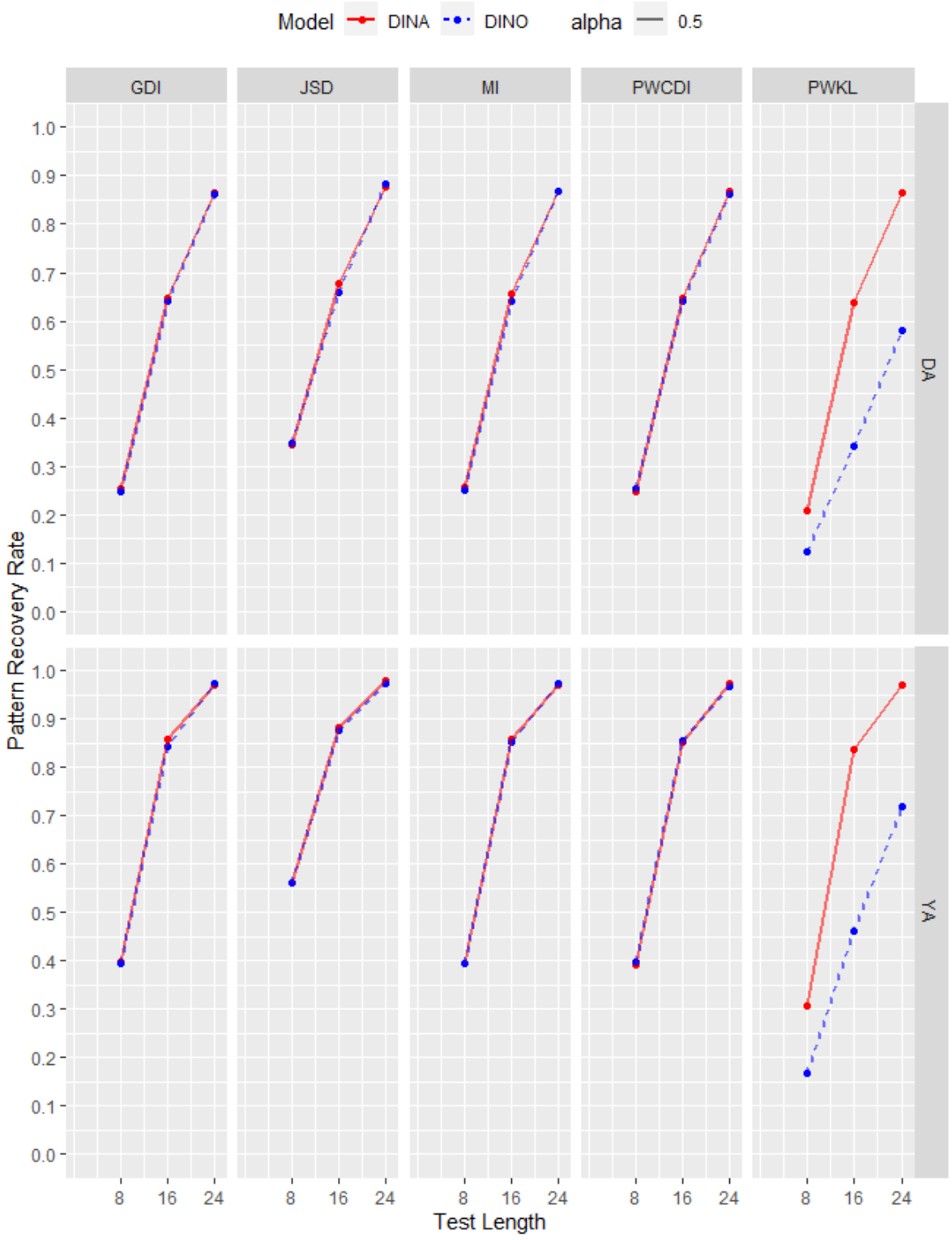


Figure 5. *Graph 6. The PRR values of the item selection algorithms according to the DINA and DINO model when the number of attributes is 8.*

Discussion

When the PRR and ARR values of the item selection algorithms are examined according to the DINA model, the PRR and ARR values of the algorithms increase as the test length and item quality increase and decrease as the number of attributes increases. These findings are parallel with the findings obtained by Cheng (2009), Wang (2013), Kaplan et al. (2015), Zheng and Chang (2016), Yigit et al. (2019). Since the PRR values of the item selection algorithms are close to 1 when the test length is 16, at the number of 5 attributes, and at the high item quality level, all algorithms with a test length of 16 for 5 attribute level can be preferred in terms of measurement accuracy. In low item quality, although high measurement accuracy was obtained from item selection algorithms in 16 test lengths, these values were found to be lower than high item quality. In this respect, when a bank consisting of low-quality items is used, longer tests should be used for more accurate classification, when the number of attributes is 8, the PRR values of the item selection algorithms decrease. The highest PRR value was found to be 0.876, even at 24 test lengths, especially in banks with items with low item discrimination. From this point of view, when the number of attributes increases, the discrimination of the items in the item bank should also be high. PRR values obtained from item selection algorithms are high at 8 attribute level, high item quality, and 24 test length termination rules. In this respect, the use of long tests and banks with highly discriminated items will increase measurement accuracy under 8 attribute level. Xu et al. (2003) stated that the pattern recovery rates of the algorithms decreased as the number of attributes increased in their study in which they examined the pattern recovery rates of item selection algorithms at the 5 and 8 attribute levels according to the DINA model. Wang (2013) examined item selection algorithms at 5 and 8 attribute levels in terms of attribute and pattern recovery rates and stated that increasing the number of attributes decreases the attribute and pattern recovery rates. Findings by studies of Xu et al. (2003) and Wang (2013) are similar to the findings obtained from this study. Similar results were obtained from the simulation study performed according to the DINO model.

In the simulation study performed for DINA and DINO models, it was observed that the measurement accuracy of the JSD algorithm was slightly higher than the measurement accuracy obtained from other algorithms under all conditions. Wang (2013) in his study comparing MI algorithm with SHE, PWKL, and KL algorithms for short tests, stated that the ARR and PRR values of the MI algorithm are higher than the values obtained from the PWKL algorithm. Yigit et al. (2019), in their study using the MC-DINA model, stated that the

measurement accuracy of the JSD algorithm is higher than the measurement accuracy obtained from the GDI and PWKL algorithms. Zheng and Chang (2016) found that the measurement accuracy values obtained from the PWCDI algorithm were higher than the PWKL algorithm in their study, which they carried out according to the DINA model at 5 attribute level. They also concluded that in the short tests, the measurement accuracy values of the MI algorithm are higher than these algorithms, and in the long tests, the measurement accuracy of these algorithms is similar. In their study, Kaplan et al. (2015) stated that higher measurement accuracy values were obtained from the GDI algorithm than the PWKL algorithm. Findings from this study correspond to the findings obtained by Wang (2013), Kaplan et al. (2015), Zheng and Chang (2016), and Yigit et al. (2019).

When the performances of the item selection algorithms are evaluated according to the cognitive diagnostic model used, the measurement accuracy values of the algorithms other than the PWKL algorithm are very close to each other for both models. However, in the study carried out according to the DINO model, the measurement accuracy value obtained from the PWKL algorithm in 8 attribute level is significantly lower than the values obtained from the DINA model. In this case, it can be said that the performance of the PWKL algorithm in terms of measurement accuracy is significantly lower than other algorithms in the DINO Model when the attribute level is high.

Conclusion

This study aims to examine the performance of item selection algorithms in DINA and DINO models, with different item quality, different number of attributes, and different fixed test length termination rule levels, according to the measurement accuracy.

In the study, it was concluded that as item discrimination of items in the bank is increased, the measurement accuracy of the item selection algorithms increased significantly. In addition, as the test length increases, the measurement accuracy of the algorithms increases, while the measurement accuracy of the algorithms decreases with the increase in the number of attributes. All item selection algorithms can be used in studies with banks with 8 attribute levels, 24 test lengths, and high item discrimination. When the number of attributes is 5, it is recommended to use algorithms with high item quality, 16 and 24 test lengths, since the measurement accuracy of the algorithms approaches 1. In all circumstances, the highest values of measurement accuracy were obtained from the JSD algorithm. The measurement accuracy values obtained from the PWKL algorithm are slightly lower than other algorithms. In the study, the performances of the algorithms except the PWKL algorithm are approximately the

same in all conditions. The measurement accuracy values obtained from the MI algorithm were slightly higher than the GDI and PWCDI algorithms. Another result drawn from the study is that the performances of other item selection algorithms are similar for DINA and DINO models except for the PWKL algorithm. It is not recommended to use the PWKL algorithm, especially in the DINO model and when the number of attributes is high.

This study considered item quality, number of attributes, and fixed test length as termination rules. A similar study can be performed using variable test length as the termination rule. DINA and DINO models, which are non-complementary and complementary, were used within the scope of the study. Both models are restrictive. Similar studies can be performed using generalized models (e.g., GDINA, GDM). This study evaluated item selection algorithms in terms of measurement accuracies (ARR and PRR). A similar study can be performed for a more holistic evaluation of the performances of algorithms using evaluation criteria such as calculation time, item exposure, etc.

Ethical Approval: Since this research is a simulative study, it is not among the studies that require "Ethics Committee Permission".

Conflict Interest: The authors declared no potential conflicts of interest concerning this article's research, authorship, and/or publication.

Authors Contributions: The authors confirm that they are responsible for the entire process of the study.

References

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

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement, 33,* 163–183.

DiBello, L., Roussos, L. A., & Stout, W. F. (2007). Handbook of Statistics. C. R. Rao ve S. Sinharay (Ed). *Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models*. 26, 979-1030.

Embretson, S.E. (2001). *The second century of ability testing: Some predictions and speculations*. Retrieved from a website like http://www.ets.org/Media/Research/pdf/PICANG7.pdf. on 6 May 2020.

Haertel, E.H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement*, 8, 333–346.

Hsu, C. L., & Wang, W. C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *Journal of Educational Measurement*, 52, 125–143.

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, 25(3), 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.

Ma. W., & de la Torre, J. (2020). GDINA: An R Package for Cognitive Diagnosis Modeling. *Journal of Statistical Software*, 93(14), 1-26. doi: 10.18637/jss.v093.i14.

Minchen, N. D., & de la Torre, J. (2016). *The continuous G-DINA model and the Jensen-Shannon divergence*. Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rupp, A. A., Henson, R. A., & Templin, J. L. (2010) *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press

Stocking, M.L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Rep. No. 94-5). Princeton: Educational Testing Service.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society*, 51, 337–350.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of Royal Statistics*, 65, 143–157.

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

Thissen, D., & Mislevy, R. J. (2000). Computerized Adaptive Testing: A primer. H. Wainer, (Ed). *Testing algorithms*, Mahwah, NH: Lawrence Erlbaum Associates, Inc, s. 101-133.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report 05-16). Princeton, NJ: Educational Testing Service.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length. *Educational and Psychological Measurement*, 73(6), 1017–1035.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Xu, X., Chang, H., & Douglas, J. (2003). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized Adaptive Testing for Cognitively Based Multiple-Choice Data. *Applied Psychological Measurement*, 43(5), 388–401.

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, 40, 608-624.