ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);79-97

Madde Tepki Kuramı'na Dayalı Madde-Uyum İndekslerinin I.Tip Hata ve Güç Oranlarının İncelenmesi*

Investigating Type I Error and Power Rates of Item Fit Indices Based on Item Response Theory

Seçil ÖMÜR SÜNBÜL** Semih AŞİRET***

Öz

Bu çalışmada, Madde Tepki Kuramı'na göre ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin, çeşitli koşullardaki (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi) I. tip hata ve güç oranlarının incelenmesi amaçlanmıştır. Çalışmada, indekslerin I. tip hata ve güç oranlarının belirlenmesi simülasyon çalışmasıyla yapılmıştır. Çalışmada, madde uyumu için geleneksel indekslerden χ^2, Q_1 ve G^2 indeksleri ile alternatif indekslerden $S\text{-}\chi^2$ indeksi kullanılmıştır. Çalışmada yer alan dört farklı madde-uyum indeksinin I. tip hata ve güç oranları, örneklem büyüklüğü (1000, 2000, 4000), test uzunluğu (20, 30, 40) ve uyumsuzluk yüzdesi (%0, %10, %30 ve %50) değişimlenerek incelenmiştir. Veriler R 3.3.2 yazılımı kullanılarak üretilmiştir ve "mirt" paketi kullanılarak analiz edilmiştir. Çalışmada üretilen ve analiz edilen model olmak üzere iki tür model kullanılmıştır. Üretilen modele uygun madde tepkileri ile analiz edilen modele uygun madde tepkileri için madde-uyum indekslerinin *p* değerleri ve serbestlik dereceleri hesaplanmıştır. Uyum indekslerinin I. tip hata ve güç oranları 0.05 anlamlılık düzeyine göre değerlendirilmiştir. Her uyum indeksinin tüm koşullardaki I. tip hata ve güç oranları hesaplanarak bu indeksler karşılaştırılmıştır. Çalışma sonucunda, tüm faktörlerde $S\text{-}\chi^2$ indeksinin diğer indekslere göre daha düşük hataya sahip olduğu görülmüştür. 2000 ve üzeri örneklem büyüklüğünde ve 20 ve daha fazla maddeden oluşan testlerde $S\text{-}\chi^2$ indeksinin diğer indekslerden daha düşük I. tip hata oranına ve daha yüksek güce sahip olduğu görülmüştür.

Anahtar Kelimeler: Madde Tepki Kuramı, madde-uyum indeksi I.tip hata, güç, $S\text{-}\chi^2$

Abstract

In this study, it was aimed to investigate type I error and power rates of the item fit indices through various conditions (sample sizes, different test lengths and different magnitudes of misfit) for dichotomously generated items based on one-, two-, and three-parameter logistic models in Item Response Theory. In this study, the type I error and power rates of these item fit indices were assessed in a simulation study. χ^2, Q_1 and G^2 indices as traditional item fit indices and $S\text{-}\chi^2$ index as alternative indices were assessed. The performance of four different item fit indices in study were compared by manipulating three different sample size (1000, 2000, 4000), three different test lengths (20, 30, 40) and four different misfit magnitude (%0, %10, %30 and %50). Item responses were generated using the R 3.3.2 software program and analyzed by using "mirt" package in R software. The p value of item fit indices and their degrees of freedom were calculated for both item responses for generating model and analysis model. Type I errors and power rates of item fit indices were examined according to significance levels of 0.05. All item fit indices in this study were compared by calculating the type I error and power rates of each item fit indices under all conditions. The findings indicated that $S\text{-}\chi^2$ index has lower type I error to detect misfit than the other indices. It can be concluded that in the case where the sample size was 2000 or more and the number of items in test are 20 and more, $S\text{-}\chi^2$ index has lower type I error rates than traditional indices and has adequate power to detect misfit items.

Keywords: Item Response Theory, item fit index, type I error, power, $S\text{-}\chi^2$.

* Bu çalışma V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde (01-03 Eylül 2016) sözlü bildiri olarak sunulmuştur.

** Yrd. Doç. Dr., Mersin Üniversitesi, Eğitim Bilimleri Bölümü, Mersin-Türkiye, e-posta:secilomur@gmail.com
*** Uzman, Mersin Üniversitesi, Eğitim Bilimleri Bölümü, Mersin-Türkiye, e-posta:semihasiret@gmail.com

DOI: 10.21031/epod.301529
Geliş Tarihi: 15.11.2016
Kabul Tarihi: 13.03.2017

GİRİŞ

Madde Tepki Kuramı (MTK), madde ve bireyin özelliklerini kullanarak, bireyin performansını kestirmede matematiksel modeller kullanan güçlü bir ölçekleme tekniğidir (Embretson ve Reise, 2000). Madde Tepki Kuramı'nın test puanlarını yorumlamada ve test sonuçlarını raporlamada birçok avantajı bulunmaktadır. Ancak, bu avantajlar seçilen model ile test verilerinin uyumlu olduğu durumlarda elde edilebilir (Hambleton ve Swaminathan, 1985). Madde Tepki Kuramı'nın güçlü yanlarından bir tanesi de değişmezlik özelliğine dayalı olmasıdır, yani madde parametrelerinin yetenek dağılımlarına veya birey parametrelerine bakılmaksızın aynı kalmasıdır (Embretson ve Reise, 2000). Ancak parametre değişmezliği, belirli MTK modellerinin tek boyutluluk, yerel bağımsızlık ve madde karakteristik eğrisinin monotonik artması gibi sayıltıları gerçekleştikten sonra geçerli olacaktır (Wells ve Hambleton, 2016). MTK modeli veriyle uyumsuzken uygulandığında değişmezlik özelliğindeki tüm önemli bilgiler kaybolacaktır. MTK modellerinin geçerli bir şekilde uygulanması ve kararlı puan ölçeklerinin elde edilebilmesinde, model veri uyumu önemli bir role sahiptir.

Aynı modelin testteki tüm maddelere uygulanma zorunluluğu yoktur. Örneğin; bir test, hem ikili hem de çoklu puanlanan maddelerden oluşabildiği gibi, aynı zamanda bazı maddeler iki parametreli lojistik model ile bazıları ise aşamalı tepki (graded response) modeliyle uyum gösterebilir (Embretson ve Reise, 2000). Bu nedenle, birçok çalışmada genel model-veri uyumunun aksine, MTK model uyumunun madde madde yargılanması önerilmektedir (Chon, Lee ve Ansley, 2007; Chon, Lee ve Dunbar, 2010; Tay, Ali, Drasgow ve Williams, 2011; Wells ve Bolt, 2008).

Belirli bir madde düzeyinde, maddelerin uyumunun değerlendirilmesinde genel strateji, gözlenen verilerle kestirilen verinin karşılaştırılmasıdır (Hambleton, Swaminathan ve Rogers, 1991). MTK'de öncelikle MTK modelinin parametreleri kestirilir. Ardından bu kestirilen parametreler kullanılarak bireylerin tepki örüntüleri kestirilir. Son olarak kestirilen tepki örüntüleri ile bireyin gerçek gözlenen tepki örüntüleri karşılaştırılır.

Madde uyumunu değerlendirme işlemi iki genel yaklaşımla gerçekleştirilir. Birincisi, çok fazla istatiksel işlem gerektirmeyen grafiksel işlemlerdir. Burada madde uyumunun yargılanması, kestirilen Madde Tepki Eğrisi (MTE) ile görgül olarak gerçek veya gözlenen verilerden elde edilen MTE'nin karşılaştırılmasına dayanır (Embretson ve Reise, 2000; Reise, 1990). Kestirilen MTE ile gözlenen MTE arasındaki fark artık (residual) olarak adlandırılır ($r=O_{ig}-E_{ig}$ şeklinde gösterilir). Artıkların görsel olarak gösterimiyle model veri uyumunun incelenmesi faydalı görülürken, öznel olma durumundan dolayı da eleştirilmiştir. Bu sebeple, MTK modellerinin veri ile uyumunu incelemek için birçok madde-uyum indeksi geliştirilmiştir (Ames, 2015; Ames ve Penfield, 2015; Lahuis, Clark ve O'brien, 2011).

Birçok madde-uyum indeksi geliştirilmiş olsa da, bu indeksler genel olarak ki-kare yaklaşımı ve olabilirlik oranı yaklaşımı olmak üzere iki yaklaşıma dayanmaktadır. İkili puanlanan maddelerde model veri uyumunu değerlendirmedeki ki-kare yaklaşımı genel olarak Eşitlik-1 de gösterilmektedir.

$$\chi^2 = \sum_{g=1}^{G} N_g \frac{(O_{ig}-E_{ig})^2}{E_{ig}(1-E_{ig})} \quad (1)$$

Eşitlik-1'de, O_{ig}, g aralığındaki i maddesi için gözlenen doğru oranını, E_{ig}, yetenek kestirim aralığındaki kestirilen MTE'ye dayalı beklenen doğru oranını, N_g, g yetenek aralığına denk gelen birey sayısını göstermektedir. Ki-kare istatistiği standardize edilmiş artıkların karesinin toplamını göstermektedir. Artık, doğrudan ki-kare eşitliğinin içinde yer almaktadır. Artık değerleri arttıkça ki-kare değeri de artmaktadır.

İkili puanlanan maddeler için olabilirlik oranı ise Eşitlik-2'de gösterilmektedir.

$$2\sum_{k=1}^{K} N_k\left[O_{ik}\ln\left(\frac{O_{ik}}{E_{ik}}\right)+(1-O_{ik})\ln\left(\frac{1-O_{ik}}{1-E_{ik}}\right)\right] \quad (2)$$

Eşitlik-2'de ilk bakışta artıklar görülmez. Ancak $\ln(\frac{O_{ik}}{E_{ik}})$'nin doğal logaritması aynı zamanda $\ln O_{ik} - \ln E_{ik}$ şeklinde ifade edildiğinden eşitlikte artıklar yer almaktadır.

Ki-kare veya olabilirlik oranı yaklaşımlarından bir tanesine dayalı olan madde uyum indeksleri iki farklı boyutta farklılaşır (Ames ve Penfield, 2015). Birincisi, grupların oluşturulma şeklidir. Gruplar yetenek düzeylerine göre farklı şekillerde oluşturulabilmektedir. İkincisi ise, kestirilen madde tepki eğrisindeki doğru cevaplama oranlarının (E_{ig}) hesaplanma şeklidir.

Lahuis, Clark ve O'brien (2011) bu indekslerin serbestlik derecelerinin hesaplanma şekline bağlı olarak farklı şekilde sınıflandırılabileceğini belirtmiştir. Buna göre indeksler, geleneksel ve alternatif madde-uyum indeksleri olmak üzere iki şekilde sınıflandırılmıştır. Geleneksel madde-uyum indekslerinde model-veri uyumu, her maddenin seçilen MTK modelindeki çeşitli alt yetenek gruplarının gözlenen performansı ile kestirilen performansı karşılaştırılarak değerlendirilir (Stone ve Zhang, 2003). Geleneksel madde-uyum indeksleri; *OUTFIT* ve *INFIT* indeksleri (Wright ve Panchapakesan, 1969), Bock'un X^2 indeksi (Bock, 1972), Yen'in Q_1 istatistiği (Yen, 1981) ve G^2 indeksidir (McKinley ve Mills, 1985).

Geleneksel yöntemlerde yetenek aralıklarının belirlenmesi genellikle keyfidir ve belirlenen farklı aralıklar oluşabilecek sonuçları da etkilemektedir (Orlando ve Thissen, 2000; Reise, 1990). Aralıklar yetenek koşuluna bağlı olduğu için gözlenen tepkiler modele bağımlı olmaktadır. Bu durum, uyum indekslerinin serbestlik derecesini etkileyebilmektedir (Orlando ve Thissen, 2000). Stone ve Zhang (2003), yetenek kestirimindeki belirsizliğin χ^2 istatistiği üzerinde olumsuz bir etkiye sahip olduğunu belirtmiştir. Ayrıca geleneksel indeksler test uzunluğu ve örneklem büyüklüğüne karşı çok duyarlıdır (Kang ve Chen, 2011). Literatürde geleneksel madde-uyum indekslerin sıklıkla kullanıldığı görülmekle birlikte bu indekslerin sınırlılıklarıyla ilgili detaylı birçok çalışmanın yapıldığı da görülmektedir (DeMars, 2005; Glas ve Su´arez Falc´on, 2003; Orlando ve Thiseen, 2000; Stone ve Zhang, 2003; von Schrader, Ansley ve Kim, 2004). Bu sebeple, birçok alternatif madde-uyum indeksleri üretilmiştir (Lahuis, Clark ve O'brien, 2011). Bu indeksler Orlando ve Thissen (2000) tarafından geliştirilen *S*-χ^2, Stone (2000) tarafından gerçekleştirilen ölçeklenmiş düzeltilmiş (scaling corrected) uyum istatistiği (χ^{2*}) ve Drasgow, Levine, Tsien, Williams ve Mead (1995) tarafından geliştirilen ayarlanmış (adjusted) χ^2-serbestlik dereceleri oranı (χ^2/dfs) indeksidir. Bu çalışma kapsamında kullanılacak yazılım farklılığından kaynaklanacak hataların oluşmaması açısından tüm madde uyum indekslerinin aynı yazılımla analiz edilmesi amaçlanmış ve bu nedenle geleneksel madde uyum indekslerinden X^2, Q_1, G^2 ve alternatif madde uyum indekslerinden *S*- X^2 kullanılmıştır. Aşağıda bu çalışma kapsamında kullanılan indekslerden kısaca bahsedilmiştir.

Bock'un Ki-Kare İndeksi (Bock, 1972)

Bock'un ki-kare eşitliği Eşitlik-3 kullanılarak elde edilmektedir. Eşitlik-3'te O_{ig}, g aralığındaki i maddesi için gözlenen doğru oranını, E_{ig}, yetenek kestirim aralığındaki ortancada kestirilen MTE'ye dayalı beklenen doğru oranını, N_g, g yetenek aralığına denk gelen birey sayısını göstermektedir. Ki-kare dağılımının serbestlik derecesi (*Gxm*) ile elde edilir (*G*, aralık sayısı ve m, madde sayısıdır) (Embretson ve Reise, 2000).

$$BCHI = \sum_{g=1}^{G} \frac{N_g(O_{ig} - E_{ig})^2}{E_{ig}(1 - E_{ig})} \qquad (3)$$

Bock'un ki-kare istatistiği beklenen frekansları kestirmek için yetenek aralıklarının ortancasını kullanır ve bu aralıkların boyutları birbirinden farklıdır (Lahuis, Clark ve O'brien, 2011).

Yen'in Q_1 İndeksi

Yen'in Q_1 indeksi (1981), yeteneği 10 eşit aralığa böler ve beklenen frekansları hesaplamak için aralıkların ortalamasını kullanır (Lahuis, Clark ve O'brien, 2011).Yen'in Q_1 istatistiği Eşitlik-4'te belirtilen şekilde hesaplanır.

$$Q_1 = \sum_{j=1}^{10} N_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})} \quad (4)$$

Yen'in Q_1 indeksi ve Bock'un χ^2 indeksi ile yapılan genel eleştirilerden bir tanesi bu iki istatistiğin grupları oluştururken yetenek değerlerini kullanmalarıdır. Model uyumu zayıf olduğunda, kestirilen yetenek değerleri yanlı olacaktır. Yanlı yetenek kestirimlerine göre oluşturulan gruplardan elde edilen istatistikler ise yanlış madde-uyum istatistiği verebilir. Bu yöntemlerin diğer bir sınırlılığı ise, gruplar oluşturulurken keyfi kesme puanı belirlenerek her gruba bireylerin yerleştirilmesidir. Bu durumda, bu istatistikler gruba dayalı istatistikler olacaktır.

Bu indeksler, mükemmel model veri uyumunun sıfır hipotezini değerlendirirler. İndekslerden elde edilen değerler ile serbestlik derecelerine denk gelen kritik değerler karşılaştırılarak hipotez red veya kabul edilir.

G^2 İndeksi

McKinley ve Mills (1985), tarafından geliştirilen bu indeks χ^2 olabilirlik oranı olarak adlandırılır. G^2 indeksi Yen'in Q_1 indeksine benzemektedir. Bu indekste, yetenek 10 eşit aralığa bölünür ve gözlenen ve beklenen frekanslar karşılaştırılır (Lahuis, Clark ve O'brien, 2011). Serbestlik derecesi grup sayısına eşittir. G^2 istatistiği Eşitlik-5'te gösterilen şekilde elde edilmektedir.

$$G_i^2 = 2\sum_{k=1}^{10} N_k \left[O_{ik} \ln\left(\frac{O_{ik}}{E_{ik}}\right) + (1 - O_{ik}) \ln\left(\frac{1 - O_{ik}}{1 - E_{ik}}\right)\right] \quad (5)$$

Bu indeks BILOG-MG ve PARSCALE yazılımlarında standart model veri indeksi olarak kullanılmaktadır. Bock'un χ^2 ve Q_1 istatistikleri gibi bu istatistikte grupların yetenek değerlerine bağlı olarak oluşturulmaktadır.

S-χ^2 İndeksi

Orlando ve Thissen (2000), tarafından ikili puanlanan maddeler için geliştirilen madde-uyum indeksidir. Bu indeks yetenek yerine toplam puanlar üzerine koşullanmıştır. Toplam puandan elde edilen gözlenen ve beklenen tepkiler, χ^2 istatistiği kullanılarak karşılaştırılır. Beklenen tepkilerin hesaplanmasında, her toplam puan için, tüm olası tepki örüntülerinin, her olası toplam puanının katışık olabilirlik dağılımı kullanılır. Yani beklenen tepkiler verilen maddenin başarıldığı ve toplam puanın üretildiği tepki örüntülerinin olabilirliğine dayalıdır. Beklenen tepkiler Thissen, Pommerich, Billeaud ve Williams (1995) tarafından geliştirilen özyinemeli algoritmalar kullanılarak geliştirilir (Lahuis, Clark ve O'brien, 2011). Beklenen tepkiler Eşitlik-6'da gösterilen şekilde hesaplanır.

$$E_{ik} = \frac{\int P_{i1}(\theta) f^{*i}(k-1|\theta) \phi(\theta) d\theta}{\int f(k|\theta) \phi(\theta) d\theta} \quad (6)$$

Eşitlik-6'da P_{i1}, i maddesi için tepki fonksiyonunu, $f(k|\theta)$; verilen yetenekteki koşullu kestirilen test puan dağılımını, $f^{*i}(k-1|\theta)$; i maddesi olmadan koşullu kestirilen test puan dağılımını ve $\phi(\theta)$; yetenek dağılım evrenini göstermektedir. Orlando ve Thissen (2000), χ^2 ve G^2 indekslerini karşılaştırarak ki-kare indeksini (S-χ^2) ve olabilirlik oran istatistiğini (S-G^2) hesaplamıştır. Bu

istatistiklerin serbestlik derecesi toplam kategori sayısından (maksimum elde edilecek puan -1) madde parametre sayısının çıkartılmasıyla elde edilir. Gerektiğinde hücrelerde minimum beklenen frekansın (1) elde edilmesi için hücreler daraltılır. Daralma olduğunda serbestlik derecesinde de düzenlemeler yapılır.

(S-χ^2) indeksinin birçok olumlu özelliği bulunmaktadır. Bu indeksin ikili ve çoklu puanlanan MTK modelleri için I. tip hatası kabul edilebilir ve bu indeks büyük örneklemlerde (N≥2000) yeterli güce sahiptir (Kang ve Chen, 2008; Orlando ve Thissen, 2000, 2003; Stone ve Zhang, 2003). Ayrıca özyinelemeli algoritmalar, ikili puanlanan maddeler için GOODFIT yazılımı (Orlando, 1997), R yazılımı ve hem ikili hem de çoklu puanlanan maddeler için IRTFIT yazılımı ile uygulanabilmektedir. (S-χ^2) indeksi yeteneği keyfi aralıklara bölmez çünkü toplam puan koşuluna bağlıdır. Ancak madde sayısı veya tepki seçenekleri arttığında olası toplam puan sayısı da artacaktır. Bu durum sorun oluşturabilmektedir. Ayrıca bu durumda indeksi hesaplamak büyük emek isteyebilir.

Bu çalışmada MTK'ye dayalı bir, iki ve üç parametreli lojistik modeller için farklı koşullarda madde-uyum indekslerinin I. tip hata ve güç oranlarının değerlendirilmesi amaçlanmıştır. İlgili literatür incelendiğinde, yapılan çalışmalarda genellikle örneklem büyüklüğü ve test uzunluğu faktörlerinin madde-uyum indeksleri üzerindeki etkilerinin incelendiği görülmüştür (Chon, Lee ve Ansley, 2007; Chon, Lee ve Dunbar, 2007; DeMars, 2005; Glas ve Su´arez Falc´on, 2003; Orlando ve Thissen, 2000, 2003; Reise, 1990; Stone ve Zhang, 2003; Wells ve Bolt, 2008). Test uzunluğu ile birlikte testte yer alan uyumsuz madde miktarının da madde-uyum indekslerinin performanslarına etki edeceği düşünülmektedir. İlgili literatürde uyumsuz (misfit) madde miktarının madde-uyum indekslerinin I. tip hata ve güç oranları ile ilgili yeterli çalışma olmadığı görülmüştür (Wells ve Bolt, 2008; Ames, 2015). Bununla birlikte ilgili faktörlerin ortak etkisini inceleyen bir çalışmaya da rastlanılmamıştır. Ayrıca madde-uyum indeksleriyle ilgili ülkemizde yapılan ilk araştırma olması nedeniyle alana katkı getireceği düşünülmektedir.

Çalışma kapsamında aşağıdaki sorulara cevap aranmaya çalışılmıştır.

1. Çeşitli faktörlerin (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi), madde-uyum indekslerinin (X^2, Q_1, G^2 ve S-X^2) I. tip hata ve güç oranlarına temel etkisi nasıldır?
2. Çeşitli faktörlerin (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi), madde-uyum indekslerinin (X^2, Q_1, G^2 ve S-X^2) I. tip hata ve güç oranlarına ortak etkisi nasıldır?

Araştırmanın Amacı

Bu çalışmada, Madde Tepki Kuramı'na göre, ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin çeşitli koşullardaki (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi) I. tip hata ve güç oranlarının incelenmesi ve hangi koşullarda hangi indeksin daha iyi sonuç verdiğinin belirlenmesi amaçlanmıştır.

YÖNTEM

Aşağıdaki bölümde araştırmanın türünden, bu araştırma kapsamında değişimlenen faktörlerden ve bunların düzeylerinden, verilerin üretiminden ve işlem adımlarından bahsedilmiştir.

Araştırmanın Türü

Bu çalışmada, Madde Tepki Kuramı'na göre, ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin çeşitli koşullardaki I. tip hata ve güç oranlarının incelenmesi ve hangi koşullarda hangi indeksin daha iyi sonuç verdiğinin belirlenmesi amaçlandığından çalışma, temel araştırma olarak değerlendirilebilir.

Araştırma Kapsamında Değişimlenen Faktörler

Bu çalışmada, indekslerin I. tip hata ve güç oranlarının belirlenmesi simülasyon çalışmasıyla yapılmıştır. Çalışma kapsamında, madde uyumu için geleneksel indekslerden χ^2, Q_1 ve G^2 indeksleri ile alternatif indekslerden (S-χ^2) indeksi kullanılmıştır. Çalışmada, örneklem büyüklüğü (1000, 2000, 4000), test uzunluğu (20, 30, 40) ve uyumsuzluk yüzdesi (%0, %10, %30 ve %50) faktörleri değişimlenmiştir. Çalışmada yer alan faktörler ve düzeyleri Tablo-1'de özetlenmiştir.

Veri Üretimi

Bu çalışma kapsamında veriler, üretilen model (GM) ve analiz edilen model (CM) olmak üzere iki farklı şekilde, değişimlenen faktörlere uygun olarak üretilmiştir. Veri üretiminde, MTK'ye dayalı 2PL ve 3PL modelleri kullanılmıştır. İlk olarak GM için, kullanılacak modele (2PL ya da 3PL), değişimlenen faktöre ve bunların düzeylerine uygun olarak veri üretilmiştir. Üretilen 2PL ve 3PL modeller için a parametre değerleri 1.00, c parametre değerleri 0.00 ve b parametre değerleri minimum -2 maksimum +2 olan uniform dağılımdan elde edilecek şekilde ayarlanmıştır. Daha sonra analiz edilen model için (CM), veri üretiminde kullanılan modele uygun olarak a, b ve c parametreleri değişimlenerek veri üretilmiştir. Analiz edilen 1PL, 2PL ve 3PL modeller için b parametre değerleri üretilen modeldeki b parametre değerlerine $\pm$ 0.75 eklenerek değişimlenmiştir. 2PL ve 3PL modeller için a parametre değerleri minimum 0, maksimum +2 olan uniform dağılımdan çekilerek değişimlenmiştir. 3PL model için ise c parametre değerleri ise 0.25 olacak şekilde değişimlenmiştir. Çalışmada bireylerin yetenek dağılımlarına ilişkin değerleri ise; ortalaması 0, standart sapması 1 olan standart normal dağılımdan *N(0,1)* elde edilmiştir. Bu yetenek ve parametre değerlerine göre 1-0 verileri üretilmiştir.

Tablo 1. Çalışmada Değişimlenen Faktörler ve Düzeyleri

Faktör	Düzey Sayısı	Düzey Değerleri
Örneklem Büyüklüğü	3	1000 2000 4000
Madde Sayısı	3	20 30 40
Uyumsuzluk Yüzdesi	4	%0 %10 %30 %50
Replikasyon Sayısı	100	

Verilerin Analizi

Verilerin üretiminden sonra ilk olarak maddelerin uyumsuzluk miktarları hesaplanmıştır. Madde modele uyumsuz olsa da uyumsuzluğun miktarı küçük olduğunda uyum indeksleri uyumsuz maddeleri tespit etmede zorlanabilmektedir. Bu sebeple, Wells ve Bolt (2008), tarafından geliştirilen *MISFIT* indeksi kullanılarak uyumsuz maddelerin uyumsuzluk miktarları hesaplanmıştır. Wells ve Bolt (2008), 0.020 ve üzeri uyumsuzluk miktarlarını orta ve yüksek uyumsuzluk olarak ele alınabileceğini ifade etmiştir. Bu sebeple çalışmada uyumsuz olarak kalibre edilmiş maddelerin uyumsuzluk miktarları 0.020 ve üzeri olanlar tercih edilmiştir. GM=3PL ve CM= 2PL, CM=1PL durumlar için üretilen a, b ve c parametreleri ile uyumsuzluk büyüklükleri Tablo 2'de verilmiştir.

Analiz edilen her bir model için maddelerdeki uyumsuzluk miktarı aşağıda belirtilen Eşitlik-7 ile hesaplanmıştır.

$$MISFIT = \sqrt{\sum_{j=1}^{601} w(\theta_j)\,(P_{GMj} - P_{CMj})^2} \qquad (7)$$

MISFIT indeksi, üretilen ve analiz edilen modelin tepki olasılıkları farkının karelerinin ağırlıklandırılmasıyla elde edilen sonucun toplamına eşittir. Yetenek -3 ile + 3 arasında 601 eşit parçaya bölünmüştür. $j=1,2,\ldots\theta$ için $w(\theta_j)$, standart normal yoğunluk tarafından tanımlanan normalleştirilmiş ağırlıktır. P_{GMj} ise θ_j yetenek düzeyindeki bireyin üretilen model için maddeyi cevaplandırma olasılığı iken, P_{CMj} , θ_j yetenek düzeyindeki bireyin analiz edilen model için maddeyi cevaplandırma olasılığıdır. Uyumsuzluk miktarı 0.020 ve üzeri olan maddeler orta ve yüksek uyumsuzluk miktarı olarak ele alınmaktadır.

Tablo 2. 3PL Model İçin Üretilen Parametre Değerleri ve Uyumsuzluk İndeksi Değerleri

	Uyumsuzluk (%0)			Uyumsuzluk (%10)			Uyumsuzluk indeksi		Uyumsuzluk (%30)			Uyumsuzluk indeksi		Uyumsuzluk (%50)			Uyumsuzluk indeksi	
M.N	a	b	c	a	b	c	2PL	1PL	a	b	c	2PL	1PL	a	b	c	2PL	1PL
1	1.00	-1.96	0.00	1.00	-1.96	0.00	-	-	**0.47**	**-1.96**	**0.25**	**0.026**	**0.034**	**0.41**	**-1.24**	**0.25**	**0.027**	**0.038**
2	1.00	-1.84	0.00	1.00	-1.68	0.00	-	-	1.00	-1.84	0.00	-	-	1.00	-1.90	0.00	-	-
3	1.00	-1.69	0.00	1.00	-1.48	0.00	-	-	1.00	-1.69	0.00	-	-	**0.57**	**-1.00**	**0.25**	**0.028**	**0.033**
4	1.00	-1.63	0.00	1.00	-1.41	0.00	-	-	1.00	-1.63	0.00	-	-	1.00	-1.73	0.00	-	-
5	1.00	-1.37	0.00	**0.16**	**-0.59**	**0.25**	**0.037**	**0.070**	**0.16**	**-0.59**	**0.25**	**0.039**	**0.085**	**0.16**	**-0.59**	**0.25**	**0.034**	**0.064**
6	1.00	-1.31	0.00	1.00	-1.02	0.00	-	-	1.00	-1.31	0.00	-	-	1.00	-1.72	0.00	-	-
7	1.00	-1.06	0.00	1.00	-0.96	0.00	-	-	1.00	-1.06	0.00	-	-	**0.47**	**-0.97**	**0.25**	**0.029**	**0.039**
8	1.00	-0.93	0.00	1.00	-0.93	0.00	-	-	1.00	-0.93	0.00	-	-	1.00	-1.48	0.00	-	-
9	1.00	-0.81	0.00	1.00	-0.92	0.00	-	-	**0.65**	**-0.81**	**0.25**	**0.041**	**0.051**	**0.65**	**-0.81**	**0.25**	**0.032**	**0.039**
10	1.00	-0.73	0.00	1.00	-0.83	0.00	-	-	1.00	-0.73	0.00	-	-	1.00	-1.38	0.00	-	-
11	1.00	-0.72	0.00	1.00	-0.72	0.00	-	-	**0.05**	**-0.61**	**0.25**	**0.039**	**0.089**	**0.05**	**-0.61**	**0.25**	**0.038**	**0.080**
12	1.00	-0.66	0.00	1.00	-0.64	0.00	-	-	1.00	-0.66	0.00	-	-	1.00	-1.32	0.00	-	-
13	1.00	-0.57	0.00	1.00	-0.62	0.00	-	-	1.00	-0.57	0.00	-	-	**0.12**	**-0.52**	**0.25**	**0.038**	**0.075**
14	1.00	-0.42	0.00	1.00	-0.61	0.00	-	-	1.00	-0.42	0.00	-	-	1.00	-1.15	0.00	-	-
15	1.00	-0.38	0.00	**0.30**	**0.15**	**0.25**	**0.042**	**0.080**	**0.30**	**0.15**	**0.25**	**0.041**	**0.100**	**0.30**	**0.15**	**0.25**	**0.037**	**0.061**
16	1.00	-0.37	0.00	1.00	-0.55	0.00	-	-	1.00	-0.37	0.00	-	-	1.00	-1.05	0.00	-	-
17	1.00	-0.28	0.00	1.00	-0.53	0.00	-	-	1.00	-0.28	0.00	-	-	**0.58**	**-0.26**	**0.25**	**0.039**	**0.052**
18	1.00	0.03	0.00	1.00	-0.46	0.00	-	-	1.00	0.03	0.00	-	-	1.00	-0.90	0.00	-	-
19	1.00	0.06	0.00	1.00	-0.33	0.00	-	-	**0.54**	**-0.07**	**0.25**	**0.048**	**0.084**	**0.54**	**-0.07**	**0.25**	**0.041**	**0.058**
20	1.00	0.10	0.00	1.00	-0.32	0.00	-	-	1.00	0.10	0.00	-	-	1.00	-0.71	0.00	-	-
21	1.00	0.10	0.00	1.00	-0.25	0.00	-	-	**1.00**	**-0.25**	**0.25**	**0.060**	**0.050**	**1.58**	**0.23**	**0.25**	**0.053**	**0.044**
22	1.00	0.11	0.00	1.00	-0.23	0.00	-	-	1.00	0.11	0.00	-	-	1.00	-0.40	0.00	-	-
23	1.00	0.15	0.00	1.00	-0.16	0.00	-	-	1.00	0.15	0.00	-	-	**1.61**	**0.50**	**0.25**	**0.059**	**0.046**
24	1.00	0.21	0.00	1.00	0.08	0.00	-	-	1.00	0.21	0.00	-	-	1.00	-0.15	0.00	-	-
25	1.00	0.25	0.00	**1.94**	**0.88**	**0.25**	**0.066**	**0.049**	**1.94**	**0.88**	**0.25**	**0.069**	**0.053**	**1.94**	**0.88**	**0.25**	**0.062**	**0.047**
26	1.00	0.49	0.00	1.00	0.25	0.00	-	-	1.00	0.49	0.00	-	-	1.00	0.14	0.00	-	-
27	1.00	0.55	0.00	1.00	0.26	0.00	-	-	1.00	0.55	0.00	-	-	**1.43**	**1.15**	**0.25**	**0.068**	**0.056**
28	1.00	0.67	0.00	1.00	0.56	0.00	-	-	1.00	0.67	0.00	-	-	1.00	0.42	0.00	-	-
29	1.00	0.71	0.00	1.00	0.75	0.00	-	-	**1.60**	**0.71**	**0.25**	**0.073**	**0.060**	**1.60**	**0.71**	**0.25**	**0.070**	**0.053**
30	1.00	0.76	0.00	1.00	0.79	0.00	-	-	1.00	0.76	0.00	-	-	1.00	0.46	0.00	-	-
31	1.00	1.13	0.00	1.00	0.83	0.00	-	-	**1.70**	**1.13**	**0.25**	**0.074**	**0.061**	**1.70**	**-**	**0.25**	**0.070**	**0.054**
32	1.00	1.17	0.00	1.00	1.05	0.00	-	-	1.00	1.17	0.00	-	-	1.00	1.14	0.00	-	-
33	1.00	1.47	0.00	1.00	1.12	0.00	-	-	1.00	1.47	0.00	-	-	**1.84**	**0.43**	**0.25**	**0.058**	**0.045**
34	1.00	1.53	0.00	1.00	1.18	0.00	-	-	1.00	1.53	0.00	-	-	1.00	1.24	0.00	-	-
35	1.00	1.63	0.00	**1.85**	**0.53**	**0.25**	**0.060**	**0.046**	**1.85**	**0.53**	**0.25**	**0.059**	**0.046**	**1.85**	**0.53**	**0.25**	**0.061**	**0.046**
36	1.00	1.68	0.00	1.00	1.41	0.00	-	-	1.00	1.68	0.00	-	-	1.00	1.39	0.00	-	-
37	1.00	1.71	0.00	1.00	1.47	0.00	-	-	1.00	1.71	0.00	-	-	**1.75**	**0.69**	**0.25**	**0.062**	**0.048**
38	1.00	1.76	0.00	1.00	1.58	0.00	-	-	1.00	1.76	0.00	-	-	1.00	1.57	0.00	-	-
39	1.00	1.93	0.00	1.00	1.68	0.00	-	-	**1.84**	**1.93**	**0.25**	**0.065**	**0.049**	**1.84**	**-**	**0.25**	**0.064**	**0.053**
40	1.00	1.98	0.00	1.00	1.77	0.00	-	-	1.00	1.98	0.00	-	-	1.00	1.99	0.00	-	-

Verilerin analizinde R 3.3.2'de yer alan "mirt" paketi kullanılmıştır. Öncelikle GM=CM ve GM>CM olduğu durumlar için her indeksin *p* değerleri hesaplanmıştır. Modeller karşılaştırılırken GM'nin parametre sayısı CM'nin parametre sayısına eşit veya daha yüksek olacak şekilde ayarlanmıştır. Maddelerin uyum gösterip göstermediği, 0.05 anlamlılık düzeyi kullanılarak belirlenmiştir. Eğer uyum indeksinin *p* değeri 0.05'den küçük ise madde modele uyumsuz olarak yorumlanmıştır. R 3.3.2 yazılımı kullanılarak her bir indeks için I. tip hata ve güç oranları hesaplanmıştır. İndekslerin I. tip hatalarının oranı, uyum göstermesi gerekirken uyumsuzluk gösteren madde sayısının toplam uyumlu olması gereken madde sayısına oranı ile hesaplanmıştır. İndekslerin güç oranları ise, uyumsuz olması gereken madde sayısı ile uyumsuz olan madde sayısının oranı ile hesaplanmıştır. GM=CM olduğu durumlarda I. tip hata ve GM>CM olduğu durumlarda güç oranları hesaplanmıştır. Bu simülasyon deseni Tablo 3'te verilmiştir. Her uyum indeksinin tüm koşullardaki I. tip hataları ve güç oranları hesaplanarak bu indeksler karşılaştırılmıştır. İndekslerin her faktör için temel ve ortak etkileri grafikleştirilmiştir.

Tablo 3. Çalışmada Kullanılan Simülasyon Deseni

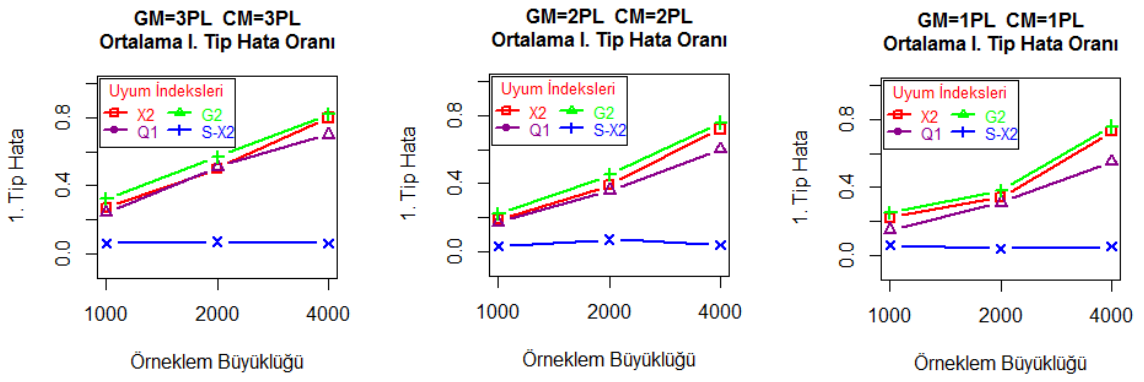
		ÜRETİLEN MODEL (GM)		
		1PL	2PL	3PL
ANALİZ MODEL (CM)	1PL	I. Tip Hata	Güç	Güç
	2PL	-	I. Tip Hata	Güç
	3PL	-	-	I. Tip Hata

BULGULAR

Aşağıdaki bölümde araştırma soruları çerçevesinde elde edilen bulgulara yer verilmiştir.

Örneklem Büyüklüğünün Madde-Uyum İndekslerinin I. Tip Hata ve Güç Oranlarına Temel Etkisine Ait Bulgular

Örneklem büyüklüğünün madde-uyum indekslerinin I. tip hata oranlarına ait temel etki grafikleri Şekil 1'de gösterilmiştir.



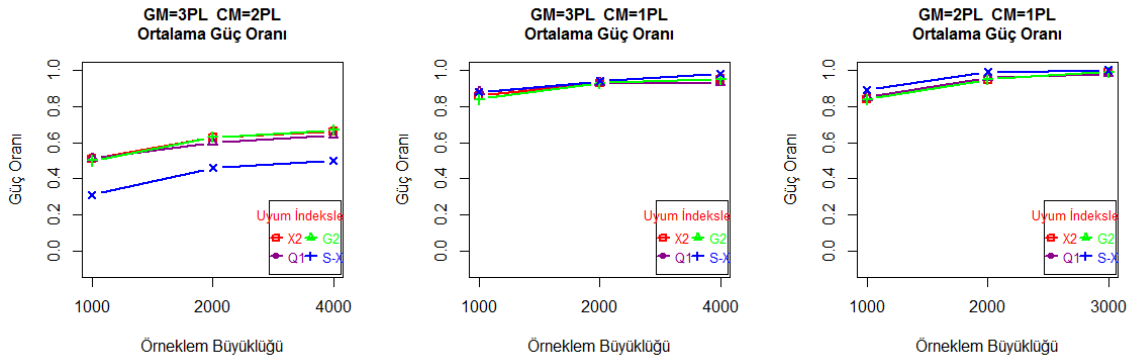
(a) (b) (c)

Şekil 1. Örneklem Büyüklüğünün Madde-Uyum İndekslerinin I. Tip Hata Oranlarına Temel Etkisi

Şekil 1 incelendiğinde örneklem büyüklüğü artıkça GM=3PL CM=3PL, GM=2PL CM=2PL ve GM=1PL CM=1PL durumları için $S\text{-}\chi^2$ indeksi hariç diğer indekslerin I. tip hata oranlarının arttığı

görülmektedir. χ^2, G^2 ve Q_1 indekslerinin I. tip hataları örneklem büyüklüğünün artması ile önemli ölçüde artış gösterirken, $S\text{-}\chi^2$ indeksinin I. tip hatasında önemsiz sayılabilecek değişimler olduğu görülmektedir. Tüm örneklem büyüklüklerinde en düşük I. tip hataya $S\text{-}\chi^2$ indeksinin sahip olduğu söylenebilir. $S\text{-}\chi^2$ indeksi ile geleneksel indeksler arasındaki I. tip hata oranları farkının 4000 örneklem büyüklüğünde önemli ölçüde artış gösterdiği görülmektedir.

Örneklem büyüklüğünün madde-uyum indekslerinin güç oranlarına etkisi ise Şekil 2'de gösterilmiştir.



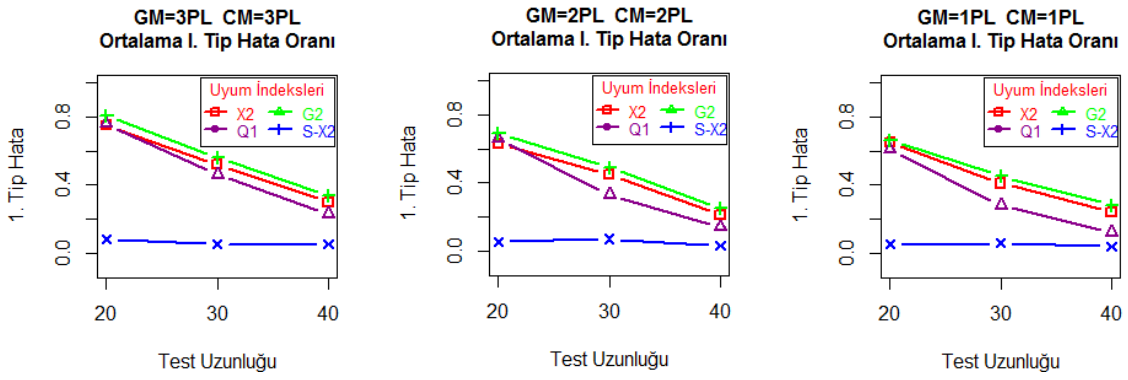
(a) (b) (c)

Şekil 2. Örneklem Büyüklüğünün Madde-Uyum Indekslerinin Güç Oranlarına Temel Etkisi

Şekil 2 incelendiğinde, örneklem büyüklüğü artıkça madde uyum indekslerinin güç oranları artmaktadır. Üretilen modelin, analiz edilen modele göre parametre farkı arttıkça indekslerin güç oranları artmakta ve birbirine yaklaşmaktadır. Şekil 2-a'da GM=3PL ve CM= 2PL durumu için en düşük güce $S\text{-}X^2$ indeksi sahipken, GM=3PL CM=1PL ve GM=2PL CM=1PL olduğu durumlarda tüm örneklem büyüklüklerinde indekslerin güç oranları arasında çok fazla fark bulunmamakla birlikte, $S\text{-}X^2$ indeksinin gücünün diğer indeksler gibi artış gösterdiği ve bu indeksin tüm örneklem büyüklüklerinde diğer indekslerin güç oranlarından daha yüksek olduğu görülmektedir. Şekil 2-a'ya göre GM=3PL ve CM=2PL olduğunda, indekslerin uyumsuz madde tespit etme güç oranlarının orta seviyede (0.30-0.60 arası), Şekil 2-b ve c'de ise, indekslerin uyumsuz maddeleri tespit etme güç oranlarının yüksek seviyede (0.80-1.00 arası) olduğu görülmektedir.

Test Uzunluğunun Madde-Uyum İndekslerinin I. Tip Hata ve Güç Oranlarına Temel Etkisine Ait Bulgular

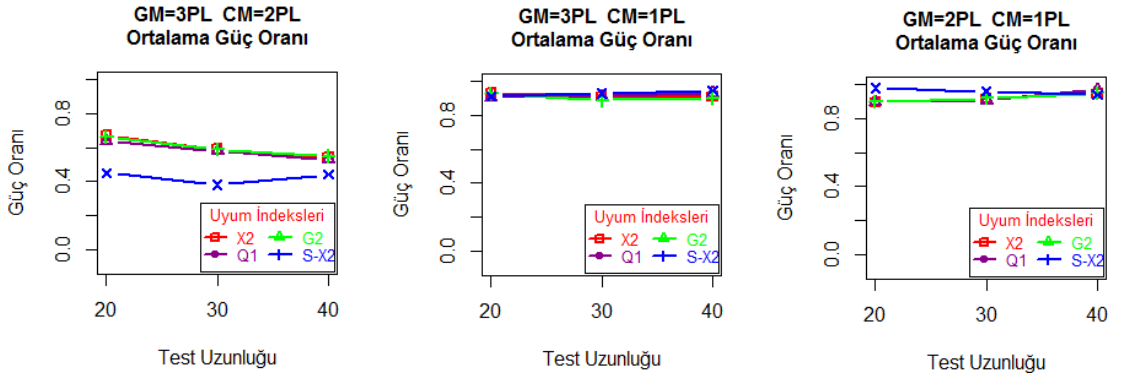
Test uzunluğunun madde-uyum indekslerinin I. tip hata oranlarına temel etkisi Şekil 3'te gösterilmiştir. Şekil 3 incelendiğinde tüm GM=CM durumları için test uzunluğu artıkça indekslerin I. tip hata oranlarının azaldığı görülmektedir. χ^2, G^2 ve Q_1 indekslerinin I. tip hata oranları, test uzunluğunun artmasıyla önemli ölçüde azalırken, $S\text{-}\chi^2$ indeksinin I. tip hata oranında çok az bir azalmanın olduğu görülmektedir. Tüm test uzunluklarında en düşük I. tip hata oranına ise $S\text{-}\chi^2$ indeksinin sahip olduğu söylenebilir. Kısa testlerde $S\text{-}\chi^2$ indeksi ile geleneksel indeksler arasındaki I. tip hata oranları farkı fazla iken, test uzunluğu arttıkça indekslerin I. tip hata oranları arasındaki farkın da azaldığı görülmektedir. Tüm GM=CM olduğu koşullarda test uzunluğu 40 iken indekslerin I. tip hata oranları yakındır.



(a) (b) (c)

Şekil 3. Test Uzunluğunun Madde-Uyum Indekslerinin I. Tip Hata Oranlarına Temel Etkisi

Test uzunluğunun madde-uyum indekslerinin güç oranlarına etkisi Şekil 4'te gösterilmiştir. Şekil 4-a incelendiğinde test uzunluğu artıkça GM=3PL CM=2PL durumu için S-χ^2 indeksi dışında madde-uyum indekslerinin güç oranları azalmaktadır. İndekslerin güç oranları 0.40 - 065 değerleri arasında değişmektedir. Ancak S-χ^2 indeksinin gücü 40 maddelik testte artış göstermiştir. Şekil 4-b ve 4-c incelendiğinde ise tüm indekslerin güç oranlarının tüm test uzunluklarında yüksek olduğu görülmektedir. İndekslerin güç oranları tüm maddelerde 0.90 - 0.95 aralığında değişmektedir. Tüm indekslerin gücü test uzunluğu arttıkça küçük değişimler göstermektedir.

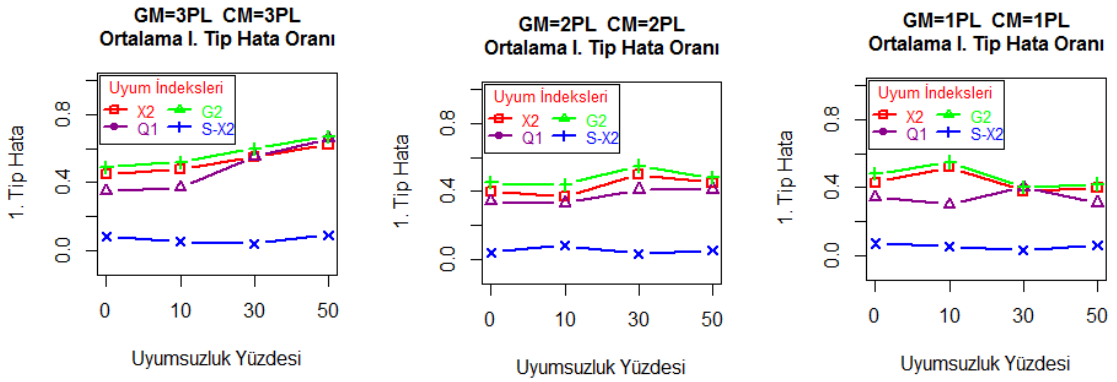


(a) (b) (c)

Şekil 4. Test Uzunluğunun Madde-Uyum İndekslerinin Güç Oranlarına Temel Etkisi

Uyumsuzluk Yüzdesinin Madde-Uyum İndekslerinin I. Tip Hata ve Güç Oranlarına Temel Etkisine Ait Bulgular

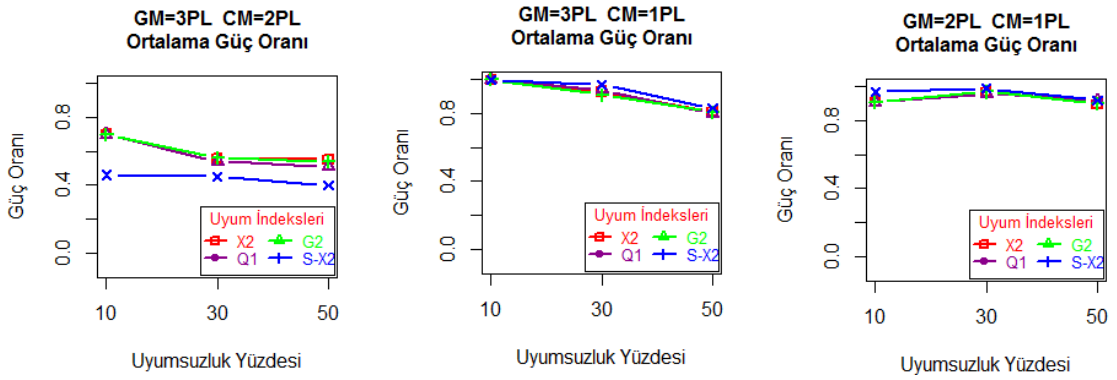
Uyumsuzluk yüzdesinin madde-uyum indekslerinin I. tip hatasına etkisi Şekil 5'te gösterilmiştir. Şekil 5-a incelendiğinde uyumsuzluk yüzdesi artıkça tüm durumlar için geleneksel indekslerin I. tip hatalarının kısmen arttığı, *S*-χ^2 indeksinin I. tip hata oranının ise küçük değişimler gösterdiği görülmektedir. Uyumsuzluk yüzdesi arttıkça I. tip hata oranı en fazla artan indeksin Q_1 indeksi olduğu görülmektedir. Ancak Şekil 5-b ve 5-c incelendiğinde, indekslerin I. tip hata oranlarının uyumsuzluk yüzdesinin artması ile düzensiz küçük değişimler gösterdiğini söyleyebiliriz. Tüm uyumsuzluk yüzdesinde *S*-χ^2 indeksi en düşük I. tip hata oranına sahiptir. Uyumsuzluk yüzdesi 50 olduğu durumlarda, *S*-χ^2 indeksi diğer indekslere göre önemli ölçüde, daha düşük I. tip hata oranına sahiptir.



(a) (b) (c)

Şekil 5. Uyumsuzluk Yüzdesinin Madde-Uyum İndekslerinin I. Tip Hatalarına Temel Etkisi

Uyumsuzluk yüzdesinin madde-uyum indekslerinin güç oranlarına etkisi Şekil 6'da gösterilmiştir. Şekil 6 incelendiğinde uyumsuzluk yüzdesi artıkça tüm durumlar için madde-uyum indekslerinin güç oranları küçük miktarda azalmaktadır. GM=3PL CM=2PL durumu için indekslerin güç oranları 0.40-065 değerleri arasında değişmektedir. Şekil 6-a'da geleneksel indekslerin güç oranları uyumsuzluk yüzdesi 10'dan 30'a artırıldığında hızlı düşüş gösterirken, 30-50 aralığında düşük miktarda azalmıştır. GM=3PL CM=2PL olduğu durumda $S\text{-}\chi^2$ tüm uyumsuzluk yüzdesinde en düşük güç oranına sahiptir. GM=3PL CM=1PL, GM=2PL CM=1PL durumunda tüm indekslerin güç oranları ise yüksek bulunmuş ve indekslerin güç oranları tüm maddelerde 0.80-1.00 aralığında değiştiği görülmüştür. Şekil 6-b ve 6-c incelendiğinde uyumsuzluk yüzdesi arttıkça indekslerin güç oranlarında az da olsa azalmanın olduğu görülmektedir.



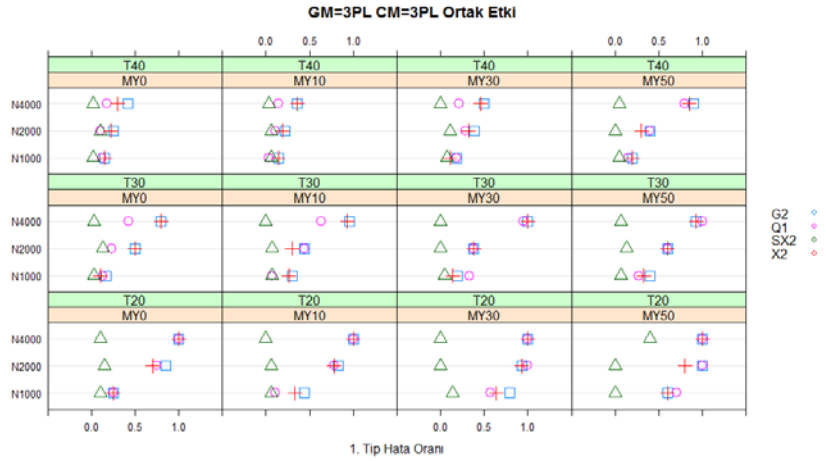
(a) (b) (c)

Şekil 6. Uyumsuzluk Yüzdesinin Madde-Uyum İndekslerinin Güç Oranlarına Temel Etki Grafiği

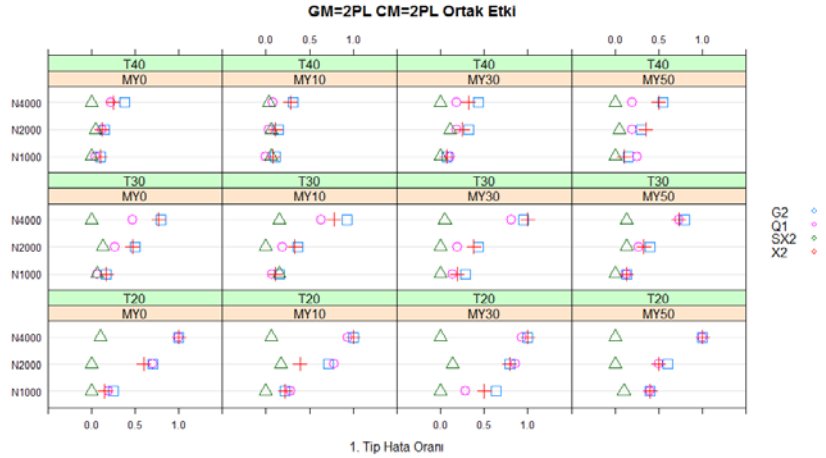
Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin I. Tip Hatalarına Ortak Etkisi

Şekil 7-a incelendiğinde (GM=CM=3PL), tüm koşullarda $S\text{-}\chi^2$ indeksinin I. tip hatası geleneksel indekslerin I. tip hatalarına göre daha düşüktür. Geleneksel indekslerin I. tip hataları örneklem büyüklüğü ve uyumsuzluk yüzdeleri arttıkça artmakta, test uzunluğu arttıkça ise I. tip hataları azalmaktadır. Geleneksel indeksler test uzunluğu 40 ve örneklem büyüklüğü 1000 olduğu

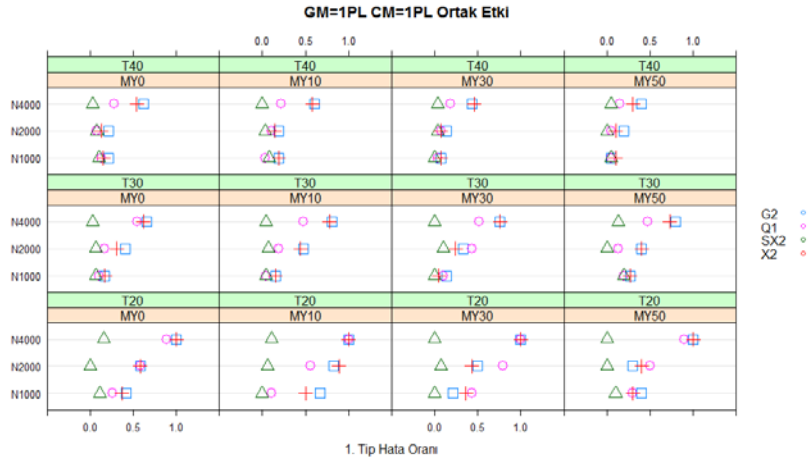
durumlarda $S\text{-}\chi^2$ indeksi ile benzer ve düşük I. tip hataya sahiptir. Ayrıca geleneksel indeksler uyumsuzluk yüzdesinin 0, örneklem büyüklüğünün 1000 olduğu tüm test uzunluklarında $S\text{-}\chi^2$



(a)

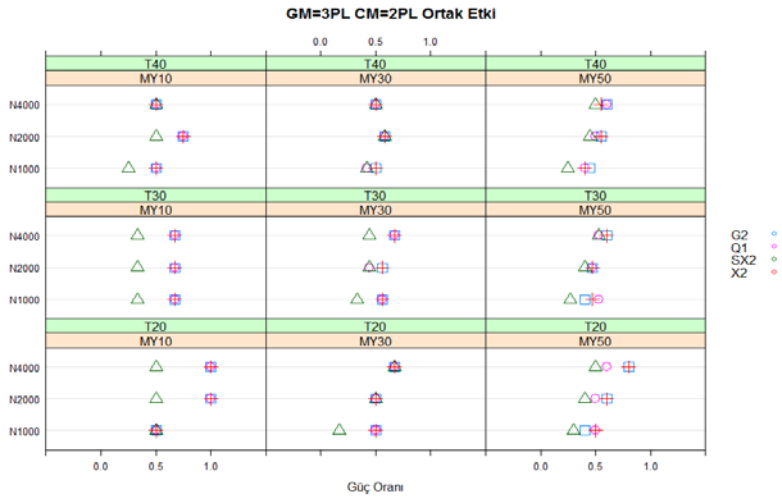


(b)

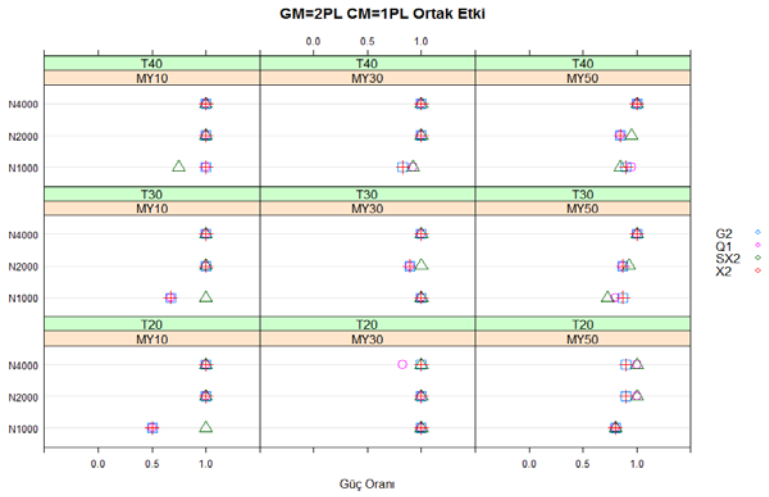


(c)

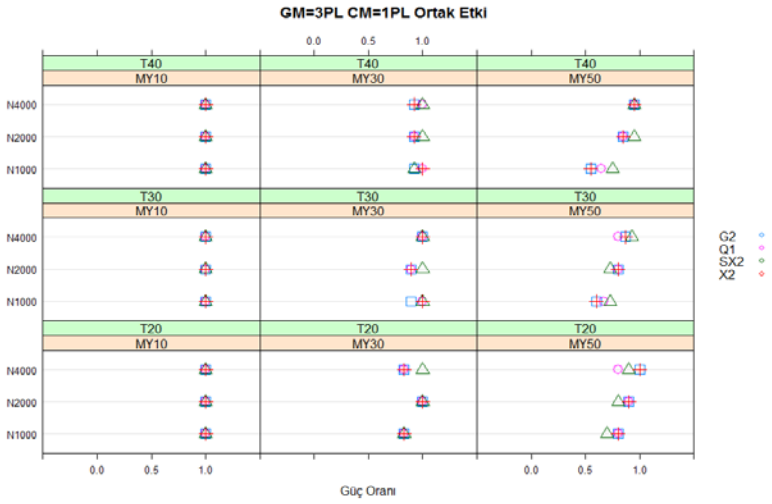
Şekil 7. Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin I. Tip Hatalarına Ortak Etkisi



(a)



(b)



(c)

Şekil 8. Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin Güç Oranlarına Ortak Etkisi

indeksine benzer ve düşük I. tip hata oranına sahiptir. Diğer tüm durumlarda $S\text{-}\chi^2$ indeksi daha düşük I. tip hata oranına sahiptir. Geleneksel indeksler, büyük örneklemlerde (2000 ve üzeri) ve yüksek uyumsuzluk yüzdesinde (%50) $S\text{-}\chi^2$ indeksine göre çok büyük I. tip hata oranına sahip olduğu görülmektedir. Benzer sonuçlar GM=CM=2PL ve GM=CM=1PL olduğu durumda da görülmektedir.

Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin Güç Oranlarına Ortak Etkisi

Örneklem büyüklüğünün, test uzunluğunun ve uyumsuzluk yüzdesinin madde-uyum indekslerinin güç oranlarına ortak etkisi incelendiğinde, uyumsuzluk yüzdesinin düşük olduğu durumlarda, test uzunluğu ve örneklem büyüklüğünün tüm düzeylerinde, indekslerin güç oranlarında önemli değişiklikler olmadığı Şekil 8'de görülmektedir. Ancak uyumsuzluk yüzdesi arttıkça, yöntemlerin güç oranlarında azalmaların olduğu, özellikle düşük örneklemlerde ve uyumsuzluk yüzdesinin yüksek olduğu durumlarda, indekslerin uyumsuz maddeleri tespit etme güçlerinin azaldığı söylenebilir. Şekil 8 incelendiğinde, test uzunluğunun değişimlenmesi diğer faktörlerin tüm düzeylerinde, indekslerin güç oranlarında anlamlı değişimler yapmadığı söylenebilir. Şekil 8-b ve 8-c incelendiğinde, CM=1PL olduğu durumlarda, indekslerin tüm koşullarda yüksek ve benzer güce sahip oldukları görülmektedir.

SONUÇLAR ve TARTIŞMA

Bu çalışmada, Madde Tepki Kuramı'na göre, ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin çeşitli koşullardaki (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi) I. tip hata ve güç oranlarının incelenerek, hangi koşulda hangi indeksin daha iyi sonuç verdiğini tespit etmek amaçlanmıştır. Bu amaçla simülasyon koşulları oluşturularak veriler üretilmiş ve sonuçları analiz edilmiştir.

Çalışmada yer alan koşullardan biri olan örneklem büyüklüğünün madde-uyum indekslerinin I. tip hata ve güç oranına etkisini incelemek için, 1000, 2000 ve 4000 olmak üzere üç farklı şekilde değişimlenmiştir. Madde-uyum indekslerinin I. tip hataları GM=CM olduğu durumlara göre değerlendirilmiştir. Orlando ve Thissen (2003), χ^2 ve Q_1 indekslerinin 500 ve 1000 örneklem büyüklüklerinde makul I. tip hataya sahip olduğunu ancak 2000 örneklem ve üzerinde I. tip

hatalarının yüksek olduğunu belirtmiştir. Ayrıca aynı çalışmada $S\text{-}\chi^2$ indeksinin 1000 örneklem büyüklüğü ve üzerinde makul hata değerine sahip olduğunu belirtmiştir. Ayrıca Orlando ve Thissen (2000), Stone ve Zhang (2003) ve Chon ve Dunbar (2010) yapmış oldukları çalışmada G^2 indeksinin 1000 ve üzeri örneklem büyüklüğünde, hata miktarının arttığını ve $S\text{-}\chi^2$ indeksinin daha makul I. tip hataya sahip olduğunu belirtmiştir. Bu çalışmada, GM=CM tüm durumları için, geleneksel indeksler olarak sınıflandırılan χ^2, Q_1 ve G^2 indekslerin I. tip hataları örneklem büyüklüğünden etkilenmiştir. Örneklem büyüklüğü arttıkça geleneksel indekslerin I. tip hata oranları da artmıştır. Ancak çalışmada alternatif indeks olarak yer alan $S\text{-}\chi^2$ indeksi örneklem büyüklüğünün değişimlenmesinden önemli ölçüde etkilenmemiştir. Tüm örneklem büyüklüğü düzeylerinde en düşük hataya $S\text{-}\chi^2$ indeksi sahiptir. Ayrıca geleneksel indeksler 1000 örneklem büyüklüğünün üzerinde yüksek hata oranına sahip olduğu görülmektedir. Elde edilen bu sonuçlar Orlando ve Thissen (2000), Orlando ve Thissen (2003), Stone ve Zhang (2003) ve Chon ve Dunbar'ın (2010) çalışmalarından elde edilen sonuçlarla benzerlik göstermektedir. Örneklem büyüklüğünün, madde-uyum indekslerinin güç oranlarına etkisi, GM=3PL CM=2Pl, GM=3PL CM=1PL ve GM=2PL CM=1PL durumları için incelenmiştir. Wells ve Bolt (2008), çalışmalarında örneklem büyüklüğünün artması ile madde-uyum indekslerinin gücünün arttığını göstermiştir. Bu çalışma sonucunda Wells ve Bolt'un (2008), çalışmasına paralel şekilde örneklem büyüklüğünün artması sonucunda indekslerin uyumsuz maddeleri tespit etme oranlarının arttığı soncuna ulaşılmıştır. Orlando ve Thissen (2000) çalışmasında, GM=3PL veya GM=2PL ve CM=1PL olduğu durumlarda indekslerin uyumsuz maddeleri tespit etme yüzdesinin %50 civarında olduğunu belirtmiştir. Ayrıca Orlando ve Thissen (2000) ve Stone ve Zhang (2003) GM=3PL ve CM=2PL olduğunda indekslerin uyumsuz maddeleri tespit etme oranlarının düştüğünü ifade etmiştir. Bu çalışmada GM=3PL ve CM=2PL olduğu durumlarda, güç oranlarının Orlando ve Thissen'in (2000) çalışmalarının sonucuna benzerlik gösterdiği söylenebilir. Ancak çalışma sonucunda GM=3PL veya GM=2PL ve CM=1PL olduğu durumlarda indekslerin güç oranlarının arttığı ve 0.80-1.00 aralığında değiştiği söylenebilir. Bu durumun, uyumsuz maddeler oluşturulurken parametrelerin uç değerlerde farklılaştırılmasından yani uyumsuzluk miktarının (misfit magnitude) değerinin yüksek olmasından kaynaklandığı düşünülmektedir.

Çalışmada yer alan bir diğer faktör olan test uzunluğunun madde-uyum indekslerinin I. tip hatalarına etkisi için madde sayısı 20, 30 ve 40 olmak üzere üç farklı düzeyde incelenmiştir. Çalışmada test uzunluğunun artması $S\text{-}\chi^2$ indeksinin I. tip hata oranını önemli ölçüde değiştirmez iken, geleneksel indekslerin I. tip hata oranını ise azalttığı sonucuna ulaşılmıştır. Mislev ve Bock (1990), G^2 indeksinin 20'den daha uzun testler için daha düşük I. tip hataya sahip olduğunu belirtmiştir. Ancak bu çalışma sonuçlarına göre, kısa testlerde geleneksel indeksler çok yüksek I. tip hata oranına sahip olduğu bulunmuştur. Ancak uzun testlerde (40 madde ve üzeri) geleneksel indeksler daha makul I. tip hata değerlerine sahip olduğu sonucuna ulaşılabilir. Bu çalışmada test uzunluğunun değişimlenen tüm düzeylerinde en küçük I. tip hata oranına $S\text{-}\chi^2$ indeksi sahiptir. Test uzunluğunun artırılması, madde-uyum indekslerin güç oranlarını önemli ölçüde etkilememektedir. GM=3PL CM=2PL olduğu durumlarda test uzunluğu arttıkça madde-uyum indeksinin gücü azalmaktadır. Ayrıca bu durum için indekslerin güç oranı orta düzeydedir. GM=3PL CM=2PL iken $S\text{-}\chi^2$ en düşük güce sahip indekstir. Ancak GM=3PL CM=1PL ve GM=2PL CM=1PL olduğu durumlar için indekslerin güç oranları test uzunluğunun artması ile önemli ölçüde değişmemektedir. Ayrıca indekslerin güç oranları test uzunluğunun tüm koşullarında yüksek bulunmuştur (0.90-1.00 arası). Elde edilen bu sonuçlar Orlando ve Thissen (2000, 2003), Stone ve Zhang'in (2003) çalışmalarının sonuçları ile tutarlılık göstermektedir.

Çalışma kapsamında çeşitli uyumsuzluk yüzde değerleri kullanılarak, uyumsuz madde yüzdesinin, madde-uyum indekslerinin, uyumsuz maddeleri tespit etme yeteneklerini nasıl etkilediği incelenmiştir. Uyumsuzluk yüzdesinin, indekslerin I. tip hatalarına etkisi incelendiğinde, uyumsuz madde sayısı arttıkça indekslerin I. tip hata oranları önemli bir şekilde değişmemektedir. Ancak GM=3PL CM=2PL olduğu durumda, yüksek uyumsuzluk yüzdesinde geleneksel indekslerin I. tip hata oranları artmaktadır soncuna ulaşılmıştır. Tüm uyumsuzluk yüzdelerinde, en düşük I. tip hataya $S\text{-}\chi^2$ indeksi sahiptir. Uyumsuzluk yüzdesinin, indekslerin güç oranlarına etkisi incelendiğinde, uyumsuzluk yüzdesi arttıkça indekslerin güç oranlarında küçük azalmalar olmaktadır. Ancak bu

değerler çok yüksek değildir. Wells ve Bolt (2008), yapmış oldukları çalışmada, uyumsuzluk yüzdesinin, indekslerin I. tip hatalarına ve güçlerine önemli bir etkisi olmadığını belirtmiştir. Bu çalışma sonucunda, Wells ve Bolt'un (2008) çalışmalarının sonucuna benzer olarak, uyumsuzluk yüzdesinin, indekslerin I. tip hata ve güç oranlarına önemli bir temel etkisinin olmadığı sonucuna ulaşılmıştır.

Çalışmada yer alan faktörlerin temel etkisinin yanı sıra ortak etkileri de I. tip hata ve güç oranlarına göre incelenmiştir. Literatür incelendiğinde (McKinley ve Mills, 1985; Yen, 1981) χ^2, Q_1 ve G^2 indekslerinin performanslarının yakın olduğu belirtilmiştir. Bu çalışmada da bu indeksler benzer performanslar göstermişlerdir. Ayrıca Orlando ve Thissen (2003) χ^2 ve Q_1 indekslerin 40 ve üzeri test uzunluğu ve 500 ile 1000 örneklem büyüklüğünde I. tip hataların makul düzeyde olduğunu belirtmiştir. Benzer bir şekilde bu çalışmada, örneklem büyüklüğü 1000, test uzunluğu 40 ve uyumsuzluk yüzdesi 0 olduğu durumlarda geleneksel indeksler S-χ^2 indeksi ile benzer ve düşük I. tip hata değerine sahiptir. Ancak büyük örneklemlerde, kısa testlerde ve yüksek uyumsuzluk yüzdelerinde S-χ^2 indeksi diğer indekslere göre daha düşük I. tip hataya sahiptir. Uyumsuzluk yüzdesinin indekslerin I. tip hata ve güç oranlarına önemli ölçüde temel etkisi olmadığı belirtilmiştir. Ancak uyumsuzluk yüzdesinin çalışmada yer alan diğer faktörlerle ortak etkisi incelendiğinde yüksek uyumsuzluk yüzdesinde (% 50) S-χ^2 indeksinin güç oranı diğer indekslere göre artmaktadır. Ortak etki açısından incelendiğinde test uzunluğunun değişimlenmesinin indekslerin güç oranına anlamlı etki etmediği sonucuna ulaşılmıştır.

Sonuç olarak, geleneksel madde-uyum indeksleri küçük örneklemlerde (1000), uzun testlerde (40 madde) ve uyumsuzluk yüzdesinin çok yüksek olmadığı durumlarda yeterli sonuçlar vermektedir. Chone, Lee ve Ansley (2007) S-χ^2 indeksinin uyum iyiliği istatistikleri arasında madde uyumunu değerlendirmede iyi bir alternatif olduğunu belirtmiştir. Ayrıca Orlando ve Thissen (2000) 1000 ve Stone ve Zhang (2003) 2000 ve üzeri örneklem büyüklüğünde S-χ^2 indeksinin tercih edilebileceğini belirtmiştir. İlgili literatürdeki çalışmalara benzer şekilde, bu çalışma sonucunda, 1000 örneklem ve üzerinde, 20 ve üzeri test uzunluklarında S-χ^2 indeksi, diğer alternatif indekslere göre daha doğru sonuçlar vermektedir. Ayrıca S-χ^2 indeksi, uyumsuzluk yüzdesinin yüksek olduğu durumlarda madde-uyumunun değerlendirildiği çalışmalarda tercih edilebilir.

Bu çalışmada, örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdelerinin, χ^2, Q_1, G^2 ve S-χ^2 indekslerinin I. tip hata ve güç oranlarına etkisi incelenmiştir. Madde uyumunu etkileyen bu faktörlerin yanı sıra indekslerin I. tip hata ve güç oranlarına etki eden uyumsuzluk büyüklüğü ve uyumsuzluk yeri (misfit location) gibi faktörler çalışmaya dahil edilerek indekslerin I. tip hata ve güç oranlarına etkileri incelenebilir. Ayrıca Stone (2000) tarafından ölçeklenmiş düzenlenmiş χ^2* indeksi veya bayese dayalı madde-uyum indeksleri kullanılarak bu faktörler altında uyumsuz maddeleri tespit etme oranları hesaplanabilir. Bu çalışmada, ikili puanlanan maddeler için simülatif veriler kullanılmıştır. Çoklu puanlanan maddeler için hem gerçek hem de simülatif veriler kullanılarak farklı çalışmalar gerçekleştirilebilir. Bununla birlikte, kayıp verili ve MTK'nın sayıltılarının yerine getirilmediği durumlara, gerçek hayattan elde edilen verilerde sıklıkla karşılaşılmaktadır. Gerçek durumlara yakın veriler elde edilerek benzer indekslerin I. tip hata ve güç oranları incelenebilir.

KAYNAKÇA

Ames, A. J. (2015). *Bayesian model criticism: Prior sensitivity of the posterior predictive checks method* (Doctoral dissertation). University of North Carolina.

Ames, A. J., & Penfield, D. R. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*(3), 39–48.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Chon, K. H., Lee, W. C., & Ansley, T. N. (2007). *Assessing IRT model-data fit for mixed format tests*. Center for Advanced Studies in Measurement and Assessment CASMA Research Report, No: 26.

Chon, K. H., Lee, W. C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement, 47*(3), 318–338.

DeMars, C. E. (2005). Type I error rates for PARSCALE's fit index. *Educational and Psychological Measurement, 65*, 42–50.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19,* 143-165.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Glas, C. A. W., & Su´arez Falc´on, J. C. (2003). A comparison of item-fit statistics for the three parameter logistic model. *Applied Psychological Measurement, 27*, 87–106.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* New York: Springer Science+Business Media, LLC.

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Kang, T., & Chen, T. T. (2008). Performance of the generalized *S-*χ^2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391-406.

Kang, T., & Chen, T. T. (2011). Performance of the generalized S-χ^2 item fit index for the graded response model. *Asia Pacific Educ. Rev., 12,* 89–96.

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of Item Response Theory item fit indices for the graded response model. *Organizational Research Methods 14(1),* 10-23.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement,* 9, 49-57.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG-W. Item analysis and test scoring with binary logistic models*. Moresville, IN: Scientific Software.

Orlando, M. (1997). *Item fit in the context of Item Response Theory.* (Doctoral dissertation, University of North Carolina, 1997). Dissertation Abstracts International, 58/04-B, 2175.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous Item Response Theory models. *Applied Psychological Measurement, 24(1),* 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of *S-*χ^2: An item fit index for use with dichotomous Item Response Theory models. *Applied Psychological Measurement, 27*, 289-298.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Pyschological Measurement, 14(2),* 127-137.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*, 58-75.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of Item Response Theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40,* 331–352.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement, 35(4),* 280–295.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item Response Theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39–49.

von Schrader, S., Ansley, T. N., & Kim, S. (2004). *Examination of item fit indices for polytomous item response models.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in Item Response Theory. *Applied Measurement in Education, 21*(1), 22-40.

Wells, C. S., & Hambleton, R. K. (2016). *Model fit with residual analyses.* In W. J. van der Linden (Ed.) *Handbook of item response theory*. Volume two, Statistical tools (pp. 395-413). NewYork: CRC Press. Taylor ve Francis Group.

Wright, B., & Panchapakesan, N. A. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29*, 23–48.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262.

EXTENDED ABSTRACT

Introduction

Item Response Theory (IRT) is a strong scaling technique that uses mathematical models, which estimate performance of examinee, by using item and person's characteristics (Embretson & Reise,

2000). IRT has many advantages in interpreting and reporting the test scores. However these advantages are valid only test data are fitted to selected model (Hambleton & Swaminathan, 1985). Beside this, the property of parameter invariance of IRT models is one of the most important issues in IRT. Parameter invariance property of IRT models is achieved only when the model and data are fit. Thus model data fit plays important role for valid applications of IRT models to obtain decisive score scales. There are many item-fit indices to examine model data fit. These indices can be classified as traditional indices and alternative indices. Traditional indices such as χ^2, Q_1, G^2 examine model data fit by comparing observed performance with estimated performance in various ability groups for each item under chosen IRT model. In traditional indices, ability interval is usually determined arbitrarily. These arbitrary ability intervals can influence the results of the study (Orlando & Thissen, 2000; Reise, 1990). Observed responses are model-dependent since discrete intervals are depended to ability. On the contrast, in alternative indices such as $S\text{-}\chi^2$, $S\text{-}G^2$, individuals do not have to be grouped arbitrary interval along the ability. These indices are conditioned on total score instead of ability. In this study, χ^2, Q_1 and G^2 indices as traditional indices and $S\text{-}\chi^2$ as an alternative index were examined.

Method

The purpose of this study was to investigate Type I error and power rates of the item fit indices through various conditions (sample sizes, different test lengths and different magnitudes of misfit) for dichotomously generated items based on one-, two-, and three-parameter logistic models in Item Response Theory. Hence, it is expected to contribute theoretical studies related to item fit in dichotomous IRT models. Thus, this study is theoretical research in this respect. In this study, the type I error and power rates of these item fit indices were assessed in a simulation study. χ^2, Q_1 and G^2 indices as traditional item fit indices and $S\text{-}\chi^2$ index as alternative indices were assessed. The performance of four different item fit indices in study were compared by manipulating three different sample size (1000, 2000, 4000), three different test lengths (20, 30, 40) and four different misfit percentage (%0, %10, %30 and %50). Item responses were generated using the R 3.3.2 software program. Data were generated for both generating models (GM) and analysis models (CM) with respect to manipulated factors and level of these factors. For all GM, item discrimination parameters were are 1, b parameters were obtained from uniform distribution which minimum was -2 and maximum was 2 and c parameters were 0. For all CM, b parameters were generated by adding +-0.75 to b parameters in GM. For 2PL and 3 PL models in CM, item discrimination parameters were obtained from uniform distribution which minimum was 0 and maximum was 2. For 3PL in analysis model, c parameter were 0.25. Ability distribution of examinees was obtained from normal distribution which mean was 0 and standard deviation was 1. In three conditions, CM has fewer parameters than GM and in other three conditions; CM and GM have identical parameters. After data generation, item fit indices were analyzed by using "mirt" package in R software. The p value of item fit indices and their degrees of freedom were calculated for both item responses for GM and the item responses for CM. Type I errors and power rates of item fit indices were examined according to significance levels of 0.05. All item fit indices in this study were compared by calculating the type I error for conditions GM=CM and power rates for conditions GM>CM of each item fit indices under all conditions.

Results and Discussion

Sample size is an important factor to assess the item fit indices. The results of this study indicated that the performance of all item fit indices were influenced from sample sizes, however, the degree of effect of sample size differs in item fit indices. Type I errors of traditional indices inflated by increasing sample size. It is not suggested to use traditional indices when the sample size was larger than 1000. On the other hand, $S\text{-}\chi^2$ had acceptable Type I error across all sample size. These results are consistent with the results of the studies of Orlando & Thissen (2000), Orlando & Thissen (2003), Stone & Zhang (2003) and Chon & Dunbar (2010). Sample size has also effects on the

power rates of the item fit indices. Item indices performed well for detecting misfit items by increasing sample size. Test length is the one of the factors, which affects the Type I errors, and power rates of item fit indices. Type I errors of traditional item fit indices decreased by increasing test length, whereas, Type I errors of $S\text{-}\chi^2$ remained same. However, $S\text{-}\chi^2$ was more acceptable than traditional indices across all test lengths. Traditional indices and $S\text{-}\chi^2$ performed closely when the test length was 40. $S\text{-}\chi^2$ performed better than other indices for 20 and 30 test length. These results are identical to findings of studies of Orlando & Thissen (2000, 2003) and Stone & Zhang (2003). Percentage of misfit had no substantial influence on either Type I errors or power rates. These results are parallel with the results of the Wells and Bolt (2008) study. Overall, among the item fit indices, $S\text{-}\chi^2$ produced more accurate results across all conditions so that it is recommended for assessing item fit. Manipulating the test length did not affect the power rates of item fit indices in a way.